Statistical Pattern Recognition: Contraction Problems in Stochastic Block Models

Eren Aldis

May 20, 2020

We are given a graph $G = \langle V, E\rangle$ from some $f \in \mathscr{F}$, where f is a Stochastic d-Block Model with parameters π and B (and n). Furthermore, we have $\dim(\pi) = d$ and B is a $d \times d$ matrix.

$$\pi = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_d \end{bmatrix}, B = \begin{bmatrix} B_{11} & \cdots & B_{1d} \\ \vdots & \ddots & \\ B_{d1} & \cdots & B_{dd} \end{bmatrix}$$

Let $V^{(i)}$ denote the set of vertices in i-th block. Hence, $V = \{V^{(1)}, V^{(2)}, \dots, V^{(d)}\}$. Then, for some $v \in V$, $\pi_i = P(v \in V^{(i)})$. Thus, we also have $\pi_1 + \pi_2 + \cdots + \pi_d = 1$. Similarly, for some $v_i \in V^{(i)}$ and $v_j \in V^{(j)}$, $B_{ij} = P(v_i \sim v_j \in E)$ where $v_i \sim v_j$ denote the edge between v_i, v_j. Therefore, $B_{ij} = B_{ji}$, so the B matrix is symmetric. Lastly, we also have $V = \{v_1, v_2, \dots, v_n\}$.

Problem I (Easy)

Say we contract the graph G into $G' = \langle V', E'\rangle$ by randomly selecting two vertices both from the **same block** into one new vertex. We will denote the contraction using the $\oplus$ operator. Hence, in this case for each $v' \in V'$ we will have $v' = v_1 \oplus v_2$ where $v_1, v_2 \in V^{(i)}$.

The goal is to figure out how to model G' using an SBM. Hence, the first question we should be asking is whether $G' \sim f$ where f is the SBM that models G with parameters π, B. Note that this would precisely mean that $G' \sim SBM(\pi, B)$. Hence, the contraction does not change the values of the parameters of the model. We will show that this is not the case. Hence, we will show $G' \sim f'$ where $f' = SBM(\pi', B')$.

Note that this still means that $f' \in \mathscr{F}$ since $\dim(B') = \dim(B) = d$ and $\dim(\pi') = \dim(\pi) = d$. This is because, since we are only contracting the vertices of G that are in the same blocks, we do not have any merging blocks. All the new vertices in G' which have been contracted from two vertices from the same block in G, will stay in the same block. In other words, $v_1, v_2 \in V^{(i)}, v' = v_1 \oplus v_2 \Leftrightarrow v' \in V'^{(i)}$. Hence, we are not going to gain a block and we will not lose a block, thus we will have the same number of blocks in the model. Hence, the dimensions of the model will stay the same.

We will first show that $\pi' = \pi$. Since $dim(\pi') = d$, we can write $\pi' = [\pi'_1, \pi'_2, \dots, \pi'_d]^T$. Then, for some $v' \in V$ such that $\exists v_1, v_2 \in V, v' = v_1 \oplus v_2$, and for $V'^{(i)})$ that denotes the set of vertices in the i-th block in the new G'.

$$\begin{aligned} \pi'_i &= P(v' \in V'^{(i)}) \\ &= P(v_1 \oplus v_2 \in V'^{(i)}) \\ &= P(v_1 \in V^{(i)}, v_2 \in V^{(i)}) \\ &= P(v_1 \in V^{(i)} | v_2 \in V^{(i)}) P(v_2 \in V^{(i)}) \\ &= 1 \times \pi_i = \pi_i \end{aligned}$$

The second to third equality, follows from the fact that $v' \in V'^{(i)} \Rightarrow \exists v_1, v_2 \in V^{(i)}, v' = v_1 \oplus v_2$ as we claimed before. The final equality follows from $P(v_1 \in V^{(i)} | v_2 \in V^{(i)}) = 1$ since we can only contract vertices from the same block. Hence given that $v_2 \in V^{(i)}$ and $v' = v_1 \oplus v_2$, $v_1 \in V^{(i)}$ must be true with probability 1. Thus, we have $\pi' = \pi$ as we had claimed.

