SPECTRAL CLUSTERING, Lesson 7.

Parameter estimation via the EM algorithm in probabilistic graph and contingency table models

Marianna Bolla, DSc. Prof. BME Math. Inst.

December 3, 2020

1 Parameter estimation in random graph models

We will discuss two basic types of parametric random graph models, and give algorithms for the maximum likelihood estimation of the parameters. The models are capable to find hidden partitions of the graph's vertices for given number of clusters. Since these special clustering algorithms do not need any preliminary information on the clusters, they correspond to the unsupervised learning of the data at hand.

1.1 EM algorithm for estimating the parameters of the stochastic block model (generatized random graph)

The so-called stochastic block model, introduced by Holland, Bickel, Karrer, Rohe was already discussed in the previous lessons. In fact, this is a generalized random graph model, formulated in terms of mixtures. Now we consider it as a parametric model, and want to estimate its parameters. The assumptions of the model are the following. Given a simple graph $G = (V, \mathbf{A})$ ($|V| = n$, with adjacency matrix $\mathbf{A}$) and k ($1 < k < n$), we are looking for the hidden k-partition $(V_1, \dots, V_k)$ of the vertices such that

- vertices are independently assigned to cluster V_a with probability π_a, $a = 1, \dots, k$; $\sum_{a=1}^{k} \pi_a = 1$;
- given the cluster memberships, vertices of V_a and V_b are connected independently, with probability

$$\mathbb{P}(i \sim j \,|\, i \in V_a, j \in V_b) = p_{ab}, \quad 1 \le a, b \le k.$$

The parameters are collected in the vector $\underline{\pi} = (\pi_1, \dots, \pi_k)$ and the $k \times k$ symmetric matrix $\mathbf{P}$ of p_{ab}'s.

Our statistical sample is the $n \times n$ symmetric, 0-1 adjacency matrix $\mathbf{A} = (a_{ij})$ of G. There are no loops, so the diagonal entries are zeros. Based on $\mathbf{A}$, we want to estimate the parameters of the above block model.

Using the theorem of mutually exclusive and exhaustive events, the likelihood function is the mixture of joint distributions of i.i.d. Bernoulli distributed entries:

$$\frac{1}{2}\sum_{1\le a,b\le k}\pi_a\pi_b\prod_{i\in V_a,j\in V_b,i\neq j}p_{ab}^{a_{ij}}(1-p_{ab})^{(1-a_{ij})}$$
$$=\frac{1}{2}\sum_{1\le a,b\le k}\pi_a\pi_b\cdot p_{ab}^{e_{ab}}(1-p_{ab})^{(n_{ab}-e_{ab})}.$$

This is the mixture of binomial distributions, where e_{ab} is the number of edges connecting vertices of V_a and V_b ($a\neq b$), while e_{aa} is twice the number of edges with both endpoints in V_a; further,

$$n_{ab}=|V_a|\cdot|V_b|\quad(a\neq b)\quad\text{and}\quad n_{aa}=|V_a|\cdot(|V_a|-1)\quad(a=1,\dots,k)\qquad(1)$$

are the numbers of possible edges between V_a, V_b and within V_a, respectively.

Here $\mathbf{A}$ is the incomplete data specification as the cluster memberships are missing. Therefore, it is straightforward to use the *Expectation-Maximization*, briefly EM algorithm, proposed by Dempster, Laird, and Rubin in 1978, for parameter estimation from incomplete data. This special application for mixtures is sometimes called *collaborative filtering*. .

First we complete our data matrix $\mathbf{A}$ with latent membership vectors $\underline{\Delta}_1,\dots,\underline{\Delta}_n$ of the vertices that are k-dimensional i.i.d. $Poly(1,\underline{\pi})$ (polinomially distributed) random vectors. More precisely, $\underline{\Delta}_i=(\Delta_{1i},\dots,\Delta_{ki})$, where $\Delta_{ai}=1$ if $i\in V_a$ and zero otherwise. Thus, the sum of the coordinates of any $\underline{\Delta}_i$ is 1, and $\mathbb{P}(\Delta_{ai}=1)=\pi_a$.

Based on these, the likelihood function above is

$$\frac{1}{2}\sum_{1\le a,b\le k}\pi_a\pi_b\cdot p_{ab}^{\sum_{i\neq j}\Delta_{ai}\Delta_{bj}a_{ij}}\cdot(1-p_{ab})^{\sum_{i\neq j}\Delta_{ai}\Delta_{bj}(1-a_{ij})}$$

that is maximized in the alternating **E** and **M** steps of the EM algorithm.

Note that that the complete likelihood would be the squareroot of

$$\begin{aligned}&\prod_{1\le a,b\le k}p_{ab}^{e_{ab}}\cdot(1-p_{ab})^{(n_{ab}-e_{ab})}\\&=\prod_{a=1}^{k}\prod_{i=1}^{n}\prod_{b=1}^{k}[p_{ab}^{\sum_{j:\,j\neq i}\Delta_{bj}a_{ij}}\cdot(1-p_{ab})^{\sum_{j:\,j\neq i}\Delta_{bj}(1-a_{ij})}]^{\Delta_{ai}}\end{aligned}\qquad(2)$$

that is valid only in case of known cluster memberships.

Starting with initial parameter values $\underline{\pi}^{(0)}$, $\mathbf{P}^{(0)}$ and membership vectors $\underline{\Delta}_1^{(0)},\dots,\underline{\Delta}_n^{(0)}$, the t-th step of the iteration is the following ($t=1,2,\dots$).

- **E**-step: we calculate the conditional expectation of each Δ_i conditioned on the model parameters and on the other cluster assignments obtained in step $t-1$ and collectively denoted by $M^{(t-1)}$. By the Bayes theorem, the responsibility of vertex i for cluster a is

$$\begin{aligned}\pi_{ai}^{(t)}&=\mathbb{E}(\Delta_{ai}\,|\,M^{(t-1)})\\&=\frac{\mathbb{P}(M^{(t-1)}|\Delta_{ai}=1)\cdot\pi_a^{(t-1)}}{\sum_{b=1}^{k}\mathbb{P}(M^{(t-1)}|\Delta_{bi}=1)\cdot\pi_b^{(t-1)}}\end{aligned}$$

$(a = 1, \dots, k;\ i = 1, \dots, n)$. For each i, $\pi_{ai}^{(t)}$ is proportional to the numerator, where

$$\mathbb{P}(M^{(t-1)} | \Delta_{ai} = 1)$$
$$= \prod_{b=1}^{k} (p_{ab}^{(t-1)})^{\sum_{j \neq i} \Delta_{bj}^{(t-1)} a_{ij}} \cdot (1 - p_{ab}^{(t-1)})^{\sum_{j \neq i} \Delta_{bj}^{(t-1)} (1 - a_{ij})}$$

is the part of the likelihood (18) effecting vertex i under the condition $\Delta_{ai} = 1$.

- **M**-step: we maximize the truncated binomial likelihood

$$p_{ab}^{\sum_{i \neq j} \pi_{ai}^{(t)} \pi_{bj}^{(t)} a_{ij}} \cdot (1 - p_{ab})^{\sum_{i \neq j} \pi_{ai}^{(t)} \pi_{bj}^{(t)} (1 - a_{ij})}$$

with respect to the parameter p_{ab}, for all a, b pairs separately. Obviously, the maximum is attained by the following estimators of p_{ab}'s comprising the symmetric matrix $\mathbf{P}^{(t)}$: $p_{ab}^{(t)} = \frac{\sum_{i,j:\, i \neq j} \pi_{ai}^{(t)} \pi_{bj}^{(t)} a_{ij}}{\sum_{i,j:\, i \neq j} \pi_{ai}^{(t)} \pi_{bj}^{(t)}}$ $(1 \leq a \leq b \leq k)$, where edges connecting vertices of clusters a and b are counted fractionally, multiplied by the membership probabilities of their endpoints.

The maximum likelihood estimator of $\underline{\pi}$ in the t-th step is $\underline{\pi}^{(t)}$ of coordinates $\pi_a^{(t)} = \frac{1}{n} \sum_{i=1}^{n} \pi_{ai}^{(t)}$ $(a = 1, \dots, k)$, while that of the membership vector $\underline{\Delta}_i$ is obtained by discrete maximization: $\Delta_{ai}^{(t)} = 1$ if $\pi_{ai}^{(t)} = \max_{b \in \{1,\dots,k\}} \pi_{bi}^{(t)}$ and 0, otherwise. (In case of ambiguity, the cluster with the smallest index is selected.) This choice of $\underline{\pi}$ will increase (better to say, not decrease) the likelihood function. Note that it is not necessary to assign vertices uniquely to the clusters, the responsibility π_{ai} of a vertex i can as well be regarded as the intensity of vertex i belonging to cluster a.

According to the general theory of the EM algorithm, in exponential families (as in the present case), convergence to a local maximum can be guaranteed (depending on the starting values), but it runs in polynomial time in the number of vertices n. However, the speed and limit of the convergence depends on the starting clustering, which can be chosen by means of preliminary application of some nonparametric multiway cut algorithm of the preceding lessons.

