Computational Statistics and Data Analysis 57 (2013) 661–671

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Variational Bayesian inference for the Latent Position Cluster Model for network data

Michael Salter-Townshend *, Thomas Brendan Murphy *

School of Mathematical Sciences & Complex and Adaptive Systems Laboratory, University College Dublin, Belfield, Dublin 4, Ireland

ARTICLE INFO

Article history:
Received 6 January 2012
Received in revised form 13 July 2012
Accepted 3 August 2012
Available online 11 August 2012

Keywords:
Social network analysis
Variational Bayes
Latent Position Cluster Model

ABSTRACT

A number of recent approaches to modeling social networks have focussed on embedding the nodes in a latent "social space". Nodes that are in close proximity are more likely to form links than those who are distant. This naturally accounts for reciprocal and transitive relationships which are commonly found in many network datasets. The Latent Position Cluster Model is one such model that also explicitly incorporates clustering by modeling the locations using a finite Gaussian mixture model. Observed covariates and sociality random effects may also be modeled. However, inference for the model via MCMC is cumbersome and thus scaling to large networks is a challenge. Variational Bayesian methods offer an alternative inference methodology for this problem. Sampling based MCMC is replaced by an optimization that requires many orders of magnitude fewer iterations to converge. A Variational Bayesian algorithm for the Latent Position Cluster Model is therefore developed and demonstrated.

1. Introduction

Hoff et al. (2002) introduced latent space models for networks. In this model, the nodes are embedded in an unobserved "social space"; nodes closer together are more likely to link than nodes far apart and inference is performed on the latent positions of the nodes. Links are modeled as occurring independently given the positions of the nodes (and optionally any observed link or node covariates). One appealing characteristic of such models is that they naturally account for reciprocity and transitivity. In addition, plotting the inferred positions of the nodes gives an intuitive visualization of the network.

Handcock et al. (2007) proposed the Latent Position Cluster Model (LPCM) which extends the latent space models to allow for model based clustering of the nodes. This is to accommodate the clustering of nodes in the network beyond that expected from simple transitivity. Clustering is thus included explicitly in the model rather than found by a post-hoc analysis of the estimated node locations. A spherical Gaussian mixture model structure is assumed for the latent positions.

1.1. Motivation for using the variational method

Currently, inference for the LPCM is via MCMC in a Bayesian setting. The disadvantage of this approach is computational and fitting the model to large or even medium size network datasets is impractical or impossible. Variational Bayesian inference offers one approximate solution to this problem. A closed form posterior is found that is "close" to the intractable posterior. This method has already been exploited successfully for other fully Bayesian social network models. We next motivate the contribution in this paper, namely to develop and assess Variational Bayesian algorithms for the LPCM.

* Corresponding authors.
E-mail address: michael.salter-townshend@ucd.ie (M. Salter-Townshend).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.08.004

In the published discussion of Handcock et al. (2007), the following contributions are amongst those made:

- David Blei and Stephen E. Feinberg: "We have appealed to variational methods for a computationally efficient approximation to the posterior (for a mixed membership blockmodel). These methods can scale to large matrices because of the simplified approximation (but at an unknown cost to accuracy). It would be interesting to understand computational trade-offs for the authors method (LPCM) as the sample size grows and when large numbers of covariates are added".
- Dirk Husmeier and Chris Glasbey: "Although a full reversible jump Markov chain Monte Carlo scheme might be computationally prohibitive, variational methods, which are currently very popular in the machine learning community, would presumably provide a much better approximation to the integration and might therefore provide a promising avenue for future research".
- David S. Leslie: "I congratulate the authors for their interesting paper. However, it seems that the Markov chain Monte Carlo sampling scheme that was used results in extremely slow mixing, requiring 2 million iterations with only every 1000th iteration being used".

This work is further prompted by Airoldi et al. (2008) when the authors state "It would be interesting to develop a variational algorithm for the latent space models".