Now, we will derive B'. Note that, we claim that $B' \neq B$ and therefore, $f' \neq f$. For some $v'_i, v'_j \in V'$ such that $v'_i \in V'^{(i)}$ and $v'_j \in V'^{(j)}$, and $\exists\ v_1, v_2 \in V^{(i)}, v_1 \oplus v_2 = v'_i$ and $\exists\ v_3, v_4 \in V^{(j)}, v_3 \oplus v_4 = v'_j$,

$$B'_{ij} = P(v'_i \sim v'_j \in E')$$

$$= P(v_1 \oplus v_2\ v'_i \sim v_3 \oplus v_4 \in E')$$

$$= P\left(v_1 \sim v_3 \in E \cup v_1 \sim v_4 \in E \cup v_2 \sim v_3 \in E \cup v_2 \sim v_4 \in E\right)$$

$$= 1 - P\left(v_1 \sim v_3 \notin E, v_1 \sim v_4 \notin E, v_2 \sim v_3 \notin E, v_2 \sim v_4 \notin E\right)$$

$$= 1 - P(v_1 \sim v_3 \notin E)P(v_1 \sim v_4 \notin E)P(v_2 \sim v_3 \notin E)P(v_2 \sim v_4 \notin E)$$

$$= 1 - (1 - B_{ij})(1 - B_{ij})(1 - B_{ij})(1 - B_{ij})$$

$$= 1 - (1 - B_{ij})^4$$

where the second equality follows since the edges between vertices are still kept after contraction. That is, at least one edge between v_1, v_2 and v_3, v_4, is enough for there to be an edge between v'_i and v'_j. The third equality follows from De Morgan's Law. The fourth equality follows from independence in sampling from the B matrix to construct the edges in the initial G graph. The penultimate equality follows since $v_1, v_2 \in V^{(i)}$ and $v_3, v_4 \in V^{(j)}$ and thus, $P(v_i \sim v_j \notin E) = 1 - P(v_i \sim v_j \in E) = 1 - B_{ij}$. Hence, $B'_{ij} = 1 - (1 - B_{ij})^4$. Note that this applies for all i, j (even when $i = j$).

An interesting extension is that, if the contractions had occured say, instead of two, between k vertices in every block, we would have $B'_{ij} = 1 - (1 - B_{ij})^{k^2}$. For different number of vertices set to be contracted in each block, i.e. k vertices to be contracted in i-th block and m vertices to be contracted in j-th block, we would have $B'_{ij} = 1 - (1 - B_{ij})^{km}$.

Hence $G' \sim SBM(\pi, B')$.

Problem II (Medium)

Now, to construct $G' = \langle V', E' \rangle$say we sample the two vertices $v_i, v_j \in V$ that we are going to contract completely randomly from V (without replacement). Hence, $v_i \in V^{(i)}, v_j \in V^{(j)}$ where i, j not necessarily equal.

We claim that $G' \sim f \in \mathscr{F}'$ where $\mathscr{F}' \neq \mathscr{F}$ since $\dim(\pi') \neq d$ and $\dim(B') \neq d$ this time. This is because if we have d blocks, and we are allowed to contract $v_i \in V^{(i)}, v_j \in V^{(j)}$, this would result in a new block when $i \neq j$. Thus, necessarily, $\dim(\pi') = \dim(B') > d$. In fact, it is quite simple to compute the new dimension,

$$d' = d + \binom{d}{2}$$

since we will not keep the first d dimensions since we will still have those blocks and there will be $\binom{d}{2}$ many new blocks from pairwise combinations of two different blocks.

