

Team Name, Semester Year

Names

February 4, 2017

Introduction

n individuals K communities Each individual belongs to exactly one community Denote by A the (n, K) membership matrix, i.e $A_{i,j} = 1$ if the i -th individual belongs to the j -th community (and 0 otherwise), for $1 \leq i \leq n$ and $1 \leq j \leq K$. Denote by X the (n, n) connectivity matrix. We use the SBM model : SBM assumes that for each $1 \leq i, j \leq n$, $X_{i,j}$ follows a Bernoulli law whose parameter only depends on $g(i)$ and $g(j)$, the respective groups of i and j . Furthermore, it assumes that the coordinates of the matrix X are all independent. Letting C be the (K, K) matrix such that $C_{i,j}$ is the parameter of connectivity between groups i and j , one can write $E[X] = ACA^T - \text{diag}(ACA^T)$. In what follows we write $X = ACA^T + \mathcal{E} - D$, where $\mathcal{E} = X - E[X]$ is a zero-mean matrix and $D = \text{diag}(ACA^T)$. The objective is to recover the membership matrix A , up to a permutation, given one realization of X , i.e given one instance of connections between the n individuals. Note that the membership matrix can be represented equivalently by the "normalized" membership (n, n) matrix B^* defined as follows : $B_{i,j} = \frac{1}{|G_k|}$ if i and j both belong to group k and $B_{i,j} = 0$ otherwise. Following the notations of [2], we now write $X = ZA^T + E$, where $Z = AC$ and $E = \mathcal{E} - D$. In doing so we can see the SBM model as a special instance of the G-latent models defined in [2]. This paper shows that the main guarantees and results of [2] can be successfully adapted to the SBM model. Denoting $\Delta(C) = \min_{j < k} (C_{kk} + C_{jj} - 2C_{jk})$, namely we show that under some conditions on $\Delta(C)$, one can recover the exact matrix B^* by solving a convex optimization problem : Let

$$\mathcal{C} = \left\{ \begin{array}{l} B \succeq 0 \\ \Sigma_a B_{ab} = 1, \forall b \\ B_{ab} \geq 0, \forall a, b \\ \text{tr}(B) = K \end{array} \right\} \subset \mathbb{R}^{p \times p}$$

Let $\hat{\Sigma} = X^t X$. PECOK algorithm :

- 1/ Estimate B^* by $\hat{B} = \arg\max_{B \in \mathcal{C}} \langle \hat{\Sigma}, B \rangle$
- 2/ Estimate G^* by applying a clustering algorithm to the columns of \hat{B} .

In this paper, we develop sufficient conditions on the SBM model, via the quantity $\Delta(C)$, so that the PECOK algorithm above recovers B^* , and hence G^* , exactly with high probability.

Our investigation follows the outline of [2], as its main arguments can be adapted to our case. Lemma 1 p.6 and its proof p.16 remain valid and so is Lemma 3 p.16. So we only need to prove that $\langle \hat{\Sigma}, B^* - B \rangle \geq 0$ for all $B \in \mathcal{C}$ such that $\text{supp}(B) \not\subseteq \text{supp}(B^*)$, with high probability. Following the decomposition (46) we write similarly $W = W_1 + W_2 + E^2$.

- [1] Lei, Rinaldo. Consistency of spectral clustering in stochastic block models.
- [2] PECOK : a convex optimization approach to variable clustering.

Lemma 1:

with probability larger than $1 - \frac{2}{n}$ it holds that :

$$|\langle W_2, B^* - B \rangle| \leq \sum_{j \neq k} (4(\ln n + |D|_\infty) |Z_{:j} - Z_{:k}|_\infty + \sqrt{6 \ln n \cdot (V_j + V_k)} |Z_{:j} - Z_{:k}|_2) |B_{G_j G_k}|_1$$

where $V_j = \max_{c \in \{1..n\}} Z_{cj}(1 - Z_{cj})$

Proof:

Consider any a and b in $[n]$ and let j and k be such that $a \in G_j$ and $b \in G_k$. If $j = k$, $(W_2)_{ab} = 0$. if $j \neq k$, then we have :

$$\begin{aligned} (W_2)_{ab} &= [E_{b:} - E_{a:}] \cdot [Z_{:j} - Z_{:k}] \\ &= \sum_{c \neq a, c \neq b} (\mathcal{E}_{bc} - \mathcal{E}_{ac}) \cdot (Z_{cj} - Z_{ck}) + \mathcal{E}_{ab}(Z_{aj} + Z_{bk} - Z_{bj} - Z_{ak}) - D_{bb} \cdot (Z_{bj} - Z_{bk}) - D_{aa} \cdot (Z_{aj} - Z_{ak}). \end{aligned}$$

$$\text{Hence } \mathbb{E}[(W_2)_{ab}] = -D_{bb} \cdot (Z_{bj} - Z_{bk}) - D_{aa} \cdot (Z_{aj} - Z_{ak}).$$

Note that the sum above is a sum of centered independent variables, as \mathcal{E}_{de} is independent of all other variables \mathcal{E}_{fg} but \mathcal{E}_{ed} . Thus we can apply Bernstein's inequality :

$$\mathbb{P}(|(W_2)_{ab} - \mathbb{E}[(W_2)_{ab}]| > t) \leq 2 \exp\left(-\frac{\frac{1}{2}t^2}{\text{Var}((W_2)_{ab}) + \frac{2}{3}|Z_{:j} - Z_{:k}|_\infty t}\right),$$

as $|\mathcal{E}_{de}| \leq 1$ for all d, e .

Hence with probability less than $\frac{2}{n^3}$, it holds that :

$$|(W_2)_{ab} - \mathbb{E}[(W_2)_{ab}]| > 4 \ln n |Z_{:j} - Z_{:k}|_\infty + \sqrt{6 \ln n \cdot \text{Var}((W_2)_{ab})}$$

Now

$$\begin{aligned} \text{Var}((W_2)_{ab}) &= \sum_{c \neq a, c \neq b} (\text{Var}(\mathcal{E}_{bc}) + \text{Var}(\mathcal{E}_{ac})) \cdot (Z_{cj} - Z_{ck})^2 + \text{Var}(\mathcal{E}_{ab}) (Z_{aj} + Z_{bk} - Z_{bj} - Z_{ak})^2 \\ &= \sum_{c \neq a, c \neq b} (Z_{ck}(1 - Z_{ck}) + Z_{cj}(1 - Z_{cj})) \cdot (Z_{cj} - Z_{ck})^2 \\ &\quad + \frac{Z_{ak}(1 - Z_{ak}) + Z_{bj}(1 - Z_{bj})}{2} (Z_{aj} + Z_{bk} - Z_{bj} - Z_{ak})^2 \end{aligned}$$

$$\text{Hence } \text{Var}((W_2)_{ab}) \leq (V_j + V_k) |Z_{:j} - Z_{:k}|_2^2.$$

By a union bound we have the inequality required.