The above algorithm gives so-called fuzzy clusters (vertices belong with certain probabilities to them). It is also possible to just relocate the vertices between the clusters in the E-step, such that a vertex is assigned to the cluster where its likelihood (with the actual cluster's parameters) is the largest. In this case the membership probabilities are not estimated during the iteration, they will be just the final relative frequencies.

2 α–β models

We will amalgamate the Rash model (for rectangular binary tables) and the newly introduced α-β models (for random undirected graphs) in the framework of a semiparametric probabilistic graph model. Our purpose is to give a partition of the vertices of an observed graph so that the generated subgraphs and bipartite graphs obey these models, where their strongly connected parameters

give multiscale evaluation of the vertices at the same time. In this way, a heterogeneous version of the stochastic block model is built via mixtures of loglinear models and the parameters are estimated with a special EM iteration. In the context of social networks, the clusters can be identified with social groups and the parameters with attitudes of people of one group towards people of the other, which attitudes depend on the cluster memberships. The algorithm is applied to randomly generated and real-word data.

So far many parametric and nonparametric methods have been proposed for community detection in networks. In the nonparametric scenario, hierarchical or spectral methods were applied to maximize the two- or multiway Newman–Girvan modularity [1, 2, 3, 4]; more generally, spectral clustering tools, based on Laplacian or modularity spectra, proved to be feasible to find community, anticommunity, or regular structures in networks [5]. In the parametric setup, certain model parameters are estimated, usually via maximizing the *likelihood function* of the graph, i.e., the joint probability of our observations under the model equations. This so-called ML estimation is a promising method of statistical inference, has solid theoretical foundations [6, 7], and also supports the common-sense goal of accepting parameter values based on which our sample is the most likely.

In the 2010s, α-β-models [8, 9] were developed as the unique graph models where the degree sequence is a *sufficient statistic*: given the degree sequence, the distribution of the random graph does not depend on the parameters any more (microcanonical distribution over the model graphs). This fact makes it possible to derive the ML estimate of the parameters in a standard way [10]. Indeed, in the context of network data, a lot of information is contained in the degree sequence, though, perhaps in a more sophisticated way. The vertices may have *clusters* (groups or modules) and their memberships may affect their affinity to make ties. We will find groups of the vertices such that the within- and between-cluster edge-probabilities admit certain parametric graph models, the parameters of which are highly interlaced. Here the degree sequence is not a sufficient statistic any more, only if it is restricted to the subgraphs. When making inference, we are partly inspired by the stochastic block model, partly by the Rasch model, the rectangular analogue of the α-β models.

The first type of block models is the homogeneous one: the probability to make ties is the same within the clusters or between the cluster-pairs. Although this probability depends on the actual cluster memberships, given the memberships of the vertices, the probability that they are connected is a given constant (parameter to be estimated). This stochastic block model, sometimes called generalized random graph or planted partition model, is thoroughly discussed in [11, 12, 13, 14, 15].

Here we propose a heterogeneous block model by carrying on the Rasch model developed more than 50 years ago for evaluating psychological tests [16, 17]. Given the number of clusters and a classification of the vertices, we will use the Rasch model for the bipartite subgraphs, whereas the α-β models for the subgraphs themselves, and process an iteration (inner cycle) to find the ML estimate of their parameters. Then, based on the overall likelihood, we find a new classification of the vertices via taking conditional expectation and using the Bayes rule. Eventually, the two steps are alternated, giving the outer cycle of the iteration. Our algorithm fits into the framework of the EM algorithm, the convergence of which is proved in exponential families under very general con-

ditions [18, 7]. The method was originally developed for missing data, and the name comes from the alternating *expectation* (E) and *maximization* (M) steps, where in the E-step (assignment phase) we complete the data by substituting for the missing data via taking conditional expectation, while in the M-step (estimation phase) we find the usual ML estimate of the parameters based on the so completed data. The algorithm naturally extends to situations, when not the data itself is missing, but it comes from a mixture, and the grouping memberships are the missing parameters. This special type of the EM algorithm developed for mixtures is often called collaborative filtering [19, 20] or Gibbs sampling [21], the roots of which method can be traced back to [22]. In the context of social networks, the clusters can be identified with social strata and the parameters with attitudes of people of one group towards people of the other, which attitude is the same for people in the second group, but depends on the individual in the first group. The number of clusters is fixed during the iteration, but an initial number can be obtained by spectral clustering tools. Together with the proof of the convergence, the algorithm is applied to randomly generated and real-word data.

This kind of model building is originated both in the statistics literature, e.g., [23, 24, 25] and in the physics literature, e.g., [2, 26, 27]. In [28], the author already considers mixing according to vertex degree. In [13] the authors introduce the degree-corrected variant of the stochastic block model, but they use Poisson edge-probabilities. In [27] the likelihood, depending on Poisson parameters, is maximized with the trick that first a likelihood maximization is performed, then the problem is traced back to the minimum-cut objective. This is not the EM algorithm, though the idea of mixed tools resembles that.

In [29], without giving an algorithm, the authors maximize the so-called likelihood modularity over k-partitions of vertices, for given k. This is rather a non-parametric way of model fitting, since, instead of parameters, they substitute the relative frequency of the edges for their Bernoulli parameters, and theoretically maximize their profile likelihood with respect to the memberships of the vertices, which is considered as unknown parameter. They also prove the consistency of their estimates. [30] considers bipartition and multipartition of dense graphs with arbitrary degree distribution. In [15], based on the adjacency matrix as a statistical sample, the authors estimate the underlying partition of the vertices, given an upper bound for the number of blocks, in the stochastic block model. They prove that the suitably modified spectral partitioning procedure is consistent. Before fitting a model, its complexity may also be investigated. In [31], the authors give the quantification of the intrinsic complexity of undirected graphs and networks, via distinguishing between randomness complexity and statistical complexity.

2.1 Multiclass binary model

Loglinear type models to describe contingency tables were proposed, e.g., by [23, 24] and widely used in statistics. Together with the Rasch model, they give the foundation of our unweighted graph and bipartite graph models, the building blocks of our EM iteration. Note that in [23], the authors also extend their model to directed graphs.

2.2 α-β models for undirected random graphs

With different parameterization, [8] and [9] introduced the following random graph model, where the degree sequence is a sufficient statistic. We have an unweighted, undirected random graph on n vertices without loops, such that edges between distinct vertices come into existence independently, but not with the same probability as in the classical Erdős–Rényi model [32]. This random graph can uniquely be characterized by its $n \times n$ symmetric adjacency matrix $\mathbf{A} = (A_{ij})$ which has zero diagonal and the entries above the main diagonal are independent Bernoulli random variables whose parameters $p_{ij} = \mathbb{P}(A_{ij} = 1)$ obey the following rule. Actually, we formulate this rule for the $\frac{p_{ij}}{1-p_{ij}}$ ratios, the so-called *odds*:

$$\frac{p_{ij}}{1-p_{ij}} = \alpha_i \alpha_j \quad (1 \le i < j \le n), \tag{3}$$

where the parameters $\alpha_1, \dots, \alpha_n$ are positive reals. This model is called α *model* in [9]. With the parameter transformation $\beta_i = \ln \alpha_i$ $(i = 1, \dots n)$, it is equivalent to the β *model* of [8] which applies to the *log-odds*:

$$\ln \frac{p_{ij}}{1-p_{ij}} = \beta_i + \beta_j \quad (1 \le i < j \le n) \tag{4}$$

with real parameters $\beta_1, \dots, \beta_n$.

Conversely, the probabilities p_{ij} and $1 - p_{ij}$ can be expressed in terms of the parameters, like

$$p_{ij} = \frac{\alpha_i \alpha_j}{1 + \alpha_i \alpha_j} \quad \text{and} \quad 1 - p_{ij} = \frac{1}{1 + \alpha_i \alpha_j} \tag{5}$$

which formulas will be intensively used in the subsequent calculations.

We are looking for the ML estimate of the parameter vector $\underline{\alpha} = (\alpha_1, \dots, \alpha_n)$ or $\underline{\beta} = (\beta_1, \dots, \beta_n)$ based on the observed unweighted, undirected graph as a statistical sample. (It may seem that we have a one-element sample here, however, there are $\binom{n}{2}$ independent random variables, the adjacencies, in the background.)

Let $\mathbf{D} = (D_1, \dots, D_n)$ denote the degree-vector of the above random graph, where $D_i = \sum_{j=1}^{n} A_{ij}$ $(i = 1, \dots n)$. The random vector $\mathbf{D}$, as a function of the sample entries A_{ij}'s, is a *sufficient statistic* for the parameter $\underline{\alpha}$, or equivalently, for $\underline{\beta}$. Roughly speaking, a sufficient statistic itself contains all the information – that can be retrieved from the data – for the parameter. More precisely, a statistic is sufficient when the conditional distribution of the sample, given the statistic, does not depend on the parameter any more. By the *Neyman–Fisher factorization theorem* [6], a statistic is sufficient if and only if the likelihood function of the sample can be factorized into two parts: one which does not contain the parameter, and the other, which includes the parameter, contains the sample entries merely compressed into this sufficient statistic. Consider this factorization of the likelihood function (joint probability of A_{ij}'s) in our case.