Now to simplify (!) the parametrization, we will define a function $\Phi : [1, d] \times [1, d] \to [1, d']$ so that we can index the parameters coherently w.r.t the indices of G. Thus,

$$\Phi(i, j) = \begin{cases} i & \text{if } i = j \\ id + j - 1 & \text{if } i \neq j \end{cases}$$

Then we can write $\pi' = [\pi'_1, \dots \pi'_d, \dots, \pi'_{d'}]^T = [\pi'_{11}, \dots \pi'_{dd}, \pi'_{12} \dots, \pi'_{(d-1)d}]^T$ where for some $v' \in V'$, $\pi'_{\Phi(i,j)} = P\left(v' \in V'^{(\Phi(i,j))}\right) = P\left(v_1 \in V^{(i)}, v_2 \in V^{(j)} \cup v_2 \in V^{(i)}, v_1 \in V^{(j)}\right)$ for $v' = v_1 \oplus v_2,\ v_1, v_2 \in V$. Then for $i, j \leq d, \mathbf{i} = \mathbf{j}$,

$$\pi'_{\Phi(i,j)} = P\left(v' \in V'^{(\Phi(i,j))}\right) = P\left(v_1 \in V^{(i)}, v_2 \in V^{(i)} \cup v_2 \in V^{(i)}, v_1 \in V^{(i)}\right)$$

$$= P\left(v_1 \in V^{(i)}, v_2 \in V^{(i)}\right)$$

$$= P(v_1 \in V^{(i)})P(v_2 \in V^{(i)})$$

$$= \pi_i^2$$

where the penultimate equality follows from the assumption of independent sampling from V while choosing the pairs to contract. Hence, $\forall i \leq d,\ \pi'_i = \pi_i^2$. Now for $i, j \leq d, \mathbf{i} \neq \mathbf{j}$,

$$\pi'_{\Phi(i,j)} = P\left(v' \in V'^{(\Phi(i,j))}\right) = P\left(v_1 \in V^{(i)}, v_2 \in V^{(i)} \cup v_2 \in V^{(i)}, v_1 \in V^{(i)}\right)$$

$$= P(v_1 \in V^{(i)}, v_2 \in V^{(i)}) + P(v_2 \in V^{(i)}, v_1 \in V^{(i)})$$

$$= 2\pi_{ij}$$

where the penultimate equality follows since the two events are obviously mutually exclusive. Hence, $\forall i, j \leq d$, $\pi'_{id+j-1} = 2\pi_{ij}$. Thus,

$$\pi' = \begin{bmatrix} \pi_1^2 \\ \vdots \\ \pi_d^2 \\ 2\pi_{12} \\ \vdots \\ 2\pi_{(d-1)d} \end{bmatrix}$$

We know that B' is a $d' \times d'$ matrix. For some $v'_1 \in V'^{\Phi(i,j)}, v'_2 \in V'^{\Phi(r,t)}$, $\exists\, v_i \in V^{(i)}, v_j \in V^{(j)}$ such that $v'_1 = v_i \oplus v_j$ and $\exists\, v_r \in V^{(r)}, v_t \in V^{(t)}$ such that $v'_2 = v_r \oplus v_t$,

$$B'_{\Phi(i,j),\Phi(r,t)} = P\left(v_1 \sim v_2 \in E'\right)$$

$$= P\left(v_i \oplus v_j \sim v_r \oplus v_t \in E'\right)$$

$$= P\left(v_i \sim v_r \in E \cup v_i \sim v_t \in E \cup v_j \sim v_r \in E \cup v_j \sim v_t \in E\right)$$

$$= 1 - P\left(v_i \sim v_r \notin E, v_i \sim v_t \notin E, v_j \sim v_r \notin E, v_j \sim v_t \notin E\right)$$

$$= 1 - (1 - B_{ir})(1 - B_{it})(1 - B_{jr})(1 - B_{jt})$$

Thus, indeed if $\mathscr{F}'$ is the set of d'-dimensional SBMs, $G' \sim f \in \mathscr{F}'$ where $f = SBM(\pi', B')$ as given above.

Empirical Evidence

For empirical evidence that the results in Problem I hold, see the code attached below.

Acknowledgments

I collaborated with Ceki Papo on deriving the SBM parameters on both problems.