CASE-CONTROL VARIATIONAL INFERENCE FOR LARGE SCALE STOCHASTIC BLOCK MODELS

Silvia Pandolfi [1], Francesco Bartolucci [1]

[1] Department of Economics, University of Perugia, IT (e-mail: silvia.pandolfi@unipg.it, francesco.bartolucci@unipg.it)

ABSTRACT: A scalable variational inference approach for stochastic block models is proposed. The approach is based on a case-control approximation of the likelihood function, which is an unbiased estimator of the full likelihood. Using the case-control likelihood under a variational inference perspective allows us to strongly reduce the computational complexity, making model estimation feasible for large networks. We evaluate the performance of the proposed algorithm using both simulated and real data coming from a Facebook derived social network.

KEYWORDS: clustering, EM algorithm, random graphs, subsampling.

1 Introduction

Stochastic block models (SBMs; e.g. Snijders & Nowicki, 1997) represent a powerful tool for modeling social network data that can discover communities and clusters of nodes according to their social behavior. Under this formulation, the nodes in the network are assumed to belong to a finite number of latent blocks, identified by individual-specific discrete latent variables, with the probability of connection between two nodes only depending on their block membership.

The predominant method of inference for these models is based on a variational approximation of the model log-likelihood (Daudin *et al.*, 2008). However, the complexity of the corresponding estimation algorithm, keeping the number of blocks fixed, is of the order of $O(n^2)$, where n is the number of nodes. This implies that model estimation is computationally intractable for large-scale networks, limiting its use to a narrow range of applications.

Here, following a previous approach (Roy *et al.*, 2019), we propose a case-control approximation of the target function maximized under the variational inference approach, which leads to a strong reduction of the computational complexity, so that the resulting estimation algorithm may be efficiently applied to large networks. The effectiveness of our proposal will be illustrated via simulation and through a real data application.

2 Stochastic block models

Let $\boldsymbol{Y}$ denote an adjacency matrix referred to n nodes and whose generic element, Y_{ij}, is a binary random variable that is equal to 1 if there is an edge between nodes i and j and to 0 otherwise; $\boldsymbol{y}$ and $y_{ij}, i, j = 1, \ldots, n$, are used to denote the realizations of $\boldsymbol{Y}$ and Y_{ij}, respectively. We focus on *binary undirected* networks with no self-loops, leading to a symmetric adjacency matrix with missing values on the main diagonal.

SBMs assume that nodes in the network belong to one out of k distinct unobserved blocks; these are described by means of independent and identically distributed, node-specific, latent variables U_i, $i = 1, \ldots, n$, defined over the discrete support $\{1, \ldots, k\}$ with probabilities $p(U_i = u) = \pi_u$, $u = 1, \ldots, k$.

SBMs also postulate a *local independence assumption* between nodes: conditional on the latent variables U_i and U_j, responses Y_{ij} are assumed to be independent Bernoulli random variables with success probabilities given by $\phi_{uv} = p(Y_{ij} = 1 | U_i = u, U_j = v)$. Therefore, the conditional distribution of Y_{ij} only depends on the block memberships of nodes involved in the relation. Moreover, parameters ϕ_{uv} must satisfy the invariance property with respect to *reflection*, that is, $\phi_{uv} = \phi_{vu}$ for all $u < v$.

2.1 Classical variational inference

Let $\boldsymbol{\theta}$ denote the vector of all model parameters. For parameter estimation, we may rely on the maximization of the following likelihood function:

$$\mathcal{L}(\boldsymbol{\theta}) = p(\boldsymbol{y}) = \sum_{\boldsymbol{u}} p(\boldsymbol{y}|\boldsymbol{u}) p(\boldsymbol{u}), \tag{1}$$

where $\boldsymbol{u} = (u_1, \ldots, u_n)'$ is a realization of the random vector $\boldsymbol{U} = (U_1, \ldots, U_n)'$, and

$$p(\boldsymbol{y}|\boldsymbol{u}) = \sqrt{\prod_{i \leq n} \prod_{j \neq i} p(y_{ij}|u_i, u_j)}, \quad p(\boldsymbol{u}) = \prod_{i \leq n} \pi_{u_i}.$$

As known, the likelihood function in equation (1) involves summation over the configurations of all latent variables in the model, so that the computational burden is prohibitive also when dealing with networks of a very limited size. Moreover, also the posterior expectation of the complete data log-likelihood, which is used within the Expectation-Maximization (EM) algorithm, is intractable. Therefore, a classical solution is to rely on a variational approximation of the EM algorithm (VEM; Daudin *et al.*, 2008), which is based on

the maximization of the following lower-bound of the likelihood function in equation (1):

$$\mathcal{J}(\boldsymbol{\theta}) = \log \mathcal{L}(\boldsymbol{\theta}) - KL[R(\boldsymbol{u}) \,||\, p(\boldsymbol{u}|\boldsymbol{y})], \tag{2}$$

where $p(\boldsymbol{u}|\boldsymbol{y})$ denotes the (intractable) posterior distribution of the latent vector $\boldsymbol{u}$ given the observed adjacency matrix $\boldsymbol{y}$, $R(\boldsymbol{u})$ denotes its approximation, and $KL[\cdot \,||\, \cdot]$ is the Kullback-Leibler divergence between these two distributions. A typical choice for $R(\boldsymbol{u})$ is that based on the conditional independence between the latent variables in the network, given the observed data, implying that $R(\boldsymbol{u}) = \prod_{i \leq n} h(u_i, \boldsymbol{\tau}_i)$, where $h(\cdot, \boldsymbol{\tau}_i)$ denotes a Multinomial probability distribution with parameters 1 and $\boldsymbol{\tau}_i = (\tau_{i1}, \ldots, \tau_{ik})'$. The generic element of $\boldsymbol{\tau}_i$, say τ_{iu}, can be interpreted as an approximation of $p(U_i = u|\boldsymbol{y})$.