Because of the symmetry of $\mathbf{A}$, this is

$$\begin{aligned}
L_{\underline{\alpha}}(\mathbf{A}) &= \prod_{i=1}^{n-1}\prod_{j=i+1}^{n} p_{ij}^{A_{ij}}(1-p_{ij})^{1-A_{ij}} \\
&= \left\{\prod_{i=1}^{n}\prod_{j=1}^{n} p_{ij}^{A_{ij}}(1-p_{ij})^{1-A_{ij}}\right\}^{1/2} \\
&= \left\{\prod_{i=1}^{n}\prod_{j=1}^{n}\left(\frac{p_{ij}}{1-p_{ij}}\right)^{A_{ij}}\prod_{i=1}^{n}\prod_{j=1}^{n}(1-p_{ij})\right\}^{1/2} \\
&= \left\{\prod_{i=1}^{n}\alpha_i^{\sum_{j=1}^{n}A_{ij}}\prod_{j=1}^{n}\alpha_j^{\sum_{i=1}^{n}A_{ij}}\prod_{i\neq j}(1-p_{ij})\right\}^{1/2} \\
&= \left\{\prod_{i\neq j}\frac{1}{1+\alpha_i\alpha_j}\right\}^{1/2}\left\{\prod_{i=1}^{n}\alpha_i^{D_i}\prod_{j=1}^{n}\alpha_j^{D_j}\right\}^{1/2} \\
&= \left\{\prod_{i<j}\frac{1}{1+\alpha_i\alpha_j}\right\}\left\{\prod_{i=1}^{n}\alpha_i^{D_i}\right\} = C_{\underline{\alpha}}\times\prod_{i=1}^{n}\alpha_i^{D_i}
\end{aligned}$$

where we used (5) and the facts that $A_{ij}=A_{ji}$, $p_{ij}=p_{ji}$ $(i<j)$ and $A_{ii}=0$, $p_{ii}=0$ $(i=1,\dots,n)$; further, used the convention $0^0=1$. Here the partition function $C_{\underline{\alpha}}=\prod_{i<j}\frac{1}{1+\alpha_i\alpha_j}$ only depends on $\underline{\alpha}$, and the whole likelihood function depends on the A_{ij}'s merely through D_i's. Therefore, $\mathbf{D}$ is a sufficient statistic. The other factor is constantly 1, indicating that the conditional joint distribution of the entries – given $\mathbf{D}$ – is uniform, but we will not make use of this fact. Note that in [13], the authors call the uniform distribution on graphs with fixed degree sequence *microcanonical*. In [8, 10] the converse statement is also proved: the above α model (reparametrized as β model) is the unique one, where the degree sequence is a sufficient statistic.

Let (a_{ij}) be the matrix of the sample realizations (the adjacency entries of the observed graph), $d_i=\sum_{j=1}^{n}a_{ij}$ be the actual degree of vertex i $(i=1,\dots,n)$ and $\mathbf{d}=(d_1,\dots,d_n)$ be the observed degree-vector. The above factorization also indicates that the joint distribution of the entries belongs to the exponential family, and hence, with natural parameterization [18], the maximum likelihood estimate $\hat{\underline{\alpha}}$ (or equivalently, $\hat{\underline{\beta}}$) is derived from the fact that, with it, the observed degree d_i equals the expected one, that is $\mathbb{E}(D_i)=\sum_{i=1}^{n}p_{ij}$. Therefore, $\hat{\underline{\alpha}}$ is the solution of the following *maximum likelihood equation*:

$$d_i=\sum_{j\neq i}^{n}\frac{\alpha_i\alpha_j}{1+\alpha_i\alpha_j}\quad(i=1,\dots,n). \tag{6}$$

The ML estimate $\hat{\underline{\beta}}$ is easily obtained from $\hat{\underline{\alpha}}$ via taking the logarithms of its coordinates.

Before discussing the solution of the system of equations (6), let us see, what conditions a sequence of nonnegative integers should satisfy so that it could be realized as the degree sequence of a graph. The sequence $d_1,\dots,d_n$

of nonnegative integers is called *graphic* if there is an unweighted, undirected graph on n vertices such that its vertex-degrees are the numbers $d_1, \dots, d_n$ in some order. Without loss of generality, d_i's can be enumerated in non-increasing order. The Erdős–Gallai theorem [33] gives the following necessary and sufficient condition for a sequence to be graphic. The sequence $d_1 \geq \cdots \geq d_n \geq 0$ of integers is graphic if and only if it satisfies the following two conditions: $\sum_{i=1}^{n} d_i$ is even and

$$\sum_{i=1}^{k} d_i \leq k(k-1) + \sum_{i=k+1}^{n} \min\{k, d_i\}, \quad k = 1, \dots, n-1. \tag{7}$$

Note that for nonnegative (not necessarily integer) real sequences a continuous analogue of (7) is derived in [8]. For given n, the convex hull of all possible graphic degree sequences is a polytope, to be denoted by $\mathcal{D}_n$. Its extreme points are the so-called *threshold graphs* [34]. It is interesting that for $n = 3$ all undirected graphs are threshold, since there are 8 possible graphs on 3 nodes, and there are also 8 vertices of $\mathcal{D}_3$; the $n = 2$ case is also not of much interest, therefore we will treat the $n > 3$ cases only. The number of vertices of $\mathcal{D}_n$ superexponentially grows with n [35], therefore the problem of characterizing threshold graphs has a high computational complexity. Its facial and cofacial sets are fully described in [10]. Apart from the trivial cases (when there is at least one degree equal to 0 or $n-1$), in [36], the authors give the following equivalent characterization of a threshold graph for $n \geq 4$: it has no four different vertices a, b, c, d such that a, b and c, d are connected by an edge, but a, c and b, d not, i.e., it has no two disjoint copies of the complete graph K_2.

The authors of [8, 9] prove that $\mathcal{D}_n$ is the topological closure of the set of expected degree sequences, and for given $n > 3$, if $\mathbf{d} \in \operatorname{int}(\mathcal{D}_n)$ is an interior point, then the maximum likelihood equation (6) has a unique solution. Later, it turned out that the converse is also true: in [10] the authors prove that the ML estimate exists if and only if the observed degree vector is an inner point of $\mathcal{D}_n$. On the contrary, when the observed degree vector is a boundary point of $\mathcal{D}_n$, there is at least one 0 or 1 probability p_{ij} which can be obtained only by a parameter vector such that at least one of the β_i's is not finite. In this case, the likelihood function cannot be maximized with a finite parameter set, its supremum is approached with a parameter vector $\underline{\beta}$ with at least one coordinate tending to $+\infty$ or $-\infty$. We also remark that, for 'large' n, the condition $\mathbf{d} \in \operatorname{int}(\mathcal{D}_n)$ is strongly related to the δ-tameness condition of [37], or to the fact that $\mathbf{d}$ has a 'scaling limit' defined in [8], also to the notion of 'there are no dominant vertices' of [38].

The authors in [9] recommend the following algorithm and prove that, provided $\mathbf{d} \in \operatorname{int}(\mathcal{D}_n)$, the iteration of it converges to the unique solution of the system (6). To motivate the iteration, we rewrite (6) as

$$d_i = \alpha_i \sum_{j \neq i} \frac{1}{\frac{1}{\alpha_j} + \alpha_i} \quad (i = 1, \dots, n).$$

Then starting with initial parameter values $\alpha_1^{(0)}, \dots, \alpha_n^{(0)}$ and using the observed degree sequence $d_1, \dots, d_n$, which is an inner point of $\mathcal{D}_n$, the iteration is as

follows:

$$\alpha_i^{(t)} = \frac{d_i}{\sum_{j\neq i} \frac{1}{\frac{1}{\alpha_j^{(t-1)}} + \alpha_i^{(t-1)}}} \quad (i = 1, \ldots, n) \tag{8}$$

for $t = 1, 2, \ldots$, until convergence.

2.3 β-γ model for bipartite graphs

This bipartite graph model traces back to Haberman [39], Lauritzen [24], and Rasch [16, 17] who applied it for psychological and educational measurements, later market research. The frequently cited Rasch model involves categorical data, mainly binary variables, therefore the underlying random object can be thought of as a contingency table. According to the Rasch model, the entries of an $m \times n$ binary table $\mathbf{A}$ are independent Bernoulli random variables, where for the parameter p_{ij} of the entry A_{ij} the following holds:

$$\ln \frac{p_{ij}}{1 - p_{ij}} = \beta_i - \delta_j \quad (i = 1, \ldots m;\, j = 1, \ldots, n) \tag{9}$$

with real parameters $\beta_1, \ldots, \beta_m$ and $\delta_1, \ldots, \delta_n$. As an example, Rasch in [16] investigated binary tables where the rows corresponded to persons and the columns to items of some psychological test, whereas the jth entry of the ith row was 1 if person i answered test item j correctly and 0, otherwise. He also gave a description of the parameters: β_i was the ability of person i, while δ_j the difficulty of test item j. Therefore, in view of the model equation (9), the more intelligent the person and the less difficult the test, the larger the success/failure ratio was on a logarithmic scale.

Given an $m \times n$ random binary table $\mathbf{A} = (A_{ij})$, or equivalently, a bipartite graph, our model is

$$\ln \frac{p_{ij}}{1 - p_{ij}} = \beta_i + \gamma_j \quad (i = 1, \ldots, m,\, j = 1, \ldots, n) \tag{10}$$

with real parameters $\beta_1, \ldots, \beta_m$ and $\gamma_1, \ldots, \gamma_n$; further, $p_{ij} = \mathbb{P}(A_{ij} = 1)$.

