Technical documentation about estimation in the ERMG model

V. Miele, F. Picard, J-J. Daudin, M. Mariadassou, S. Robin

August 3, 2007

$\mathbf{X}$ is the adjacency matrix of size $n \times n$ defined such that $X_{ij} = 1$ if nodes i and j are connected. $\mathbf{Z}$ is defined such that $\{Z_{iq} = 1\}$ if node i belongs to class q. We distinguish formulas for graphs with undirected or directed vertices ($X_{ij} = X_{ji}$ or $X_{ij} \neq X_{ji}$), and formulas for graphs with or without self loops ($X_{ii} \neq 0$ or $X_{ii} = 0$).

1 Initialization with Hierarchical clustering

1.1 Distances

1.1.1 Undirected graphs

Distances between vertices. This distance represents the number of discordances between vertices i and j.

$$d(i,j) = \sum_k (x_{ik} - x_{jk})^2 = \|x_i - x_j\|^2. \tag{1}$$

Distance between groups. Denoting g_q the barycenter of group q defined such that

$$\forall i \in \{1,\dots,n\},\ g_{qi} = \frac{\sum_{k \in q} x_{ki}}{n_q},$$

and $n_q = \#(k \in q)$, we define the following distance between groups:

$$\Delta(q,\ell) = \frac{n_q n_\ell}{n_q + n_\ell} \|g_q - g_\ell\|^2. \tag{2}$$

This distance is the classical Ward distance between groups.

1.1.2 Directed graphs

Distances between vertices.

$$\begin{aligned} d(i,j) &= \sum_k (x_{ik} - x_{jk})^2 + \sum_k (x_{ki} - x_{kj})^2 \\ &= d^+(i,j) + d^-(i,j) \end{aligned}$$

Distance between groups. Denoting (g_q^+, g_q^-) the barycenters of group q for rows and columns, defined such that

$$\forall i \in \{1,\dots,n\}, \begin{cases} g_{qi}^+ = \frac{\sum_{k\in q} x_{ik}}{n_q}, \\ g_{qi}^- = \frac{\sum_{k\in q} x_{ki}}{n_q}, \end{cases} \tag{3}$$

and $n_q = \#(k \in q)$. Similarly we define the following distance between groups:

$$\Delta(q,\ell) \quad = \quad \frac{n_q n_\ell}{n_q + n_\ell}\left(\|g_q^+ - g_\ell^+\|^2 + \|g_q^- - g_\ell^-\|^2\right).$$

1.2 k-means algorithm

In order to reduce the computational burden for the hierarchical clustering step, we reduce the dimension of the dataset using a k-means algorithm, the number of centers being fixed by the user. This number is denoted N_{max}.

1. Choose N_{max} regularly ordered centers at random $(\mathrm{mod}(n/N_{max}))$,
2. Calculate the distances of vertices with the centers,
3. Cluster vertices to the nearest center. If *ex-aequo*, centers are chosen randomly,
4. Iterate (1)-(2)-(3) until no change of centers.

At the end of the k-means step, matrix $\mathbf{Z}$ is filled with the class for each node.

1.3 Hierarchical clustering algorithm

1. Initialization: calculate Δ the distance between the N_{max} groups defined by the k-means step.
2. Merging step: two groups are merged if their distance Δ is the smallest. If two distances are equal, groups to merge are randomly chosen. The label of the new formed group is the smallest of the two previous label.
3. Calculate distance between groups,
4. Iterate (1)-(2)-(3) until the number of classes equals 1.

2 Variational algorithm

Definitions. (t) is the current index for iterations, Q the number of classes, τ the matrix of *posterior* probabilities (n, Q) defined such that:

$$\tau_{iq} = \Pr\{Z_{iq} = 1|\mathbf{X}\} \tag{4}$$

where $Z_{iq} = 1$ if $i \in class(q)$

$$\forall i \sum_{q=1,Q} Z_{iq} = 1 \tag{5}$$

and

$$\forall i \sum_{q=1,Q} \tau_{iq} = 1 \tag{6}$$

Note. In the following for the undirected case, $\sum_{i<j}$ is equivalent to $\frac{1}{2}\sum_{i\neq j}$.

