Model-Based Clustering of Large Random Graphs With High-Dimensional Predictors

Duy Q. Vu
Department of Mathematics and Statistics
University of Melbourne
duy.vu@unimelb.edu.au

Michael Schweinberger
Department of Statistics
Rice University
m.s@rice.edu

Abstract

Model-based clustering of random graphs has attracted much attention in machine learning and statistics. We improve the predictive power of model-based clustering by exploiting high-dimensional predictors and encouraging sparsity by using a penalized Bayesian approach. We estimate the high-dimensional posterior by using variational methods. To implement variational methods, we exploit the minorization-maximization (MM) principle to construct minorizing functions which separate the parameters and facilitate the exploration of the high-dimensional posterior. The resulting Bayesian Variational GEM-MM algorithm is less prone to be trapped at local maxima than conventional Variational EM algorithms. We demonstrate the success of the model and method by using two real-world networks with more than 10,000 nodes and 15,000 predictors.

1 Introduction

Data which can be represented by random graphs arise in computer science (e.g., social networks, World Wide Web), engineering (e.g., power networks), the life sciences (e.g., biological networks), the health sciences (e.g., sexual networks), and the social sciences (e.g., terrorist networks). An attractive approach to modeling random graphs model is based on the assumption that edge variables are exchangeable. An extension of de Finetti's theorem to random graphs suggests that exchangeable edge variables can be modeled by mixture models [1]. A simple mixture model is a finite mixture model, which assumes that edges are governed by a finite mixture of distributions. Such models are known as stochastic block models [2, 3].

Scalable methods to estimate stochastic block models are variational methods, which give rise to approximate maximum likelihood estimates [4, 5, 6, 7]. An alternative is based on spectral clustering [8, 9, 10, 11], but spectral clustering neither allows users to specify the features of the data on the basis of which nodes are to be clustered nor allows users to incorporate available predictors. In contrast, variational methods can be used to estimate a wide range of models capturing interesting features of real-world networks and incorporating available predictors. The consistency and asymptotic normality of variational estimators was established by [12, 13]. An important, remaining challenge stems from the fact that the high-dimensional posterior and its variational approximations tend to be non-concave and multimodal and algorithms tend to be trapped at local maxima.

We improve the predictive power of model-based clustering by introducing a novel model which incorporates available predictors and address the non-concave and multimodal nature of the high-dimensional posterior and its variational approximations. In most applications, additional information in the form of covariates is available, but—despite the fact that covariates may improve the predictive power of models—it is common practice to discard them. A notable exception is [14] and related work on latent variable models by [15, 16, 17], but the work of [14] and others relies on Bayesian Markov chain Monte Carlo methods which are not scalable and therefore have been

limited to small networks with up to 200 nodes. We introduce a novel model which incorporates available covariates and allows covariate parameters to depend on blocks, thus allowing a covariate to be active in some blocks while being inactive in others and varying in strength across the blocks in which it is active. We address the non-concave and multimodal nature of the high-dimensional posterior and its variational approximations by using the MM principle [18, 19] to derive a Bayesian Variational GEM-MM algorithm. We demonstrate the success of the model and method by using two data sets with more than 10,000 nodes and 15,000 block-dependent covariate terms.

The paper is structured as follows. We introduce a novel model with predictors in Section 2, discuss a penalized Bayesian approach in Section 3, and introduce a novel Bayesian Variational GEM-MM in Section 4. Experiments are presented in Section 5.

2 Data-generating process

We extend stochastic block models to include predictors and model both the sequence of degrees and block-dependent predictors.

We assume that the set of nodes is partitioned into K subsets, called blocks, and that the block memberships are unobserved. We assume that node-dependent categorical covariates are available and denote by $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ the set of covariate vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of nodes $1, \ldots, n$. We denote by $Y_{i,j}$ undirected edge variables, where $Y_{i,j}$ means that there is an undirected edge between nodes i and j and $Y_{i,j} = 0$ otherwise. Since edge variables are undirected and self-edges are meaningless, we assume that $Y_{i,j} = Y_{j,i}$ and $Y_{i,i} = 0$ with probability 1.

The data-generating process can be described as follows.

