A stochastic block model for hypergraphs

LUCA BRUSA *(luca.brusa@unimib.it)*

University of Milano-Bicocca - Department of Economics, Management and Statistics

CATHERINE MATIAS *(catherine.matias@math.cnrs.fr)*

Sorbonne Université, Université de Paris Cité, Centre National de la Recherche Scientifique

August 24, 2022

Outline

1. Hypergraphs
2. Stochastic blockmodel for hypergraphs
   - Model formulation
   - Model identifiability
   - Parameter estimation
3. Simulation studies
   - Performance of VEM algorithm
   - Performance of model selection
4. Conclusions

Higher-order interactions

- Over the past two decades a broad variety of models has been developed for pairwise interactions, encoded in graphs
- Modern applications highlight the need to account for higher-order interactions, to include the information deriving from groups of three or more nodes
- Simple examples include triadic and larger groups interactions in social networks, scientific co-authorship, interactions between more than two species in ecological systems, and higher-order interactions between neurons in brain networks
- A graph description lacks a proper interpretation: it is impossible to state weather any higher-order interaction is actually present or not

Hypergraphs

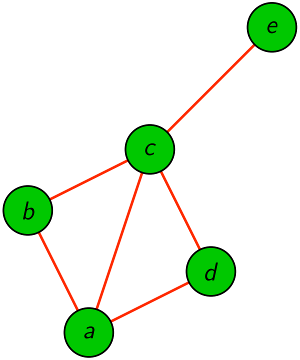
Definition

A (simple) hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ is defined as a set of nodes $\mathcal{V} \neq \emptyset$ and a set of hyperedges $\mathcal{E}$. Each hyperedge is a non-empty collection of m distinct nodes taking part within an interaction

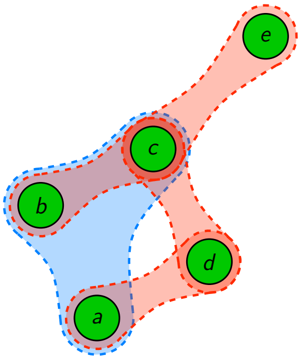
- Hypergraphs naturally include the entity of graphs, by simply considering $m = 2$ for each hyperedge $e \in \mathcal{E}$
- A hypergraph can contain a hyperedge of size 3 $[a, b, c]$ without any requirement on the existence of the hyperedges of size 2 $[a, b]$, $[a, c]$, and $[b, c]$

Graph vs Hypergraph representation

Set of higher-order interactions: $\{[a, b, c], [a, d], [c, d], [c, e]\}$



(a) Graph representation



(b) Hypergraph representation

Outline

1. Hypergraphs
2. Stochastic blockmodel for hypergraphs
   - Model formulation
   - Model identifiability
   - Parameter estimation
3. Simulation studies
   - Performance of VEM algorithm
   - Performance of model selection
4. Conclusions

Notation - Observable component

- $\mathcal{H} = (\mathcal{V}, \mathcal{E})$, with $\mathcal{V} = \{1, \ldots, n\}$ set of nodes and $\mathcal{E}$ set of hyperedges
- $M = \max_{e \in \mathcal{E}} |e| \geq 2$, largest size of hyperedges in $\mathcal{E}$
- $\mathcal{V}^{(m)} = \big\{\{i_1, \ldots, i_m\} : i_1, \ldots, i_m \in \mathcal{V} \text{ and } i_1 \neq \ldots \neq i_m\big\}$, set of unordered node tuples of size m
- $\mathcal{E}^{(m)} = \big\{\{i_1, \ldots, i_m\} \in \mathcal{V}^{(m)} : \{i_1, \ldots, i_m\} \in \mathcal{E}\big\}$, set of hyperedges of size m
- $Y_{i_1,\ldots,i_m} = \mathbb{1}_{\{i_1,\ldots,i_m\} \in \mathcal{E}}$ for each $\{i_1, \ldots, i_m\} \in \mathcal{V}^{(m)}$

Notation - Latent component

- $1, \dots, Q$, latent groups
- $\boldsymbol{Z} = (Z_1, \dots, Z_n)$, independent and identically distributed latent variables having a discrete distribution with support points $\{1, \dots, Q\}$
- $Z_i \rightarrow (Z_{i1}, \dots, Z_{iQ})$, with $Z_{iq} = 1$ if node i belongs to latent group q and $Z_{iq} = 0$ otherwise
- Model parameters:
  - $\pi_q = \mathbb{P}(Z_i = q)$: prior probability of latent group q
  - $B^{(m)}_{q_1,\dots,q_m}$: probability that m unordered nodes with latent configuration $\{q_1, \dots, q_m\}$ are connected into a hyperedge
- $Y_{i_1,\dots,i_m} | \{Z_1 = q_1, \dots, Z_m = q_m\} \overset{i.i.d.}{\sim} \text{Bern}(B^{(m)}_{q_1,\dots,q_m})$