In terms of the transformed parameters $b_i = e^{\beta_i}$ and $g_j = e^{\gamma_j}$, the model (10) is equivalent to

$$\frac{p_{ij}}{1 - p_{ij}} = b_i g_j \quad (i = 1, \ldots, m,\, j = 1, \ldots, n) \tag{11}$$

where $b_1, \ldots, b_m$ and $g_1, \ldots, g_n$ are positive reals.

Conversely, the probabilities can be expressed in terms of the parameters:

$$p_{ij} = \frac{b_i g_j}{1 + b_i g_j} \quad \text{and} \quad 1 - p_{ij} = \frac{1}{1 + b_i g_j}. \tag{12}$$

Observe that if (10) holds with the parameters β_i's and γ_j's, then it also holds with the transformed parameters $\beta_i' = \beta_i + c$ $(i = 1, \ldots, m)$ and $\gamma_j' = \gamma_j - c$ $(j = 1, \ldots, n)$ with some $c \in \mathbb{R}$. Equivalently, if (11) holds with the positive parameters b_i's and g_j's, then it also holds with the transformed parameters

$$b_i' = b_i \kappa \quad \text{and} \quad g_j' = \frac{g_j}{\kappa} \tag{13}$$

with some $\kappa > 0$. Therefore, the parameters b_i and g_j are arbitrary to within a multiplicative constant.

Here the row-sums $R_i = \sum_{j=1}^{n} A_{ij}$ and the column-sums $C_j = \sum_{i=1}^{m} A_{ij}$ are the sufficient statistics for the parameters collected in $\mathbf{b} = (b_1, \ldots, b_m)$ and $\mathbf{g} = (g_1, \ldots, g_n)$. Indeed, the likelihood function is factorized as

$$\begin{aligned} L_{\mathbf{b},\mathbf{g}}(\mathbf{A}) &= \prod_{i=1}^{m}\prod_{j=1}^{n} p_{ij}^{A_{ij}}(1-p_{ij})^{1-A_{ij}} \\ &= \left\{\prod_{i=1}^{m}\prod_{j=1}^{n}\left(\frac{p_{ij}}{1-p_{ij}}\right)^{A_{ij}}\right\}\prod_{i=1}^{m}\prod_{j=1}^{n}(1-p_{ij}) \\ &= \left\{\prod_{i=1}^{m} b_i^{\sum_{j=1}^{n} A_{ij}}\right\}\left\{\prod_{j=1}^{n} g_j^{\sum_{i=1}^{m} A_{ij}}\right\}\prod_{i=1}^{m}\prod_{j=1}^{n}(1-p_{ij}) \\ &= \left\{\prod_{i=1}^{m}\prod_{j=1}^{n}\frac{1}{1+b_i g_j}\right\}\left\{\prod_{i=1}^{m} b_i^{R_i}\right\}\left\{\prod_{j=1}^{n} g_j^{C_j}\right\}. \end{aligned}$$

Since the likelihood function depends on $\mathbf{A}$ only through its row- and column-sums, by the Neyman–Fisher factorization theorem, $R_1, \ldots, R_m, C_1, \ldots, C_n$ is a sufficient statistic for the parameters. The first factor (including the partition function) depends only on the parameters and the row- and column-sums, whereas the seemingly not present factor – which would depend merely on $\mathbf{A}$ – is constantly 1, indicating that the conditional joint distribution of the entries, given the row- and column-sums, is uniform (microcanonical) in this model. Note that in [37], the author characterizes random tables sampled uniformly from the set of 0-1 matrices with fixed margins. Given the margins, the contingency tables coming from the above model are uniformly distributed, and a typical table of this distribution is produced by the β-γ model with parameters estimated via the row- and column sums as sufficient statistics. In this way, here we obtain another view of the typical table of [37].

Based on an observed binary table (a_{ij}), since we are in exponential family, and $\beta_1, \ldots, \beta_m, \gamma_1, \ldots, \gamma_n$ are natural parameters, the likelihood equation is obtained by making the expectation of the sufficient statistic equal to its sample value. Therefore, with the notation $r_i = \sum_{j=1}^{n} a_{ij}$ $(i = 1, \ldots, m)$ and $c_j = \sum_{i=1}^{m} a_{ij}$ $(j = 1, \ldots, n)$, the following *system of likelihood equations* is yielded:

$$\begin{aligned} r_i &= \sum_{j=1}^{n}\frac{b_i g_j}{1+b_i g_j} = b_i\sum_{j=1}^{n}\frac{1}{\frac{1}{g_j}+b_i}, \quad i = 1, \ldots m; \\ c_j &= \sum_{i=1}^{m}\frac{b_i g_j}{1+b_i g_j} = g_j\sum_{i=1}^{m}\frac{1}{\frac{1}{b_i}+g_j}, \quad j = 1, \ldots n. \end{aligned} \tag{14}$$

Note that for any sample realization of $\mathbf{A}$,

$$\sum_{i=1}^{m} r_i = \sum_{j=1}^{n} c_j \tag{15}$$

holds automatically. Therefore, there is a dependence between the equations of the system (14), indicating that the solution is not unique, in accord with

our previous remark about the arbitrary scaling factor $\kappa > 0$ of (13). We will prove that apart from this scaling, the solution is unique if it exists at all. For our convenience, let $(\tilde{\mathbf{b}}, \tilde{\mathbf{g}})$ denote the equivalence class of the parameter vector $(\mathbf{b}, \mathbf{g})$, which consists of parameter vectors $(\mathbf{b}', \mathbf{g}')$ satisfying (13) with some $\kappa > 0$. So that to avoid this indeterminacy, we may impose conditions on the parameters, for example,

$$\sum_{i=1}^{m} \beta_i + \sum_{j=1}^{n} \gamma_j = 0. \tag{16}$$

Like the graphic sequences, here the following sufficient conditions can be given for the sequences $r_1 \geq \cdots \geq r_m > 0$ and $c_1 \geq \cdots \geq c_n > 0$ of integers to be row- and column-sums of an $m \times n$ matrix of 0-1 entries (see, e.g., [40]):

$$\begin{aligned} \sum_{i=1}^{k} r_i &\leq \sum_{j=1}^{n} \min\{c_j, k\}, \quad k = 1, \ldots, m; \\ \sum_{j=1}^{k} c_j &\leq \sum_{i=1}^{m} \min\{r_i, k\}, \quad k = 1, \ldots, n. \end{aligned} \tag{17}$$

Observe that the $k = 1$ cases imply $r_1 \leq n$ and $c_1 \leq m$; whereas the $k = m$ and $k = n$ cases together imply $\sum_{i=1}^{m} r_i = \sum_{j=1}^{n} c_j$. This statement is the counterpart of the Erdős-Gallai conditions for bipartite graphs, where – due to (15) – the sum of the degrees is automatically even. In fact, the conditions in (17) are redundant: one of the conditions – either the one for the rows, or the one for the columns – suffices together with (15) and $c_1 \leq m$ or $r_1 \leq n$. The so obtained necessary and sufficient conditions define *bipartite realizable sequences* with the wording of [36]. Already in 1957, the author [41] determined arithmetic conditions for the construction of a 0-1 matrix having given row- and column-sums. The construction was given via swaps. More generally, [42] referred to the transportation problem and the Ford–Fulkerson max flow–min cut theorem [43].

The convex hull of the bipartite realizable sequences $\mathbf{r} = (r_1, \ldots, r_m)$ and $\mathbf{c} = (c_1, \ldots, c_n)$ form a polytope in $\mathbb{R}^{m+n}$, actually, because of (15), in an $(m+n-1)$-dimensional hyperplane of it. It is called *polytope of bipartite degree sequences* and denoted by $\mathcal{P}_{m,n}$ in [36]. It is the special case of the transportation polytope describing margins of contingency tables with nonnegative integer entries. There is an expanding literature on the number of such matrices, e.g., [44], and on the number of 0-1 matrices with prescribed row and column sums, e.g., [45].

Analogously to the considerations of the α-β models, and applying the thoughts of the proofs in [8, 9, 10], $\mathcal{P}_{m,n}$ is the closure of the set of the expected row- and column-sum sequences in the above model. In [36] it is proved that an $m \times n$ binary table, or equivalently a bipartite graph on the independent sets of m and n vertices, is on the boundary of $\mathcal{P}_{m,n}$ if it does not contain two vertex-disjoint edges. In this case, the likelihood function cannot be maximized with a finite parameter set, its supremum is approached with a parameter vector with at least one coordinate β_i or γ_j tending to $+\infty$ or $-\infty$, or equivalently, with at least one coordinate b_i or g_j tending to $+\infty$ or 0. Based on the proofs of [10], and stated as Theorem 6.3 in the supplementary material of [10], the maximum likelihood estimate of the parameters of model (11) exists if and only if the observed row- and column-sum sequence $(\mathbf{r}, \mathbf{c}) \in \mathrm{ri}\,(\mathcal{P}_{m,n})$, the relative

interior of $\mathcal{P}_{m,n}$, satisfying (15). In this case for the probabilities, calculated by the formula (12) through the estimated positive parameter values $\hat{b}_i$'s and $\hat{g}_j$'s (solutions of(14)), $0 < p_{ij} < 1$ holds $\forall i, j$.

