Technical documentation about online estimation in the ERMG model

V. Miele, C.Ambroise, F. Picard, H.Zanghi

September 13, 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 vertices ($X_{ij} = X_{ji}$) for graphs with or without self loops ($X_{ii} \neq 0$ or $X_{ii} = 0$).

1 Initialization with Hierarchical clustering on a randomly-drawn subgraph

1.1 Distance

We use the classical Ward distance between groups.

1.1.1 Undirected graphs

Distance 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:

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

1.1.2 Directed graphs

Distance 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

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

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 Hierarchical clustering algorithm

We perform the hierarchical clustering step on a randomly-drawn subgraph to reduce the computational burden.

1. Shuffle the vertices,
2. Build the adjacency matrix X_O of size $n_0 \times n_0$ (i.e. the subgraph) from the edges connecting the n_0 first vertices, with $n_0 = \min(\max(n/3, 200), n)$ or a user-specified n_0.

then

1. Initialization: calculate Δ the distance between the n_0 vertices considered as groups.
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 Online Classification algorithm

Définitions. We begin at $(m) = n_0$ with (m) the current index for iterations. Q the number of classes, $(n_q)_{1\leq q\leq Q}$ and $(n_{ql})_{1\leq q,l\leq Q}$ such that:

- $n_{ql} = \sum_{i>j} x_{ij} z_{iq} z_{jl}$, the number of egdes having nodes in class q and l,
- $n_q = \sum_i z_{iq}$, the number of nodes of class q.

We define $i(m) = mod(m-1, n)$.

2.1 M-step

If $m \leq n$

$$n_q^{(m)} = n_q^{(m-1)} + z_{mq}^{(m)}, \quad (4)$$

$$n_{ql}^{(m)} = n_{ql}^{(m-1)} + \sum_{j \neq m} z_{mq}^{(m)} z_{jl}^{(m)} x_{mj}, \quad (5)$$

else

$$n_q^{(m)} = n_q^{(m-1)} - z_{i(m)q}^{(m-1)} + z_{i(m)q}^{(m)}, \quad (6)$$

$$n_{ql}^{(m)} = n_{ql}^{(m-1)} + (z_{i(m)q}^{(m)} - z_{i(m)q}^{(m-1)}) \sum_{j \neq i(m)} z_{jl^{(m)}} x_{i(m)j}, \quad (7)$$

In any case, we have $\alpha_q^{(m)} = \frac{n_q^{(m)}}{min(n,m)}$ and the estimator for parameter π_{ql} is such that

$$\pi_{ql}^{(m)} = \frac{n_{ql}^{(m)}}{n_q^{(m)} n_l^{(m)}}, \quad (8)$$

$$(9)$$

Without self loop:

$$\pi_{qq}^{(m)} = \frac{n_{qq}^{(m)}}{\frac{1}{2} * n_q^{(m)} (n_q^{(m)} - 1)}. \quad (10)$$

With self loop:

$$\pi_{qq}^{(m)} = \frac{n_{qq}^{(m)}}{\frac{1}{2} * n_q^{(m)} (n_q^{(m)} - 1)) + n_q}. \quad (11)$$

- 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 $(1 - \epsilon_\pi)$.

2.2 E-step

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

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

Note that π_{ql} is bounded in the M-step.
We recall $i(m) = mod(m - 1, n)$.

At step (m), assign the node $i(m)$ to the class q^* such that $q^* = \arg\max_q L_q$ where:

Without self loop:
If $m \leq n$

$$L_q = \log \alpha_q^{(m-1)} + \sum_{l=1,Q} \sum_{j<m} z_{jl}^{(m-1)} \beta_{mjql}^{(m-1)}.$$

else

$$L_q = \log \alpha_q^{(m-1)} + \sum_{l=1,Q} \sum_{j \neq i(m)} z_{jl}^{(m-1)} \beta_{i(m)jql}^{(m-1)}.$$

With self loop:
If $m \leq n$

$$L_q = \log \alpha_q^{(m-1)} + \sum_{l=1,Q} \sum_{j<m} z_{jl}^{(m-1)} \beta_{mjql}^{(m-1)} + \beta_{mmqq}^{(m-1)} \tag{12}$$

else

$$L_q = \log \alpha_q^{(m-1)} + \sum_{l=1,Q} \sum_{j \neq i(m)} z_{jl}^{(m-1)} \beta_{i(m)jql}^{(m-1)} + \beta_{i(m)i(m)qq}^{(m-1)} \tag{13}$$

2.3 Stopping rule and Likelihoods.

Stopping rule The EM algorithm stops when $m = N * n$ where N is user-specified.