Parameter identifiability

Generic identifiability: a parameter θ almost surely (*w.r.t.* Lebesgue measure) uniquely defines the distribution $\mathbb{P}_\theta$ up to label switching on the node groups

Theorem

For any $m \geq 2$ and any Q, the parameter $\theta^{(m)} = (\pi_q, B^{(m)}_{q_1,\dots,q_m})_{q,q_1,\dots,q_m}$ of the HSBM restricted to m-uniform (simple) hypergraphs over n nodes, is generically identifiable for large enough n

Corollary

For any Q, the parameter $\theta = (\pi_q, B^{(m)}_{q_1,\dots,q_m})_{m,q,q_1,\dots,q_m}$ of the HSBM for (simple) hypergraphs over n nodes, is generically identifiable for large enough n

Variational approximation

- EM algorithm is not feasible because latent variables are not independent conditional on observed ones
- Variational approximation to EM algorithm: replace the intractable posterior distribution by the best approximation (with respect to Kullback-Leibler divergence) in a class of simpler distributions:

$$\mathbb{Q}_\tau(Z_1, \ldots, Z_n) = \prod_{i=1}^{n} \mathbb{Q}_\tau(Z_i) = \prod_{i=1}^{n} \prod_{q=1}^{Q} \tau_{iq}^{Z_{iq}},$$

with the variational parameter $\tau_{iq} = \mathbb{Q}_\tau(Z_i = q) \in [0, 1]$ and $\sum_{q=1}^{Q} \tau_{iq} = 1$, for any $i = 1, \ldots, n$ and $q = 1, \ldots, Q$

Evidence lower bound

Define the following function based on the variational distribution $\mathbb{Q}_\tau$:

$$\mathcal{J}(\theta, \tau) = \mathbb{E}_{\mathbb{Q}_\tau}[\log \mathbb{P}_\theta(\boldsymbol{Y}, \boldsymbol{Z})] - \mathbb{E}_{\mathbb{Q}_\tau}[\log \mathbb{Q}_\tau(\boldsymbol{Z})]$$

- $\mathcal{J}(\theta, \tau)$ satisfies $\mathcal{J}(\theta, \tau) = \log \mathbb{P}_\theta(\boldsymbol{Y}) - \mathrm{KL}(\mathbb{Q}_\tau(\boldsymbol{Z}) || \mathbb{P}_\theta(\boldsymbol{Z}|\boldsymbol{Y}))$, where KL denotes the Kullback-Leibler divergence
- It follows that $\mathcal{J}(\theta, \tau) \leq \log \mathbb{P}_\theta(\boldsymbol{Y})$, with equality iff $\mathbb{Q}_\tau(\boldsymbol{Z})$ is the true posterior $\mathbb{P}_\theta(\boldsymbol{Z}|\boldsymbol{Y})$
- Maximise the lower bound $\mathcal{J}(\theta, \tau)$ (with respect to τ and θ) instead of the intractable log-likelihood $\log \mathbb{P}_\theta(\boldsymbol{Y})$

VEM algorithm

- **VE-Step** maximizes $\mathcal{J}(\theta, \tau)$ with respect to τ:

$$\widehat{\tau}^{(t)} = \arg\max_{\tau} \mathcal{J}(\theta^{(t-1)}, \tau); \quad \text{s.t.} \ \sum_{q=1}^{Q} \tau_{iq} = 1 \qquad \forall i = 1, \ldots, n.$$

This is equivalent to minimising the Kullback-Leibler divergence
In practice this step is obtained by a fixed-point algorithm

- **M-Step** maximizes $\mathcal{J}(\theta, \tau)$ with respect to θ:

$$\widehat{\theta}^{(t)} = \arg\max_{\theta} \mathcal{J}(\theta, \tau^{(t-1)}), \quad \text{s.t.} \ \sum_{q=1}^{Q} \pi_q = 1,$$

thus updating the value of the model parameters π_q and $B^{(m)}_{q_1,\ldots,q_m}$.