Under these conditions, we define an algorithm that converges to the unique (up to the above equivalence) solution of the maximum likelihood equation (14). More precisely, we will prove that if $(\mathbf{r}, \mathbf{c}) \in \mathrm{ri}\,(\mathcal{P}_{m,n})$, then our algorithm gives a unique equivalence class of the parameter vectors as the fixed point of the iteration, which therefore provides the ML estimate of the parameters.

Starting with positive parameter values $b_i^{(0)}$ $(i = 1, \ldots, m)$ and $g_j^{(0)}$ $(j = 1, \ldots, n)$ and using the observed row- and column-sums, the iteration is as follows:

$$I. \quad b_i^{(t)} = \frac{r_i}{\sum_{j=1}^{n} \frac{1}{\frac{1}{g_j^{(t-1)}} + b_i^{(t-1)}}}, \quad i = 1, \ldots m$$

$$II. \quad g_j^{(t)} = \frac{c_j}{\sum_{i=1}^{m} \frac{1}{\frac{1}{b_i^{(t)}} + g_j^{(t-1)}}}, \quad j = 1, \ldots n$$

for $t = 1, 2, \ldots$, until convergence.

2.4 The EM iteration

In the several clusters case, we are putting the bricks together. The above discussed α-β and β-γ models will be the building blocks of a heterogeneous block model. Here the degree sequences are not any more sufficient for the whole graph, only for the building blocks of the subgraphs.

Given $1 \leq k \leq n$, we are looking for k-partition, in other words, clusters $C_1, \ldots, C_k$ of the vertices such that

- different vertices are independently assigned to a cluster C_u with probability π_u $(u = 1, \ldots, k)$, where $\sum_{u=1}^{k} \pi_u = 1$;
- given the cluster memberships, vertices $i \in C_u$ and $j \in C_v$ are connected independently, with probability p_{ij} such that

$$\ln \frac{p_{ij}}{1 - p_{ij}} = \beta_{iv} + \beta_{ju},$$

for any $1 \leq u, v \leq k$ pair. Equivalently,

$$\frac{p_{ij}}{1 - p_{ij}} = b_{ic_j} b_{jc_i},$$

where c_i is the cluster membership of vertex i and $b_{iv} = e^{\beta_{iv}}$.

The parameters are collected in the vector $\underline{\pi} = (\pi_1, \ldots, \pi_k)$ and the $n \times k$ matrix $\mathbf{B}$ of b_{iu}'s $(i \in C_u,\ u = 1, \ldots, k)$. The likelihood function is the following mixture:

$$\sum_{1 \leq u, v \leq k} \pi_u \pi_v \prod_{i \in C_u, j \in C_v} p_{ij}^{a_{ij}} (1 - p_{ij})^{(1 - a_{ij})}.$$

Here $\mathbf{A} = (a_{ij})$ is the incomplete data specification as the cluster memberships are missing. Therefore, it is straightforward to use the EM algorithm, proposed by [18], also discussed in [47, 7], for parameter estimation from incomplete data.

This special application for mixtures is sometimes called *collaborative filtering*, see [20, 19], which is rather applicable to fuzzy clustering.

First we complete our data matrix $\mathbf{A}$ with latent membership vectors $\mathbf{m}_1, \dots, \mathbf{m}_n$ of the vertices that are k-dimensional i.i.d. $Multy(1, \underline{\pi})$ (multinomially distributed) random vectors. More precisely, $\mathbf{m}_i = (m_{i1}, \dots, m_{ik})$, where $m_{iu} = 1$ if $i \in C_u$ and zero otherwise. Thus, the sum of the coordinates of any $\mathbf{m}_i$ is 1, and $\mathbb{P}(m_{iu} = 1) = \pi_u$.

Note that, if the cluster memberships where known, then the complete likelihood would be

$$\prod_{u=1}^{k} \prod_{i=1}^{n} \prod_{v=1}^{k} \prod_{j=1}^{n} [p_{ij}^{m_{jv} a_{ij}} \cdot (1 - p_{ij})^{m_{jv}(1-a_{ij})}]^{m_{iu}} \tag{18}$$

that is valid only in case of known cluster memberships.

Starting with initial parameter values $\underline{\pi}^{(0)}$, $\mathbf{B}^{(0)}$ and membership vectors $\mathbf{m}_1^{(0)}, \dots, \mathbf{m}_n^{(0)}$, the t-th step of the iteration is the following ($t = 1, 2, \dots$).

- **E**-step: we calculate the conditional expectation of each $\mathbf{m}_i$ conditioned on the model parameters and on the other cluster assignments obtained in step $t-1$, and collectively denoted by $M^{(t-1)}$.

  The responsibility of vertex i for cluster u in the t-th step is defined as the conditional expectation $\pi_{iu}^{(t)} = \mathbb{E}(m_{iu} \mid M^{(t-1)})$, and by the Bayes theorem, it is

  $$\pi_{iu}^{(t)} = \frac{\mathbb{P}(M^{(t-1)} | m_{iu} = 1) \cdot \pi_u^{(t-1)}}{\sum_{v=1}^{k} \mathbb{P}(M^{(t-1)} | m_{iv} = 1) \cdot \pi_v^{(t-1)}}$$

  ($u = 1, \dots, k$; $i = 1, \dots, n$). For each i, $\pi_{iu}^{(t)}$ is proportional to the numerator, therefore the conditional probabilities $\mathbb{P}(M^{(t-1)} | m_{iu} = 1)$ should be calculated for $u = 1, \dots, k$. But this is just the part of the likelihood (18) effecting vertex i under the condition $m_{iu} = 1$. Therefore,

  $$\begin{aligned}
  &\mathbb{P}(M^{(t-1)} | m_{iu} = 1) \\
  &= \prod_{v=1}^{k} \prod_{j \in C_v,\, j \sim i} \frac{b_{iv}^{(t-1)} b_{ju}^{(t-1)}}{1 + b_{iv}^{(t-1)} b_{ju}^{(t-1)}} \prod_{j \in C_v,\, j \not\sim i} \frac{1}{1 + b_{iv}^{(t-1)} b_{ju}^{(t-1)}} \\
  &= \prod_{v=1}^{k} \left\{ \frac{b_{iv}^{(t-1)} b_{ju}^{(t-1)}}{1 + b_{iv}^{(t-1)} b_{ju}^{(t-1)}} \right\}^{e_{vi}} \left\{ \frac{1}{1 + b_{iv}^{(t-1)} b_{ju}^{(t-1)}} \right\}^{|C_v| \cdot (|C_v| - 1)/2 - e_{vi}},
  \end{aligned}$$

  where e_{vi} is the number of edges within C_v that are connected to i.

- **M**-step: We update $\underline{\pi}^{(t)}$ and $\mathbf{m}^{(t)}$: $\pi_u^{(t)} := \frac{1}{n} \sum_{i=1}^{n} \pi_{iu}^{(t)}$ and $m_{iu}^{(t)} = 1$ if $\pi_{iu}^{(t)} = \max_v \pi_{iv}^{(t)}$ and 0, otherwise (in case of ambiguity, we select the smallest index for the cluster membership of vertex i). This is an ML estimation (discrete one, in the latter case, for the cluster membership). In this way, a new clustering of the vertices is obtained.

  Then we estimate the parameters in the actual clustering of the vertices. In the within-cluster scenario, we use the parameter estimation of model (3), obtaining estimates of b_{iu}'s ($i \in C_u$) in each cluster separately

$(u = 1, \ldots, k)$; as for cluster u, b_{iu} corresponds to α_i and the number of vertices is $|C_u|$. In the between-cluster scenario, we use the bipartite graph model (11) in the following way. For $u < v$, edges connecting vertices of C_u and C_v form a bipartite graph, based on which the parameters b_{iv} $(i \in C_u)$ and b_{ju} $(j \in C_v)$ are estimated with the above algorithm; here b_{iv}'s correspond to b_i's, b_{ju}'s correspond to g_j's, and the number of rows and columns of the rectangular array corresponding to this bipartite subgraph of $\mathbf{A}$ is $|C_u|$ and $|C_v|$, respectively. With the estimated parameters, collected in the $n \times k$ matrix $\mathbf{B}^{(t)}$, we go back to the E-step, etc.

By the general theory of the EM algorithm, since we are in exponential family, the iteration will converge. Note that here the parameter β_{iv} with $c_i = u$ embodies the affinity of vertex i of cluster C_u towards vertices of cluster C_v; and likewise, β_{ju} with $c_j = v$ embodies the affinity of vertex j of cluster C_v towards vertices of cluster C_u. By the model, this affinities are added together on the level of the log-odds. This so-called k-β model, introduced in [48], is applicable to social networks, where attitudes of individuals in the same social group (say, u) are the same toward members of another social group (say, v), though, this attitude also depends on the individual in group u. The model may also be applied to biological networks, where the clusters consist, for example, of different functioning synopses or other units of the brain, see [49].