Complete-data loglikelihood

Without self loop:

$$\mathcal{Q}_Q = \sum_q \log \alpha_q + \sum_q \sum_l \sum_{j<i} z_{iq} z_{jl} \log \left(\pi_{ql}^{x_{ij}} (1 - \pi_{ql})^{1-x_{ij}}\right). \tag{14}$$

With self loop:

$$\mathcal{Q}_Q = \sum_q \log \alpha_q + \sum_q \sum_l \sum_{j<i} z_{iq} z_{jl} \log \left(\pi_{ql}^{x_{ij}} (1 - \pi_{ql})^{1-x_{ij}}\right) + \sum_{i,q} z_{iq} \log \left(\pi_{qq}^{x_{ii}} (1 - \pi_{qq})^{1-x_{ii}}\right). \tag{15}$$

3 Online Stochastic Classification algorithm

TO DO.

4 Online Variational algorithm

Définitions. We begin at $(m) = n_0$ with (m) the current index for iterations. Q the number of classes, τ the matrix of *posterior* probabilities (n, Q) defined such that:

$$\tau_{iq} = \Pr\{Z_{iq} = 1|\mathbf{X}\} \tag{16}$$

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

$$\forall i \sum_{q=1,Q} Z_{iq} = 1 \tag{17}$$

and

$$\forall i \sum_{q=1,Q} \tau_{iq} = 1 \tag{18}$$

We define $i(m) = mod(m-1, n)$.

4.1 M-step

If $m \leq n$

$$\tau_{\bullet q}^{(m)} = \tau_{\bullet q}^{(m-1)} + \tau_{mq}^{(m)} \tag{19}$$

else

$$\tau_{\bullet q}^{(m)} = \tau_{\bullet q}^{(m-1)} + \tau_{i(m)q}^{(m)} - \tau_{i(m)q}^{(m-1)} \tag{20}$$

Without self loop:
If $m \leq n$

$$\begin{aligned} \gamma_{ql}^{(m)} &= \gamma_{ql}^{(m-1)} + \tau_{mq}^{(m)} \tau_{\bullet l}^{(m-1)} && (21) \\ \theta_{ql}^{(m)} &= \theta_{ql}^{(m-1)} + \tau_{mq}^{(m)} \sum_{j<m} x_{mj} \tau_{jl}^{(m-1)} && (22) \\ & && (23) \end{aligned}$$

else

$$\begin{aligned} \gamma_{ql}^{(m)} &= \gamma_{ql}^{(m-1)} + (\tau_{i(m)q}^{(m)} - \tau_{i(m)q}^{(m-1)}) \tau_{\bullet l}^{(m-1)} && (24) \\ \theta_{ql}^{(m)} &= \theta_{ql}^{(m-1)} + (\tau_{i(m)q}^{(m)} - \tau_{i(m)q}^{(m-1)}) \sum_{j \neq i(m)} x_{i(m)j} \tau_{jl}^{(m-1)} && (25) \\ & && (26) \end{aligned}$$

With self loops:
If $m \leq n$

$$\begin{aligned}
\gamma^{(m)}_{ql,q\neq l} &= \gamma^{(m-1)}_{ql} + \tau^{(m)}_{mq}\tau^{(m-1)}_{\bullet l} &(27)\\
\gamma^{(m)}_{qq} &= \gamma^{(m-1)}_{qq} + \tau^{(m)}_{mq}(\tau^{(m-1)}_{\bullet q} + 1) &(28)\\
\theta^{(m)}_{ql,q\neq l} &= \theta^{(m-1)}_{ql} + \tau^{(m)}_{mq}\sum_{j<m} x_{mj}\tau^{(m-1)}_{jl} &(29)\\
\theta^{(m)}_{qq} &= \theta^{(m-1)}_{qq} + \tau^{(m)}_{mq}\sum_{j<m} x_{mj}\tau^{(m-1)}_{jq} + \tau^{(m)}_{mq}x_{mm} &(30)\\
& &(31)
\end{aligned}$$