Outline

1. Hypergraphs
2. Stochastic blockmodel for hypergraphs
   - Model formulation
   - Model identifiability
   - Parameter estimation
3. Simulation studies
   - Performance of VEM algorithm
   - Performance of model selection
4. Conclusions

Simulation setting

- 10 Hypergraphs are simulated from the HSBM with $Q = 2$ latent groups ($\pi_1 = 0.6$ and $\pi_2 = 0.4$), $M = 3$, and $n \in \{50, 100, 150, 200\}$
- Various scenarios according to a simplified submodel:

$$B_{q_1,\dots,q_m} = \begin{cases} \alpha & \text{if } q_1 = \dots = q_m \\ \beta & \text{if there exist at least } q_i \neq q_j \end{cases} \qquad \forall m = 2, \dots, M$$

A. *Communities*: case of high intra-groups and low inter-groups connection probabilities ($\alpha = 0.7 > \beta = 0.3$);

B. *Disassortative*: case of low intra-groups and high inter-groups connection probabilities ($\alpha = 0.3 < \beta = 0.7$)

C. *Erdös-Rényi-like*: diffcult case of very similar intra-groups and inter-groups connection probabilities ($\alpha = 0.25 \approxeq \beta = 0.35$)

Recovery of the correct clustering (ARI)

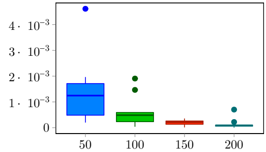
We rely on the Adjusted Rand Index, measuring the similarity between the correct node clustering and the estimated one

n	Scenario A	Scenario B	Scenario C
50	1.00	1.00	**0.50**
100	1.00	1.00	**0.90**
150	1.00	1.00	1.00
200	1.00	1.00	1.00

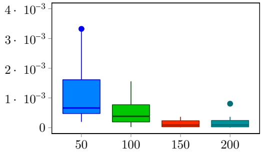
- Scenarios A and B: all values are equal to 1 and the correct clusters are perfectly recovered in all cases
- Scenario C: the proposed approach sometimes fails to recover the optimal clustering, in particular in the case with $n = 50$ nodes, where the average ARI is rather low

Estimation of the model parameters (MSE)

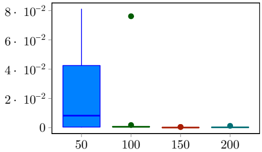
We rely on an aggregated Mean Squared Error over all the components of θ measuring the distance between true and estimated parameters



(a) Scenario A



(b) Scenario B



(c) Scenario C

- Again, scenarios A and B provide the best results, with values of the MSE that are always lower than 0.5%
- Scenario C confirms to be the most difficult from the estimation perspective, showing the highest MSE for each value of n (up to 8%)

Model selection setting

- We simulate 50 hypergraphs from the HSBM with $Q = 3$ latent states and assuming the same simplified formulation for the latent structure (with $\alpha = 0.7$ and $\beta = 0.3$)
- Two different values are tested for the number of nodes, $n = 100$ and $n = 200$
- The largest size M of hyperedges is set equal to 3
- The simulated data is then fitted with the HSBM with a number of latent states ranging from 1 to 5
- We rely on the Integrated Classification Likelihood: $\hat{q} = \arg\max_{q} ICL(q)$

Model selection results

	$n = 100$		$n = 200$	
Q	Percentage	ARI for 3 groups	Percentage	ARI for 3 groups
2	0%	-	2%	0.55
3	68%	1.00	90%	1.00
4	22%	0.57	6%	0.60
5	10%	0.58	2%	0.61

- The correct model is selected in 68% of cases for $n = 100$ and in 90% of cases for $n = 200$
- The ARI value of the classification obtained with 3 clusters is equal to 1 when the correct model is recovered
- When an incorrect number of groups is selected, values of ARI are quite low (around or smaller than 0.60)

Outline

1 Hypergraphs

2 Stochastic blockmodel for hypergraphs
- Model formulation
- Model identifiability
- Parameter estimation

3 Simulation studies
- Performance of VEM algorithm
- Performance of model selection

4 Conclusions

Conclusions

- We propose a Stochastic Blockmodel for clustering the nodes of a (simple) hypergraph
- We establish (generic) identifiability of the parameters of the model
- Estimation and nodes clustering is performed through VEM algorithm
- ICL criterion is used to select the number of groups
- `R` package (`https://github.com/LB1304/HyperSBM`) and preprint available very soon (write me an email!)

Any questions ?