1.2. Specification of the latent position cluster model

In the LSM and LPCM, a binary interactions sociomatrix $\boldsymbol{Y}$ is modeled using logistic regression in which the probability of a link between two nodes depends on the distance between the nodes in the latent space.

$$\log - \text{odds}(y_{i,j} = 1 | \underline{z}_i, \underline{z}_j, \beta) = \log\left(\frac{\mathbb{P}\{y_{i,j} = 1\}}{\mathbb{P}\{y_{i,j} = 0\}}\right) = \beta - |\underline{z}_i - \underline{z}_j|, \tag{1}$$

where $y_{i,j} = 1$ if node i links with node j and $y_{i,j} = 0$ otherwise, β is an intercept parameter and $|\underline{z}_i - \underline{z}_j|$ is the Euclidean distance between the latent positions $\underline{z}_i$ and $\underline{z}_j$ of nodes i and j. The links are assumed to be independent conditional on the latent positions of the nodes in the latent space.

Hence, the probability of the observed network $\boldsymbol{Y}$ given the latent positions $\boldsymbol{Z} = (\underline{z}_1, \ldots, \underline{z}_N)$ of all of the nodes is

$$\mathbb{P}(\boldsymbol{Y}|\beta, \boldsymbol{Z}) = \prod_{i=1}^{N} \prod_{\substack{j \neq i \\ j=1}}^{N} \left[\frac{\exp(\beta - |\underline{z}_i - \underline{z}_j|)}{1 + \exp(\beta - |\underline{z}_i - \underline{z}_j|)}\right]^{y_{i,j}} \left[\frac{1}{1 + \exp(\beta - |\underline{z}_i - \underline{z}_j|)}\right]^{(1-y_{i,j})}.$$

Note that if the network is undirected then the product term is taken over $i < j$ instead of $i \neq j$.

For the LPCM, in order to represent clustering of nodes in the network, the latent positions $\boldsymbol{Z}$ are modeled as coming from a mixture of G multivariate normal distributions.

$$\underline{z}_i \sim \sum_{g=1}^{G} \lambda_g \, \text{MVN}_d(\underline{\mu}_g, \sigma_g^2 \mathbf{I}_d), \tag{2}$$

where λ_g is the probability that a node belongs to the gth group and $\mathbf{I}_d$ is the $d \times d$ identity matrix. We let $\underline{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_G)$, $\boldsymbol{\mu} = (\underline{\mu}_1, \underline{\mu}_2, \ldots, \underline{\mu}_G)$ and $\underline{\sigma}^2 = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_G^2)$.

In order to fit the model in a Bayesian setting, the following hierarchical priors are assumed for β, $\underline{\lambda}$, $\underline{\sigma}$ and $\boldsymbol{\mu}$:

$$\underline{\lambda} \sim \text{Dirichlet}(\underline{\nu}), \tag{3}$$

$$\beta \sim \text{Normal}(\xi, \psi^2), \tag{4}$$

$$\underline{\mu}_g \sim \text{MVN}_d(0, \omega^2 \mathbf{I}_d), \tag{5}$$

and

$$\sigma_g^2 \sim \sigma_0^2 \text{ Inverse } \chi_\alpha^2 \tag{6}$$

where the values ξ, ψ^2, $\underline{\nu}$, σ_0^2, α and ω^2 are fixed prior hyper-parameters.

Hence, the posterior of the latent positions and the model parameters is given by,

$$p(\boldsymbol{Z}, \underline{\lambda}, \beta, \boldsymbol{\mu}, \underline{\sigma}^2 | \boldsymbol{Y}) = Cp(\boldsymbol{Y}|\beta, \boldsymbol{Z}) p(\boldsymbol{Z}|\underline{\lambda}, \boldsymbol{\mu}, \underline{\sigma}^2 \mathbf{I}_d) p(\underline{\lambda}|\underline{\nu}) p(\beta|\xi, \psi^2) p(\boldsymbol{\mu}|0, \omega^2 \mathbf{I}_d) p(\underline{\sigma}^2|\sigma_0^2, \alpha) \tag{7}$$

where the proportionality constant C is unknown and therefore the posterior is only known up to proportionality.

1.3. Variational Bayesian inference

We develop a Variational Bayesian inference procedure for approximating the posterior distribution of the latent variables in the LPCM. This approach facilitates the application of the LPCM to larger networks than is currently possible