A novel spectral algorithm for the identification of overlapping communities

Emilie Kaufmann, Thomas Bonald and Marc Lelarge
EMILIE.KAUFMANN@INRIA.FR, THOMAS.BONALD@TELECOM-PARISTECH.FR, MARC.LELARGE@ENS.FR
Inria and Ecole Normale Supérieure, Telecom ParisTech

1. Introduction

Spectral algorithms are popular methods for finding a partition of a network into groups of nodes that are densely connected, called communities. This contribution presents *combinatorial spectral clustering* (CSC), a simple spectral algorithm that can handle overlapping communities. The algorithm is motivated by a random graph model called *stochastic blockmodel with overlap* (SBMO), under which it is proved to be consistent.

A network with n nodes is drawn from the SBMO if its observed adjacency matrix $\hat{A}$ satisfies $\mathbb{E}[\hat{A}] = A$, with

$$A = \frac{\alpha_n}{n} Z B Z^T, \qquad (1)$$

for K the number of communities, and $B \in \mathbb{R}^{K\times K}$, $Z \in \{0,1\}^{n\times K}$ two full rank matrices. α_n controls the sparsity of the graph and tends slowly to infinity. Each row $Z_{i,\cdot}$ is a membership vectors indicating the communities to which node i belongs. The SBMO generalizes the stochastic blockmodel (in which $\forall i, ||Z_{i,\cdot}|| = 1$). It is reminiscent of other models proposed in the literature to describe overlapping communities, but has never been introduced in this form, to the best of our knowledge. Special features include the binary membership vectors, and the fact that B and Z are fixed and not drawn from some distribution.

2. Combinatorial spectral clustering

Let U be a $\mathbb{R}^{n\times K}$ matrix whose columns $u_1, \dots, u_K$ are independent normalized eigenvectors associated to the K non-zero eigenvalues of matrix A in (1) . As the (u_k) form a basis of $\mathrm{Im}(A) \subseteq \mathrm{Im}(Z)$, there exists $C \in \mathbb{R}^{K\times K}$ such that $U = ZC$. This motivates the following estimation procedure for Z, based on an observed adjacency matrix $\hat{A}$ drawn under the SBMO:

$$(\hat{C}, \hat{Z}) = \underset{\substack{C'\in R^{K\times K},\ Z'\in\{0,1\}^{n\times K}:\\ \forall i, 1\le ||Z'_{i,\cdot}|| \le O_{\max}}}{\operatorname{argmin}} ||\hat{U} - Z'C'||_F^2, \qquad (2)$$

where $\hat{U}$ is the matrix of K leading eigenvectors of $\hat{A}$ and $O_{\max}$ is an upper bound on the maximal number of communities to which a node belongs (it can be set to K). If K is unknown, for $\eta, r > 0$ some parameters, we let $\hat{U}$ be a matrix of $\hat{K}$ eigenvectors of $\hat{A}$ associated to eigenvalues λ satisfying $|\lambda| \ge \sqrt{2\,(1+\eta)\,\hat{d}_{\max}\log(4n)}$, where $\hat{d}_{\max}$ is the maximal degree in the observed network.

3. Theoretical results

Let $\mathcal{T} = \{z \in \{0,1\}^{1\times K} : \exists i \in \{1,n\} : Z_{i,\cdot} = z\}$ be the set of mixtures of communities in the network. We assume that there exists $\epsilon > 0 : \forall z \in \mathcal{T}, |\{i : Z_{i,\cdot} = z\}|/n \ge \epsilon$. For every $r > 0$, if $\alpha_n \ge d_0(\log n)^{1+r}$, for some constant d_0 that depends only on B, on the fractions of pairwise overlaps between communities and on the parameters η, we exhibit a set $\mathcal{N}_n$ such that, for n large enough, $\exists \sigma \in \mathfrak{S}_K$: $\forall i \in \mathcal{N}_n,\ \forall k \in \{1,\dots,K\}, \quad \hat{Z}_{i,\sigma(k)} = Z_{i,k}$.

We give an upper bound on the cardinality of $\mathcal{N}_n^c$, showing that $|\mathcal{N}_n^c|/n$ goes to zero with high probability. It permits also to prove the consistency of the method under the SBMO : if P_σ is the permutation matrix associated to σ,

$$\inf_{\sigma\in\mathfrak{S}_K} \frac{1}{nK} \left|\left|\hat{Z}P_\sigma - Z\right|\right|_F \xrightarrow[n\to\infty]{a.s.} 0.$$

4. Implementation and experimental results

In practice, we use alternate minimization to approximate the solution of (2).The minimization in Z' consists in letting the i-th row be $\operatorname{argmin}_{z\in\{0,1\}^{1\times K}} ||zC' - \hat{U}_{i,\cdot}||$ (with the extra condition $1 \le ||z|| \le O_{\max}$). The minimization in C' has a closed form: $C' = (Z'^T Z')^{-1} Z'^T \hat{U}$. We use a k-means++ initialization, and let the first centroid be $\hat{U}_{i,\cdot}$, where i is a random node whose degree is smaller than the median degree in the network.

For networks generated from SBMOs with $n = 500$, $K = 5$, $\alpha_n = \log^{1.5} n$ that have a variable fraction p of nodes belonging to a single community, we compare the CSC algorithm to normalized spectral clustering (SC) and to another spectral method recently proposed by (Zhang et al., 2014) (OC. for OCCAM), in terms of the fraction of wrong entries in $\hat{Z}$ (up to a columns permutation). CSC appears to perform best, even with a small fraction of 'pure' nodes.

p	0.1	0.2	0.25	0.5	0.7	0.9
SC	0.300	0.246	0.229	0.148	0.088	0.030
OC.	0.283	0.261	0.099	0.054	0.046	0.003
CSC	0.053	0.054	0.045	0.032	0.014	0.003

References

Zhang, Y., Levina, E., and Zhu, J. Detecting Overlapping Communities in Networks with Spectral Methods. *arXiv:1412.3432v1*, 2014.