Exact recovery and Bregman hard clustering of node-attributed Stochastic Block Model

Maximilien Dreveton — EPFL (Switzerland)
Felipe Fernandes — Federal University of Rio de Janeiro (Brazil)
Daniel Figueiredo — Federal University of Rio de Janeiro (Brazil)

Motivation

Setup

- Hidden data: Nodes are divided into clusters.
- Observed data: Interactions between node pairs (weighted graph) and node attributes (feature vector).

Contributions

- Theoretical: how much information brought by the network and the attributes is required for exact recovery of clusters?
- Practical: an algorithm that combines *network data* (almost always sparse and possibly weighted) and *attribute data* (discrete or continuous) to place nodes into *clusters* (unsupervised learning).

Model

- n nodes are divided into K blocks. We denote by $z \in [K]^n$ the block membership vector. $z_1, \cdots, z_n$ are iid such that $\mathbb{P}(z_i = a) = \pi_a$.
- $f_{ab}(X_{ij})$: probability of observing an interaction X_{ij} between node i in block a and node j in block b;
- $h_a(Y_i)$: probability of observing an attribute Y_i for node i in block a;
- Pairwise interactions $(X_{ij})_{1\le i,j\le n}$ and node attributes $(Y_i)_{1\le i\le n}$ are independent conditionally on the block membership.

Conditional distribution of the data (X, Y) given block membership z:

$$\mathbb{P}(X, Y \mid z) = \prod_{1\le i<j\le n} f_{z_i z_j}(X_{ij}) \prod_{i=1}^{n} h_{z_i}(Y_i). \tag{1}$$

Exact recovery threshold

Denote $\mathrm{D}_t(f; g) = \frac{1}{t-1}\log \int f^t g^{1-t}$ the *Rényi divergence* of order t between pdf f and g.
A key information-theoretic divergence concerning a set of K blocks is

$$I = \min_{a\neq b\in[K]} \mathrm{CH}(a, b).$$

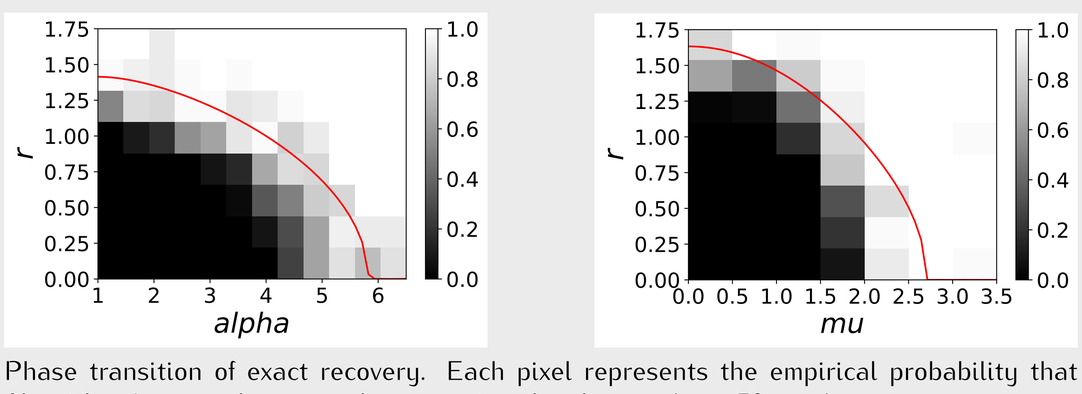
where $\mathrm{CH}(a, b) = \sup_{t\in(0,1)} (1-t)\Bigg[\underbrace{\sum_{c=1}^{K} \pi_c \,\mathrm{D}_t\left(f_{bc}; f_{ac}\right)}_{\text{information from the network}} + \underbrace{\frac{1}{n}\mathrm{D}_t\left(h_b; h_a\right)}_{\text{information from the attributes}} \Bigg].$

Theorem. Suppose $K = \Theta(1)$ and $\pi_a > 0$ for all $a \in [K]$. Denote by a^*, b^* the two hardest blocks to distinguish, that is $\mathrm{CH}(a^*, b^*) = I$. Suppose for all $t \in (0, 1)$, $\lim_{n\to\infty} \frac{n}{\log n}\mathrm{CH}_t(a^*, b^*)$ exists and is strictly concave. Then the following holds:

1. exact recovery of z is impossible if $\lim_{n\to\infty} \frac{n}{\log n} I < 1$;
2. exact recovery of z is possible if $\lim_{n\to\infty} \frac{n}{\log n} I > 1$.

Information from the network and attributes can be exchanged to ensure exact recovery!

Illustration of the exact recovery threshold



Phase transition of exact recovery. Each pixel represents the empirical probability that Algorithm 1 succeeds at exactly recovering the clusters (over 50 runs).
red line: theoretical curve for exact recovery threshold
(left) $n = 500$, $K = 2$, $f_{\text{in}} = \mathrm{Ber}(\alpha n^{-1}\log n)$, $f_{\text{out}} = \mathrm{Ber}(n^{-1}\log n)$. The attributes are 2d-spherical Gaussian with radius $(\pm r\sqrt{\log n}, 0)$ and identity covariance matrix.
(right) $n = 600$, $K = 3$, $f_{\text{in}} = (1-\rho)\delta_0 + \rho\,\mathrm{Nor}(\mu, 1)$, $f_{\text{out}} = (1-\rho)\delta_0 + \rho\,\mathrm{Nor}(0, 1)$ with $\rho = 5n^{-1}\log n$. The attributes are 2d-spherical Gaussian whose means are the vertices of a regular polygon on the circle of radius $r\sqrt{\log n}$.