2.1 M-step

In any case, we have $\alpha_q^{(t)} = \sum_i \tau_{iq}^{(t)}/n$. The estimator for parameter π_{ql} is such that:

Undirected without self loop:

$$\pi_{ql}^{(t)} = \frac{\sum_{i<j} \tau_{iq}^{(t)} x_{ij} \tau_{jl}^{(t)}}{\sum_{i<j} \tau_{iq}^{(t)} \tau_{jl}^{(t)}}$$

Directed without self loop:

$$\pi_{ql}^{(t)} = \frac{\sum_{i\neq j} \tau_{iq}^{(t)} x_{ij} \tau_{jl}^{(t)}}{\sum_{i\neq j} \tau_{iq}^{(t)} \tau_{jl}^{(t)}}$$

Undirected with self loops:

$$\pi_{ql}^{(t)} = \begin{cases} \text{if } q \neq l & \frac{\sum_{i<j} \tau_{iq}^{(t)} x_{ij} \tau_{jl}^{(t)}}{\sum_{i<j} \tau_{iq}^{(t)} \tau_{jl}^{(t)}} \\ \\ \text{otherwise} & \frac{\sum_i \tau_{iq}^{(t)} (\sum_{j<i} x_{ij} \tau_{jl}^{(t)} + x_{ii})}{\sum_i \tau_{iq}^{(t)} (\sum_{j<i} \tau_{jl}^{(t)} + 1)} \end{cases}$$

Directed with self loops:

$$\pi_{ql}^{(t)} = \begin{cases} \text{if } q \neq l & \frac{\sum_{i\neq j} \tau_{iq}^{(t)} x_{ij} \tau_{jl}^{(t)}}{\sum_{i\neq j} \tau_{iq}^{(t)} \tau_{jl}^{(t)}} \\ \\ \text{otherwise} & \frac{\sum_i \tau_{iq}^{(t)} (\sum_{j\neq i} x_{ij} \tau_{jl}^{(t)} + x_{ii})}{\sum_i \tau_{iq}^{(t)} (\sum_{j\neq i} \tau_{jl}^{(t)} + 1)} \end{cases}$$

- α_qs are bounded at ϵ_α such that no empty class is created.
- π_{ql} is left and right bounded with ϵ_π and $(1 - \epsilon_\pi)$.
- if the denominator $\to 0$, π_{ql} is set to 0.5. This configuration corresponds to the case where one class tends to contain only one node.

2.2 E-step

We define $\beta_{ijql}^{(t)}$, such that:

$$\beta_{ijql}^{(t)} = x_{ij} \ln(\pi_{ql}^{(t)}) + (1 - x_{ij}) \ln(1 - \pi_{ql}^{(t)}).$$

Note that π_{ql} is bounded in the M-step. Posterior probabilities are calculated using a fixed point algorithm. Let (h) denote the current index for iterations.

Undirected case without self loop:

$$\log \tau_{iq}^{(h+1)} = \log \alpha_q^{(t)} + \sum_{j<i} \sum_{l=1,Q} \tau_{jl}^{(h+1)} \beta_{ijql}^{(t)} + \sum_{j>i} \sum_{l=1,Q} \tau_{jl}^{(h)} \beta_{ijql}^{(t)}, \tag{7}$$

Directed case without self-loop:

$$\log \tau_{iq}^{(h+1)} = \log \alpha_q^{(t)} + \sum_{j<i} \sum_{l=1,Q} \tau_{jl}^{(h+1)} (\beta_{ijql}^{(t)} + \beta_{jilq}^{(t)}) + \sum_{j>i} \sum_{l=1,Q} \tau_{jl}^{(h)} (\beta_{ijql}^{(t)} + \beta_{jilq}^{(t)}), \tag{8}$$

Undirected case with self loops:

$$\log \tau_{iq}^{(h+1)} = \log \alpha_q^{(t)} + \sum_{j<i} \sum_{l=1,Q} \tau_{jl}^{(h+1)} \beta_{ijql}^{(t)} + \sum_{j>i} \sum_{l=1,Q} \tau_{jl}^{(h)} \beta_{ijql}^{(t)} + \beta_{iiqq}^{(t)}, \tag{9}$$