Parameter estimates are obtained by alternating two separate steps until convergence of the algorithm. In the variational E-step, $\mathcal{J}(\boldsymbol{\theta})$ is maximized with respect to $\boldsymbol{\tau}_i, i = 1, \ldots, n$, with $\boldsymbol{\theta}$ fixed at the values obtained from the previous iteration, under the constraints that these quantities are non-negative and $\sum_u \tau_{iu} = 1$ for all i. In the variational M-step, $\mathcal{J}(\boldsymbol{\theta})$ is maximized with respect to $\boldsymbol{\theta}$, with the $\boldsymbol{\tau}_i$'s fixed at the values obtained from the E-step.

Besides the several advantages of the variational approximation procedure, the complexity of the iterative algorithm used for deriving parameter estimates, as already mentioned, is of order $O(n^2)$ and this may lead to a excessive computational effort when dealing with large-scale networks.

3 Proposed case-control variational inference

The case-control idea derives from cohort studies where the aim is to compare a group having the outcome of interest ("case") with a control group with regard to one or more characteristics. Usually, the presence of case subjects is relatively rare compared to that of control subjects, and it is impossible or too expensive to select a simple random sample with enough cases to draw conclusions. Accordingly, in a case-control study, all available cases are collected and the corresponding controls are sampled from the corresponding cohort.

In the context of network data, we can view the presence of connections (that is, the 1's) as cases and the absence of connections (the 0's) as controls, and we can rely on this analogy to propose a case-control approximation of the target function in (2). In particular, for every node i, let $\mathcal{A}_i$ denotes the random subset of $\{j : y_{ij} = 0,\ j \neq i\}$, with $n_{i0} = \sum_{j \neq i}(1 - y_{ij})$ being the total number of nodes that are not connected with node i. We also define $\mathcal{B}_i$ as the random subset of $\{j : y_{ij} = 1,\ j \neq i\}$, with $n_{i1} = \sum_{j \neq i} y_{ij}$ being the total number of nodes

connected with i. We may derive the following approximation of $p(\boldsymbol{y}|\boldsymbol{u})$:

$$\tilde{p}(\boldsymbol{y}|\boldsymbol{u}) = \sqrt{\prod_{i\leq n}\left[\left(\prod_{j\in\mathcal{A}_i} p(y_{ij}|u_i,u_j)\right)^{n_{i0}/|\mathcal{A}_i|}\left(\prod_{j\in\mathcal{B}_i} p(y_{ij}|u_i,u_j)\right)^{n_{i1}/|\mathcal{B}_i|}\right]}.$$

Since $\tilde{p}(\boldsymbol{y}|\boldsymbol{u})$ is based on random samples from the 1's and 0's, we get an unbiased estimator of $p(\boldsymbol{y}|\boldsymbol{u})$. The case-control approximate likelihood is then defined as $\tilde{\mathcal{L}}(\boldsymbol{\theta}) = \sum_{\boldsymbol{u}} \tilde{p}(\boldsymbol{y}|\boldsymbol{u})p(\boldsymbol{u})$ and the corresponding lower bound may be derived as in equation (2), leading to the approximate target function $\tilde{\mathcal{J}}(\boldsymbol{\theta})$. Given the assumption of *a posteriori* independence of the latent variables and denoting by $w_{i0} = n_{i0}/|\mathcal{A}_i|$ and $w_{i1} = n_{i1}/|\mathcal{B}_i|$ the sampling rates, we have

$$\begin{aligned}\tilde{\mathcal{J}}(\boldsymbol{\theta}) &= \sum_{\boldsymbol{u}} R(\boldsymbol{u})\log[p(\boldsymbol{u})\tilde{p}(\boldsymbol{y}|\boldsymbol{u})] - \sum_{\boldsymbol{u}} R(\boldsymbol{u})\log R(\boldsymbol{u}) = \sum_{i\leq n}\sum_{u\leq k}\tau_{iu}\log\pi_u \\ &\quad + \frac{1}{2}\sum_{i\leq n}\sum_{u\leq k}\tau_{iu}\left[w_{i0}\sum_{j\in\mathcal{A}_i}\sum_{v}\tau_{jv}\log(1-\phi_{uv}) + w_{i1}\sum_{j\in\mathcal{B}_i}\sum_{v}\tau_{jv}\log\phi_{uv}\right] \\ &\quad - \sum_{i\leq n}\sum_{u\leq k}\tau_{iu}\log\tau_{iu}.\end{aligned}$$

Parameter estimation may be then obtained by means of a modified VEM algorithm that maximizes $\tilde{\mathcal{J}}(\boldsymbol{\theta})$. Denoting by $m < n$ the average number of 1's and 0's selected for each node, the complexity of the proposed estimation algorithm reduces to $O(n \times m)$. For large networks that are usually sparse, we can randomly choose a very small subset of 0's, so as to obtain a strong reduction of the computing time. Moreover, alternative sampling schemes based on descriptive network statistics may also be considered in order to increase the efficiency of the algorithm and the accuracy of the estimates.

References

DAUDIN, J-J., PICARD, F., & ROBIN, S. 2008. A mixture model for random graphs. *Statistics and Computing*, **18**, 173–183.

ROY, S., ATCHADÉ, Y., & MICHAILIDIS, G. 2019. Likelihood inference for large scale stochastic blockmodels with covariates based on a divide-and-conquer parallelizable algorithm with communication. *Journal of Computational and Graphical Statistics*, **28**, 609–619.

SNIJDERS, T.A., & NOWICKI, K. 1997. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, **14**, 75–100.