Sparse weighted node-attributed networks

We consider the model defined in (1), such that f_{ab} are *zero-inflated distributions* given by

$$f_{ab}(x) = (1 - p_{ab})\delta_0(x) + p_{ab}\tilde{f}_{ab}(x), \tag{2}$$

where $p_{ab} \in [0, 1]$ is the interaction probability between blocks a and b, $\delta_0(x)$ is the Dirac delta at zero, and $\tilde{f}_{ab}$ is a probability density with no mass at zero.
We suppose that the distributions $\{\tilde{f}_{ab}\}$ and $\{h_a\}$ belong to *exponential families*, *i.e.*,

$$\tilde{f}_{ab}(x) = e^{<\theta_{ab}, T_f(x)> - \psi(\theta_{ab})} \quad \text{and} \quad h_a(y) = e^{<\eta_a, T_h(y)> - \phi(\eta_a)}, \tag{3}$$

for some parameters θ_{ab}, η_a and functions T_f, T_g, ψ, ϕ.

Log-likelihood of the model

Definition. Given a convex function ψ, the *Bregman divergence* $d_\psi\colon \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_+$ is defined by

$$d_\psi(x, y) = \psi(x) - \psi(y) - < x - y, \nabla\psi(y) > .$$

The log-likelihood of the density $p_{\theta,\psi}(x) = e^{<\theta, x> - \psi(\theta)}$ of an exponential family distribution equals

$$\log p_{\psi,\theta}(x) = -d_{\psi^*}(x, \mu) + \psi^*(x),$$

Lemma. Suppose that X, Y follow the model (1) with probability distributions given by (2)-(3). Let A be a binary matrix such that $A_{ij} = 1(X_{ij} \neq 0)$. We have

$$-\log \mathbb{P}(X, Y \mid z) = \sum_i \left\{ \frac{1}{2}\sum_{j\neq i} \left[\mathrm{d}_{\mathrm{KL}}(A_{ij}, p_{z_i z_j}) + A_{ij} d_{\psi^*}\left(X_{ij}, \mu_{z_i z_j}\right)\right] + d_{\phi^*}(Y_i, \nu_{z_i}) \right\} + c,$$

Expression is relatively simple to compute as d_{ψ^*} has a closed form for many well-known distributions.

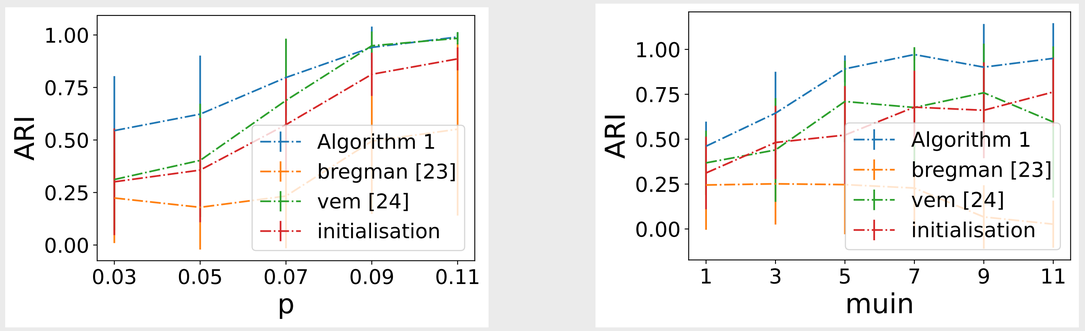
Bregman hard clustering

Input: Interactions X, attributes Y, convex functions ψ^*, ϕ^*, initial clustering Z_0.
repeat
  Estimate $\hat{p}, \hat{\mu}, \hat{\nu}$ using the current prediction for the cluster memberships Z;
  for $i = 1, \ldots, n$ **do**
    Let $Z^{(ia)}$ be the membership matrix obtained from Z by placing node i in cluster a
    Find $k^* = \arg\max_{a\in[K]} L_{ia}\left(Z^{(ia)}\right)$, where

$$L_{ia}(Z) = \frac{1}{2}\mathrm{d}_{\mathrm{KL}}\left(A_{i\cdot}, \left(Z\hat{p}Z^T\right)_{i\cdot}\right) + \frac{1}{2}d'_{\psi^*}\left(X_{i\cdot}, \left(Z\hat{\mu}Z^T\right)_{i\cdot}\right) + d_{\phi^*}\left(Y_i, \left(Z^T\hat{\nu}\right)_i\right);$$

    Let $Z^{\text{new}}_{ik} = 1(k = k^*)$ for all $k = 1, \ldots, K$
  Let $Z = Z^{\text{new}}$.
until *convergence*;
Return: Node-membership matrix Z

Numerical experiments



Comparison of Algorithm 1 with algorithms of [23] and [24]. Error bars show the standard deviations over 25 runs. Attributes are 2d-spherical Gaussian attributes with radius $(\pm 1, 0)$.
(a) $n = 100$, $K = 2$, $f_{\text{in}} = (1 - p_{\text{in}})\delta_0 + p_{\text{in}}\,\mathrm{Poi}(5)$, $f_{\text{out}} = (1 - 0.03)\delta_0 + 0.03\,\mathrm{Poi}(1)$.
(b) $n = 100$, $K = 2$, $f_{\text{in}} = (1 - 0.07)\delta_0 + 0.07\,\mathrm{Poi}(\mu_{\text{in}})$, $f_{\text{out}} = (1 - 0.04)\delta_0 + 0.04\,\mathrm{Poi}(1)$.

[23] Bo Long, Zhongfei Mark Zhang, and Philip S Yu. A probabilistic framework for relational clustering. In *ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 470–479, 2007.
[24] Mahendra Mariadassou, Stéphane Robin, and Corinne Vacher. Uncovering latent structure in valued graphs: A variational approach. *The Annals of Applied Statistics*, 4(2):715 – 742, 2010.