Directed case with self-loops:

$$\log \tau_{iq}^{(h+1)} = \log \alpha_q^{(t)} + \sum_{j<i} \sum_{l=1,Q} \tau_{jl}^{(h+1)} \left(\beta_{ijql}^{(t)} + \beta_{jilq}^{(t)}\right) + + \sum_{j>i} \sum_{l=1,Q} \tau_{jl}^{(h)} \left(\beta_{ijql}^{(t)} + \beta_{jilq}^{(t)}\right) + \beta_{iiqq}^{(t)}, \tag{10}$$

In any case, $\tau_{iq}s$ are normalized such that:

$$\tau_{iq} = \frac{\tau_{iq}}{\sum_l \tau_{il}}.$$

- $\tau_{iq}s$ are bounded such that $\epsilon_\tau < \tau_{iq} < 1 - \epsilon_\tau$,
- A factorization is used to avoid numerical zeros in the calculus of *posterior* probabilities. Considering that $\tau_{iq} \propto \exp(-\delta_{iq})$, and denoting $\delta_i^\star = \max_q \delta_{iq}$, τ_{iq} is calculated such that:

$$\tau_{iq} \propto \frac{e^{-(\delta_{iq} - \delta_i^\star)}}{\sum_l e^{-(\delta_{il} - \delta_i^\star)}}$$

- the stopping rule is :

$$\begin{cases} \max_{iq} |\tau_{iq}^{(h+1)} - \tau_{iq}^{(h)}| \leq \delta_\tau \\ h \geq h_{max} \end{cases}$$

2.3 Stopping rule and Likelihoods.

Stopping rule on parameters. Denoting $\theta = (\alpha, \pi)$, the EM algorithm stops when

$$\begin{cases} \max |(\theta^{(t+1)} - \theta^{(t)})/\theta^{(t)}| \leq \delta_\theta \\ t \geq t_{max} \end{cases} \tag{11}$$

Incomplete-data log-likelihood approximation.

$$J_Q = \mathcal{Q}_Q - \mathcal{H}_Q$$

Complete-data log-likelihood.

Undirected case without self loop:

$$\mathcal{Q}_Q = \sum_i \sum_q \tau_{iq} \log \alpha_q + \sum_i \sum_{j<i} \sum_{q,l} \tau_{iq}\tau_{jl}\beta_{ijql}$$

Directed case without self-loop:

$$\mathcal{Q}_Q = \sum_i \tau_{iq} \log \alpha_q + \sum_i \sum_{j<i} \sum_{q,l} \tau_{iq}\tau_{jl}(\beta_{ijql} + \beta_{jilq}) + \sum_i \sum_{j>i} \sum_{q,l} \tau_{iq}\tau_{jl}(\beta_{ijql} + \beta_{jilq}),$$

Undirected case with self loops:

$$\mathcal{Q}_Q = \sum_i \sum_q \tau_{iq} \log \alpha_q + \sum_i \sum_{j<i} \sum_{q,l} \tau_{iq}\tau_{jl}\beta_{ijql} + \sum_{i,q} \tau_{iq}\beta_{iiqq},$$

Directed case with self-loops:

$$\mathcal{Q}_Q = \sum_i \sum_q \tau_{iq} \log \alpha_q + \sum_i \sum_{j<i} \sum_{q,l} \tau_{iq}\tau_{jl}\left(\beta_{ijql} + \beta_{jilq}\right) + \sum_i \sum_{j>i} \sum_{q,l} \tau_{iq}\tau_{jl}\left(\beta_{ijql} + \beta_{jilq}\right) + \sum_{i,q} \tau_{iq}\beta_{iiqq},$$

Entropy.

$$\mathcal{H}_Q = \sum_i \sum_q \tau_{iq} \log \tau_{iq}$$

BIC.

Undirected case:

$$BIC_Q = J_Q - \frac{Q(Q+1)}{4} \log \frac{n(n-1)}{2} - \frac{(Q-1)}{2} \log n$$

Directed case:

$$BIC_Q = J_Q - \frac{Q^2}{2} \log n^2 - \frac{(Q-1)}{2} \log n$$

ICL.

Undirected case:

$$ICL_Q = \mathcal{Q}_Q - \frac{Q(Q+1)}{4} \log \frac{n(n-1)}{2} - \frac{(Q-1)}{2} \log n$$

Directed case:

$$ICL_Q = \mathcal{Q}_Q - \frac{Q^2}{2} \log n^2 - \frac{(Q-1)}{2} \log n$$