After normalizing the β_{iv} $(i \in C_u)$ and β_{ju} $(j \in C_v)$ to meet the requirement of (16) for any $u \neq v$ pair, the sum of the parameters will be zero, and the sign and magnitude of them indicates the affinity of nodes of C_u to make ties with the nodes of C_v, and vice versa:

$$\sum_{i \in C_u} \beta_{iv} + \sum_{j \in C_v} \beta_{ju} = 0.$$

This becomes important when we want to compare the parameters corresponding to different cluster pairs. For selecting the initial number of clusters we can use considerations of [46], while for the initial clustering, spectral clustering tools of [5].

3 Biclassified blockmodels and mixtures of standardized contingency tables with continuously distributed entries

The blockmodel defined here is built of standardized random contingency table models, where the entries are independent beta-distributed with parameters depending on their row and column labels. Sufficient statistics are specified, and based on them, a convergent algorithm is introduced to find the MLE of the parameters. The model is extended to the multiclass scenario, where for fixed number of biclusters, the parameters of the beta-distributed entries also depend on their row and column cluster memberships. To find the clusters and estimate the parameters, an EM iteration for mixtures of exponential-family distributions is used. The algorithm is applicable to microarrays, and a genetic example is presented.

3.1 The model

Let $\mathbf{W} = (w_{ij})$ be an $n \times m$ contingency table of entries transformed into the (0,1) interval. Our model is the following: w_{ij} obeys beta-distribution with parameters $a_i > 0$ and $b_j > 0$. The parameters are collected in the vectors $\mathbf{a} = (a_1, \dots a_n)$ and $\mathbf{b} = (b_1, \dots b_m)$, or briefly, in $(\mathbf{a}, \mathbf{b})$. Here a_i can be thought of as the potential of row-item i to be connected to the column-items, and b_j as the resistance of column-item j to be connected to the row-items; whereas, w_{ij} is the weight of their connection.

The likelihood function is factorized as

$$\begin{aligned} L_{\mathbf{a},\mathbf{b}}(\mathbf{W}) &= \prod_{i=1}^{n}\prod_{j=1}^{m} \frac{\Gamma(a_i+b_j)}{\Gamma(a_i)\Gamma(b_j)} w_{ij}^{a_i-1}(1-w_{ij})^{b_j-1} \\ &= C(\mathbf{a},\mathbf{b}) \prod_{i=1}^{n}\prod_{j=1}^{m} \exp[(a_i-1)\ln w_{ij} + (b_j-1)\ln(1-w_{ij})] \\ &= \exp\left[\sum_{i=1}^{n}(a_i-1)\sum_{j=1}^{m}\ln w_{ij} + \sum_{j=1}^{m}(b_j-1)\sum_{i=1}^{n}\ln(1-w_{ij}) - Z(\mathbf{a},\mathbf{b})\right], \end{aligned}$$

where $C(\mathbf{a},\mathbf{b})$ is the normalizing constant, and $Z(\mathbf{a},\mathbf{b}) = -\ln C(\mathbf{a},\mathbf{b})$ is the log-partition (cumulant) function. Since the likelihood function depends on $\mathbf{W}$ only through the row-sums s_i's of the $n \times m$ matrix $\mathbf{U} = \mathbf{U}(\mathbf{W})$ of general entry $\ln w_{ij}$ and the column-sums z_j's of the $n \times m$ matrix $\mathbf{V} = \mathbf{V}(\mathbf{W})$ of general entry $\ln(1-w_{ij})$, by the Neyman–Fisher factorization theorem, the row-sums of $\mathbf{U}$ and column-sums of $\mathbf{V}$ are sufficient statistics for the parameters. In formulas,

$$s_i = \sum_{j=1}^{m}\ln w_{ij} \quad (i = 1, \dots, n); \qquad z_j = \sum_{i=1}^{n}\ln(1-w_{ij}) \quad (j = 1, \dots, m).$$

With them, the system of likelihood equations is

$$\begin{aligned} \frac{\partial \ln L_{\mathbf{a},\mathbf{b}}(\mathbf{W})}{\partial a_i} &= \sum_{s=1}^{m}\frac{\Gamma'(a_i+b_s)}{\Gamma(a_i+b_s)} - m\frac{\Gamma'(a_i)}{\Gamma(a_i)} + s_i = 0, \quad i = 1, \dots, n; \\ \frac{\partial \ln L_{\mathbf{a},\mathbf{b}}(\mathbf{W})}{\partial b_i} &= \sum_{s=1}^{n}\frac{\Gamma'(a_s+b_i)}{\Gamma(a_s+b_i)} - n\frac{\Gamma'(b_i)}{\Gamma(b_i)} + z_i = 0, \quad i = 1, \dots, m. \end{aligned} \tag{19}$$

The theory of the exponential families guarantees that the system (19) has a unique solution (MLE) if the sufficient statistic is an inner point of a closed manifold (convex hull of all possible sufficient statistics). But in case of absolutely continuous distributions, so in the present situation, it happens with probability 1. Let $\hat{\underline{\theta}}$ denote this unique (with probability 1) MLE, where $\underline{\theta}$ is the shorthand for the parameter pair $(\mathbf{a}, \mathbf{b})$ to be estimated.

In practice, we use a fixed point iteration, for which purpose we rewrite the system (19) in the form $\underline{\theta} = f(\underline{\theta})$ as follows:

$$\begin{aligned} a_i &= \psi^{-1}\left[\frac{1}{m}s_i + \frac{1}{m}\sum_{s=1}^{m}\psi(a_i+b_s)\right] =: g_i(\mathbf{a},\mathbf{b}), \quad i = 1, \dots, n; \\ b_i &= \psi^{-1}\left[\frac{1}{n}z_i + \frac{1}{n}\sum_{s=1}^{n}\psi(a_s+b_i)\right] =: h_i(\mathbf{a},\mathbf{b}), \quad i = 1, \dots, m. \end{aligned} \tag{20}$$

Here $\psi(x) = \frac{d \ln \Gamma(x)}{dx} = \frac{\Gamma'(x)}{\Gamma(x)}$ for $x > 0$ is the *digamma function*; further, f is the shorthand for the function pair (g, h), $g : \mathbb{R}^{n+m} \to \mathbb{R}^n$ and $h : \mathbb{R}^{n+m} \to \mathbb{R}^m$ with coordinate functions g_i's and h_j's, as in Equation (20), respectively.

Then, starting with $\underline{\theta}^{(0)}$, we use the successive approximation $\underline{\theta}^{(t)} := f(\underline{\theta}^{(t-1)})$ for $t = 1, 2, \ldots$ until convergence. As for the starting, let

$$M := \max \left\{ \max_{i \in \{1,\ldots,n\}} \left(-\frac{s_i}{n} \right), \max_{i \in \{1,\ldots,m\}} \left(-\frac{z_i}{m} \right) \right\}$$

and $\varepsilon > 0$ be the solution of the equation $\psi(2x) - \psi(x) = M$ (it is unique as the function $\psi(2x) - \psi(x)$, $x \in (0, \infty)$ is strictly decreasing). With this ε, the convergence of the above iteration follows from the fact that the sequence $\underline{\theta}^{(t)}$ is coordinate-wise increasing and is bounded from above by $\hat{\underline{\theta}}$. By continuity of f, the limit is clearly the fixed point of f. As the MLE is the solution of the equivalent system (20) of the maximum likelihood equations, this fixed point cannot be else but the uniquely existing $\hat{\underline{\theta}}$. This closed neighborhood K of $\hat{\underline{\theta}}$ is only theoretically guaranteed. However, the vector $\varepsilon \mathbf{1} \in \mathbb{R}^{n+m}$ can be a good starting. Indeed, starting with it, an iterate sooner or later gets into K. From that point, the iteration speeds up, and converges at a geometric rate.

We also applied the algorithm to migration data between 34 countries. Here w_{ij} is proportional to the number of people in thousands who moved from country i to country j (to find jobs) during the year 2011, and it is normalized so that be in the interval (0,1). The estimated parameters are in Table 1.

In this context, a_i's are related to the emigration and and b_i's to the counter-immigration potentials. When a_i is large, country i has a relatively large potential for emigration. On the contrary, when b_i is large, country i tends to have a relatively large resistance against immigration.

i	Country	a_i	b_i	i	Country	a_i	b_i
1	Australia	0.26931	1475.75242	18	Japan	0.23211	9926.91644
2	Austria	0.27403	632.81653	19	Korea	0.22310	4199.25005
3	Belgium	0.33380	46.01197	20	Luxembourg	0.17543	107.91399
4	Canada	0.27383	2363.23435	21	Mexico	0.26706	4655.95370
5	Chile	0.21236	28940.59777	22	Netherlands	0.37754	39.52320
6	Czech Rep.	0.31188	470.28651	23	New Zealand	0.20542	2568.00582
7	Denmark	0.26514	847.34887	24	Norway	0.22646	519.12451
8	Estonia	0.23235	25602.33371	25	Poland	0.62846	1106.55946
9	Finland	0.29357	1100.00568	26	Portugal	0.31011	1606.59979
10	France	0.52721	37.92122	27	Slovak Rep.	0.27871	42451.19093
11	Germany	0.62020	1.64064	28	Slovenia	0.19720	6824.54028
12	Greece	0.29708	6319.19184	29	Spain	0.39732	182.47160
13	Hungary	0.31443	32750.88310	30	Sweden	0.39627	57.34509
14	Iceland	0.18051	2950.72653	31	Switzerland	0.33611	4524.67821
15	Ireland	0.27555	364.52781	32	Turkey	0.25900	146175.82805
16	Israel	0.25854	1926.04551	33	United Kingdom	0.49301	48.61626
17	Italy	0.50522	135.14076	34	United States	0.38019	2433.78269

Table 1: Estimated parameters for migration data, 2011

It should be noted again that edge-weighted graphs of this type very frequently model real-world directed networks.