else

$$\begin{aligned}
\gamma^{(m)}_{ql,q\neq l} &= \gamma^{(m-1)}_{ql} + (\tau^{(m)}_{i(m)q} - \tau^{(m-1)}_{i(m)q})\tau^{(m-1)}_{\bullet l} &(32)\\
\gamma^{(m)}_{qq} &= \gamma^{(m-1)}_{qq} + (\tau^{(m)}_{i(m)q} - \tau^{(m-1)}_{i(m)q})(\tau^{(m-1)}_{\bullet q} + 1) &(33)\\
\theta^{(m)}_{ql,q\neq l} &= \theta^{(m-1)}_{ql} + (\tau^{(m)}_{i(m)q} - \tau^{(m-1)}_{i(m)q})\sum_{j\neq i(m)} x_{i(m)j}\tau^{(m-1)}_{jl} &(34)\\
\theta^{(m)}_{qq} &= \theta^{(m-1)}_{qq} + (\tau^{(m)}_{i(m)q} - \tau^{(m-1)}_{i(m)q})(\sum_{j\neq i(m)} x_{i(m)j}\tau^{(m-1)}_{jq} + x_{i(m)i(m)}) &(35)
\end{aligned}$$

where

$$\begin{aligned}
\tau^{(m)}_{\bullet q} &= \sum_{i=1}^{m}\tau^{(m)}_{iq}\\
\gamma^{(m)}_{ql} &= \sum_{i=1}^{m}\sum_{j<i}\tau^{(m)}_{iq}\tau^{(m)}_{jl} \text{ (only for } q \neq l \text{ with self loops)}\\
\gamma^{(m)}_{qq} &= \sum_{i=1}^{m}\tau^{(m)}_{iq}(\sum_{j<i}\tau^{(m)}_{jq} + 1) \text{ (with self loops)}\\
\theta^{(m)}_{ql} &= \sum_{i=1}^{m}\sum_{j<i}\tau^{(m)}_{iq}x_{ij}\tau^{(m)}_{jl} \text{ (only for } q \neq l \text{ with self loops)}\\
\theta_{qq} &= \sum_{i=1}^{m}\tau^{(m)}_{iq}(\sum_{j<i}x_{ij}\tau^{(m)}_{jl} + x_{ii}) \text{ (with self loops)}
\end{aligned}$$

In any case, we have $\alpha^{(m)}_q = \frac{\tau^{(m)}_{\bullet q}}{min(n,m)}$ and the estimator for parameter π_{ql} is such that

$$\pi^{(m)}_{ql} = \frac{\theta^{(m)}_{ql}}{\gamma^{(m)}_{ql}}. \quad (36)$$

- 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 $\sum_{i\neq j} \tau_{iq}^{(m)}\tau_{jl}^{(m)} \to 0$ π_{ql} is set to 0.5. This configuration corresponds to the case where one class tends to contain only one node.

4.2 E-step

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

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

Note that π_{ql} is bounded in the M-step.
We recall $i(m) = mod(m-1, n)$.

Without self loop:

$$\log \tau_{i(m)q}^{(m)} = \log \alpha_q^{(m-1)} + \sum_{j\neq i(m)} \sum_{l=1,Q} \tau_{jl}^{(m-1)} \beta_{i(m)jql}^{(m-1)}, \tag{37}$$

With self loops:

$$\log \tau_{i(m)q}^{(m)} = \log \alpha_q^{(m-1)} + \sum_{j\neq i} \sum_{l=1,Q} \tau_{jl}^{(m-1)} \beta_{i(m)jql}^{(m-1)} + \beta_{i(m)i(m)qq}^{(m-1)}, \tag{38}$$

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:

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

4.3 Stopping rule and Likelihoods.

Stopping rule The EM algorithm stops when $m = N * n$ where N is user-specified.

Incomplete-data log-likelihood approximation.

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

Complete-data log-likelihood.

Undirected case without self loop:

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

Undirected case with self loops:

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

Entropy.

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

5 Criteria

BIC.

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

ICL.

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