3.2 The multiclass contingency table model

In the several clusters case, we are putting blocks of Section 3.1 together. Here the statistics are sufficient only within the blocks. Given the integers $1 \leq k \leq n$

and $1 \leq l \leq m$, we are looking for k-partition, in other words, clusters $R_1, \ldots, R_k$ of the rows and $C_1, \ldots, C_l$ of the columns such that the row and column items are assigned to the clusters independently, and given the cluster memberships, the weight of the connection of row-item $u \in R_i$ to column-item $v \in C_j$ is $w_{uv} \sim \mathcal{B}eta(a_{uj}, b_{vi})$; further, all these assignments are done independently.

The parameters are stored in the $n \times l$ matrix $\mathbf{A}$ and the $m \times k$ matrix $\mathbf{B}$, where the jth column of $\mathbf{A}$ contains the parameters a_{uj} in the block $u \in R_i$, for $j = 1, \ldots, l$; $i = 1, \ldots, k$. Likewise, the ith column of $\mathbf{B}$ contains the parameters b_{vi} in the block $v \in C_j$, for $i = 1, \ldots, k$; $j = 1, \ldots, l$. Here a_{uj} can be thought of as the potential of row-item u of cluster R_i to be connected to C_j, and b_{vi} as the potential of column-item v of cluster C_j to be connected to R_i.

This is a mixture of exponential-family distributions, and as the mixing can be supervised by two multinomially distributed random variables (responsible for the memberships), the general theory of mixtures, and the iteration of the EM algorithm can be used to estimate the parameters. With the terminology of the EM algorithm, $\mathbf{W}$ is the incomplete data specification. If the missing memberships were known, we would be able to write the complete log-likelihood in the following form:

$$\sum_{i=1}^{k} \sum_{j=1}^{l} \sum_{u \in R_i} \sum_{v \in C_j} \left[\ln \frac{\Gamma(a_{uj} + b_{vi})}{\Gamma(a_{uj})\Gamma(b_{vi})} + (a_{uj} - 1) \ln w_{uv} + (b_{vi} - 1) \ln(1 - w_{uv}) \right]. \tag{21}$$

Starting with an initial clustering $R_1^{(0)}, \ldots, R_k^{(0)}$ of the rows and $C_1^{(0)}, \ldots, C_l^{(0)}$ of the columns, the t-th step of the iteration is as follows ($t = 1, 2, \ldots$).

- **Maximization step within the blocks**: We update estimates of the parameters $\mathbf{A}^{(t)}$, $\mathbf{B}^{(t)}$ within the kl blocks, separately. As for the block $R_i^{(t)} \times C_j^{(t)}$, we use the algorithm of Section 3.1 to find the estimates $a_{uj}^{(t)}$ for $u \in R_i^{(t)}$ and $b_{vi}^{(t)}$ for $v \in C_j^{(t)}$. As each row u and column v uniquely corresponds to exactly one row- and column-cluster, respectively, in this way, the parameter blocks, estimated from $R_i^{(t)} \times C_j^{(t)}$, for $i = 1, \ldots, k$, $j = 1, \ldots, l$ will fill in the $\mathbf{A}^{(t)}$, $\mathbf{B}^{(t)}$ parameter matrices.

- **Relocation step between the blocks**: Given the new estimates of the parameters $\mathbf{A}^{(t)}$, $\mathbf{B}^{(t)}$, we relocate u into the row-cluster R_{i^*} and v into the column-cluster C_{j^*} for which the contribution of w_{uv} to the overall likelihood (21) is maximal. We do that separately for the rows and columns. For this purpose, we write the overall likelihood in terms of membership vectors. Let the $n \times k$ matrix $\mathbf{R} = (r_{ui})$ contain the membership vectors of the rows, i.e., $r_{ui} = 1$ if $u \in R_i$ and $r_{ui'} = 0$ for $i' \neq i$. Likewise, let the $m \times l$ matrix $\mathbf{C} = (c_{vj})$ contain the membership vectors of the columns, i.e., $c_{vj} = 1$ if $v \in C_j$ and $c_{vj'} = 0$ for $j' \neq j$.

  - *Relocation of the rows*: for each u ($u = 1, \ldots n$), take the maximum of the following over i ($i = 1, \ldots, k$):

$$\sum_{v=1}^{m} \sum_{j=1}^{l} c_{vj} \left[\ln \frac{\Gamma(a_{uj}) + b_{vi})}{\Gamma(a_{uj})\Gamma(b_{vi})} + (a_{uj} - 1) \ln w_{uv} + (b_{vi} - 1) \ln(1 - w_{uv}) \right]. \tag{22}$$

If it is maximum for i^*, then we relocate u into the row-cluster R_{i^*}. This is a discrete maximization. Break ties arbitrarily.

– *Relocation of the columns*: for each v ($v = 1, \dots m$), take the maximum of the following over j ($j = 1, \dots, l$):

$$\sum_{u=1}^{n}\sum_{i=1}^{k} r_{ui}\left[\ln\frac{\Gamma(a_{uj})+b_{vi})}{\Gamma(a_{uj})\Gamma(b_{vi})} + (a_{uj}-1)\ln w_{uv} + (b_{vi}-1)\ln(1-w_{uv})\right]. \tag{23}$$

If it is maximum for j^*, then we relocate v into the column-cluster C_{j^*}. This is a discrete maximization. Break ties arbitrarily.

In this way, we get a new clustering $R_1^{(t)}, \dots, R_k^{(t)}$ of the rows and $C_1^{(t)}, \dots, C_l^{(t)}$ of the columns, with which we go back to the maximization step.

As in both steps we increase the likelihood, and the likelihood function is bounded from above with the sum of the existing maxima over the blocks, the iteration must converge to a local maximum of it. A good starting, for example, with spectral biclustering helps a lot.

In fact, the relocation corresponds to the E-step of the classical EM algorithm . Indeed, given $\mathbf{W}$, k, and l, the complete log-likelihood is

$$\sum_{i=1}^{k}\sum_{j=1}^{l}\sum_{u=1}^{n}\sum_{v=1}^{m} r_{ui}c_{vj}\left[\ln\frac{\Gamma(a_{uj})+b_{vi})}{\Gamma(a_{uj})\Gamma(b_{vi})} + (a_{uj}-1)\ln w_{uv} + (b_{vi}-1)\ln(1-w_{uv})\right]. \tag{24}$$

If we fix i, j, then we maximize the inner double summation, i.e., we find the ML estimate of the parameters in the $R_i \times C_j$ block (M-step) as in (21). Reordering the summation as

$$\sum_{u=1}^{n}\sum_{i=1}^{k} r_{ui}\left[\sum_{v=1}^{m}\sum_{j=1}^{l} c_{vj}\left(\ln\frac{\Gamma(a_{uj})+b_{vi})}{\Gamma(a_{uj})\Gamma(b_{vi})} + (a_{uj}-1)\ln w_{uv} + (b_{vi}-1)\ln(1-w_{uv})\right)\right], \tag{25}$$

for any fixed row u, we maximize the inner double summation, which is in the brackets and is identical to (22), for $i = 1, \dots, k$. It is equivalent to maximizing

$$\mathbb{E}(r_{ui} \,|\, \mathcal{M}^{(t-1)}) = \mathbb{P}(r_{ui} = 1 \,|\, \mathcal{M}^{(t-1)})$$

with respect to i, where $\mathcal{M}^{(t-1)}$ contains the model parameters and the cluster assignments after step $t-1$. By the Bayes rule it is equivalent to maximizing

$$\mathbb{P}(\mathcal{M}^{(t-1)} \,|\, r_{ui} = 1)\mathbb{P}(r_{ui} = 1)$$

with respect to i. But, under uniform law of prior memberships, by the Bayes rule, it is obtained by maximizing $\mathbb{P}(\mathcal{M}^{(t-1)} \,|\, r_{ui} = 1)$ for $i = 1, \dots, k$, that is (25). If it is maximum for i^*, then we relocate u into the row-cluster R_{i^*}. It is also possible to maximize

$$\mathbb{P}(\mathcal{M}^{(t-1)} \,|\, r_{ui} = 1) \cdot \frac{n_i^{(t-1)}}{n},$$

where $n_i^{(t-1)}$ is the number of rows in the cluster R_i after the $(t-1)$th iteration.

Then we do the relocation for the columns:

$$\sum_{v=1}^{m}\sum_{j=1}^{l} c_{vj}\left[\sum_{u=1}^{n}\sum_{i=1}^{k} r_{ui}\left(\ln\frac{\Gamma(a_{uj})+b_{vi})}{\Gamma(a_{uj})\Gamma(b_{vi})}+(a_{uj}-1)\ln w_{uv}+(b_{vi}-1)\ln(1-w_{uv})\right)\right]. \tag{26}$$

For any fixed column v, we maximize the inner double summation, which is in the brackets and is identical to (23), for $j = 1, \ldots, l$. If it is maximum for j^*, then we relocate v into the column-cluster C_{j^*}. It is also equivalent to maximizing

$$\mathbb{E}(c_{vj} \,|\, \mathcal{M}^{(t-1)}) = \mathbb{P}(c_{vj} = 1 \,|\, \mathcal{M}^{(t-1)})$$

with respect to j. Actually, the last maximization already corresponds to the M-step. Note that similar idea appears in papers related to the collaborative filtering.

References

[1] Clauset,A., Newman,M.E.J., and Moore,C., Finding community structure in very large networks, *Physical Review E* **70**, 066111 (2004).

[2] Newman,M.E.J. *Networks, An Introduction.* Oxford University Press (2010).

[3] Fortunato,S. Community detection in graphs. *Phys. Rep.* **486**, 75–174 (2010).

[4] Bolla,M., Penalized versions of the Newman–Girvan modularity and their relation to multiway cuts and k-means clustering, *Physical Review E* **84**, 016108 (2011).

[5] Bolla, M., *Spectral Clustering and Biclustering. Learning Large Graphs and Contingency Tables.* Wiley (2013).

[6] Rao,C.R. *Linear Statistical Inference and its Applications.* Wiley (1973).

[7] McLachlan,G.J., *The EM Algorithm and Extensions.* Wiley (1997).

[8] Chatterjee,S., Diaconis,P. and Sly, A., Random graphs with a given degree sequence, *Ann. Statist.* **21**, 1400-1435 (2010).

[9] Csiszár,V., Hussami,P., Komlós,J., Móri,T.F., Rejtő,L. and Tusnády,G., When the degree sequence is a sufficient statistic, *Acta Math. Hung.* **134**, 45-53 (2011).

[10] Rinaldo,A., Petrovic,S. and Fienberg,S.E., Maximum likelihood estimation in the β-model, *Ann. Statist.* **41**, 1085-1110 (2013).

[11] Holland,P.W., Laskey,K.B. and Leinhardt,S., Stochastic blockmodels: some first steps, *Social Networks* **5**, 109-137 (1983).

[12] Rohe,K., Chatterjee,S. and Yu,B., Spectral clustering and the high-dimensional stochastic blockmodel, *Ann. Statist.* **39** (4), 1878–1915 (2011).

[13] Karrer,B. and Newman,M.E.J., Stochastic blockmodels and community structure in networks, *Phys. Rev. E* **83**, 016107 (2011).

[14] Choi,D.S„ Wolfe,P.J. and Airoldi,E.M., Stochastic blockmodels with growing number of classes, *Biometrika* **99** (2), 273–284 (2012).

[15] Fishkind,D.E., Sussman,D.L., Tang,M., Vogelstein,J.T. and Priebe,C.E., Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown, *Siam J. Matrix Anal. Appl.* **34** (1), 23-39 (2013).

[16] Rasch,G.,*Studies in Mathematical Psychology: I. Probabilistic Models for Some Intelligence and Attainment Tests.* Nielsen and Lydiche, Oxford, UK (1960).

[17] Rasch,G., On general laws and the meaning of measurement in psychology. In *Proc. of the Fourth Berkeley Symp. on Math. Statist. and Probab.*, pp. 321-333, University of California Press (1961).

[18] Dempster,A.P., Laird,N.M. and Rubin,D.B., Maximum likelihood from incomplete data via the EM algorithm, *J. R. Statist. Soc. B* **39**, 1-38 (1977).

[19] Ungar,L.H. and Foster,D.P., A Formal Statistical Approach to Collaborative Filtering. In *Proc. Conference on Automatical Learning and Discovery (CONALD 98)* (1998).

[20] Hofmann,T. and Puzicha,J., Latent class models for collaborative filtering. In *Proc. 16th International Joint Congress on Artificial Intelligence (IJCAI 99)* (ed. Dean T), Vol. 2, pp. 688-693. Morgan Kaufmann Publications Inc., San Francisco CA (1999).

[21] Casella,G. and George,E.I., Explaining the Gibbs sampler, *The American Statistician* **46**, 167–174 (1992).

[22] Metropolis,N., Rosenblut,A., Rosenbluth,M., Teller,A. and Teller,E., Equation of state calculation by fast computing machines, *J. Chem. Physics* **21**, 1087–1092 (1953).

[23] Holland,P.W. and Leinhardt,S., An exponential family of probability distributions for directed graphs, *J. Amer. Statist. Assoc.* **76**, 33-50 (1981).

[24] Lauritzen,S.L., *Extremal families and systems of sufficient statistics.* Lecture Notes in Statistics **49**, Springer (1988).

[25] Daudin,J-J., Picard,F. and Robin,S., A mixture model for random graphs, *Statistics and Computing* **18**, 173-183 (2008).

[26] Newman,M.E.J., Analysis of weighted networks, *Physical Review E* **70**, 056131 (2004).

[27] Newman,M.E.J., Community detection and graph partitioning, *Europhysics Letters* **103**, 28003 (2013).

[28] Newman,M.E.J., Mixing patterns in networks, *Physical Review E* **67**, 026126 (2003).

[29] Bickel,P.J. and Chen,A., A nonparametric view of network models and Newman-Girvan and other modularities, *Proc. Natl. Acad. Sci. USA* **106** (50), 21068-21073 (2009).

[30] Reichardt,J. and Bornholdt,S., Partitioning and modularity of graphs with arbitrary degree distribution, *Physical Review E* **76**, 015102(R) (2007).

[31] Escolano,F., Hancock,E.R. and Lozano,M.A., Heat diffusion: Thermodynamic depth complexity of networks, *Physical Review E* **85**, 036206 (2012).

[32] Erdős,P. and Rényi,A., On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 17-61 (1960).

[33] Erdős,P. and Gallai, T., Graphs with given degree of vertices (in Hungarian), *Matematikai Lapok* **11**, 264-274 (1960).

[34] Mahadev,N.V.R. and Peled,U.N., Threshold graphs and related topics, *Ann. Discrete. Math.* **56**, North-Holland, Amsterdam (1995).

[35] Stanley, L. P., A zonotope associated with graphical degree sequences. In *Applied Geometry and Discrete Mathematics. DIMACS Series in Discrete Mathematics and Theoretical Computer Science* **4**, pp. 555-570. Amer. Math. Soc., Providence, RI. (1991).

[36] Hammer,P.L., Peled,U.N. and Sun,X., Difference graphs, *Discrete Applied Mathematics* **28**, 35-44 (1990).

[37] Barvinok,A., What does a random contingency table look like? preprint, arXiv:0806.3910 [math.CO] (2009).

[38] Borgs,C., Chayes,J.T., Lovász,L, T.-Sós,V. and Vesztergombi,K., Convergent sequences of dense graphs II: Multiway cuts and statistical physics. *Ann. Math.* **176**, 151-219 (2012).

[39] Haberman,S.J., Log-linear models and frequency tables with small expected counts, *Ann. Statist.* **5**, 1148-1169 (1977).

[40] Barvinok,A., Matrices with prescribed row and column sums, preprint, arXiv:1010.5706 [math.CO] (2010).

[41] Gale,D., A theorem on flows in networks, *Pacific J. Math.* **7**, 1073-1082 (1957).

[42] Ryser,H.J., Combinatorial properties of matrices of zeros and ones, *Canad. J. Math.* **9**, 371-377 (1957).

[43] Ford,L.R. and Fulkerson,D.R., Maximal flow through a network, *Canad. J. Math.* **8**, 399-404 (1956).

[44] Barvinok,A., Hartigan,J.A., An asymptotic formula for the number of nonnegative integer matrices with prescribed row and column sums, preprint, arXiv:0910.2477 [math.CO] (2009).

[45] Barvinok,A., On the number of matrices and a random matrix with prescribed row and column sums and 0-1 entries, preprint, arXiv:0806.1480 [math.CO] (2009).

[46] Yan,D., Chen,A. and Jordan,M.I., Cluster forests, *Comput. Statist. and Data Anal.* **66**, 178–192 (2013).

[47] Hastie,T., Tibshirani,R. and Friedman,J., *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* Springer (2001).

[48] Csiszár,V., Hussami,P., Komlós,J., Móri,T.F., Rejtő,L. and Tusnády,G., Testing goodness of fit of random graph models, *Algorithms* **5**, 629-635 (2012).

[49] Négyessy,L., Nepusz,T., Zalányi,L. and Bazsó,F., Convergence and divergence are mostly reciprocated properties of the connections in the network of cortical areas, *Proc. R. Soc. B: Biological Sciences* **275**, 2403-2410 (2008).

[50] Bernard,H.R., Killworth,P.D. and Sayler,L., Informant accuracy in social-network data V. An experimental attempt to predict actual communication from recall data, *Social Science Research* **11**, 30-66 (1982).

[51] Pálovics, R., Benczúr, A., Kocsis, L., Kiss, T. and Frigó, E., Exploiting temporal influence in online recommendation, In Proc. of RecSys'14, 8th ACM Conference on Recommender Systems pp. 273-280, ACM, New York, NY, USA (2014).

[52] Liu,Y-Y., Slotine,J-J. and Barabási,A-L., Controllability of complex networks, *Nature* **473**, 167-173 (2011).