1. Generate mixing probabilities:

$$\alpha_1, \ldots, \alpha_K \mid \omega_1, \ldots, \omega_K \quad \sim \quad \text{Dirichlet}(\omega_1, \ldots, \omega_K). \tag{1}$$

2. Generate block membership indicators:

$$Z_{i1}, \ldots, Z_{iK} \mid \alpha_1, \ldots, \alpha_K \quad \overset{\text{iid}}{\sim} \quad \text{Multinomial}(1; \alpha_1, \ldots, \alpha_K). \tag{2}$$

3. Generate edge variables:

$$Y_{ij} \mid Z_{ik} = 1, Z_{jl} = 1, \boldsymbol{x} \quad \overset{\text{iid}}{\sim} \quad \text{Bernoulli}(\pi_{y_{ij};kl}(\boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{x})), \tag{3}$$

where

$$\log\left[\frac{\pi_{y_{ij};kl}(\boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{x})}{1 - \pi_{y_{ij};kl}(\boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{x})}\right] \quad = \quad \delta_{kl} + \boldsymbol{\beta}_{kl}^{\top}\, \boldsymbol{f}(x_i, x_j),$$

where δ_{kl} are block-dependent degree parameters, $\boldsymbol{\beta}_{kl}$ are block-dependent covariate parameters, and $\boldsymbol{f}$ are functions of covariates.

The model, in its full generality, is not an attractive model of large random graphs for at least two reasons.

First, estimating and interpreting $O(K^2)$ block-dependent degree parameters δ_{kl} is challenging when K is large. It is desirable to impose constraints to facilitate estimation and interpretation [7]. We consider here constraints of the form

$$\delta_{kl} \quad = \quad \delta_k + \delta_l, \tag{4}$$

where δ_k and δ_l are the degree parameters of blocks k and l, respectively, implying that there are $O(K)$ rather than $O(K^2)$ degree parameters.

Second, and worse, the number of block-dependent covariate parameters $\boldsymbol{\beta}_{kl}$ is $O(p\, K^2)$, where p is the dimension of the covariate vectors. To deal with the large number of block-dependent covariate parameters, we encourage sparsity by using a penalized Bayesian approach.

3 Penalized Bayesian approach

A Bayesian approach has advantages over non-Bayesian approaches, being able to deal with parameter uncertainty and average with respect to nuisance parameters, such as the mixing proportions $\alpha_1, \ldots, \alpha_K$.

We discuss here a penalized Bayesian approach which encourages sparsity. In Bayesian fashion, the penalty is imposed through the prior. We encourage most covariate-related parameters to be 0 by using a spike-and-slab prior of the following form:

- Bock-dependent degree parameters: $\delta_k \mid \mu_d, \sigma_d^2 \overset{\text{iid}}{\sim} N(\mu_d, \sigma_d^2)$.
- Block-dependent covariate parameters: $\beta_{kl}^s \mid \tau^2, I_{kl}^s \overset{\text{iid}}{\sim} \left[N(0,\tau^2)\right]^{1-I_{kl}^s} + \left[N(0, g\,\tau^2)\right]^{I_{kl}^s}$, where $g \gg 1$ is given.
- Block-dependent activeness indicators: $I_{kl}^s \mid \gamma_{kl} \overset{\text{iid}}{\sim} \text{Bernoulli}(\gamma_{kl})$.
- Block-dependent activeness expectation: $\gamma_{kl} \mid a_\gamma, b_\gamma \overset{\text{iid}}{\sim} \text{Beta}(a_\gamma, b_\gamma)$.
- Variance parameter: $\tau^2 \sim \text{Inverse-Gamma}(a_{\tau^2}, b_{\tau^2}^{-1})$.

That is, the covariate parameters are governed by a two-component mixture distribution, where both component distributions are Gaussians centered at 0, one having a small variance τ^2 (spike at 0) and the other having a large variance $g\,\tau^2$ with $g \gg 1$ (slab). The indicator I_{kl}^s is 1 if covariate s is active in pair of blocks $\{k, l\}$ and 0 otherwise. The idea is that the posterior mass of inactive covariate parameters is concentrated in a small neighborhood of 0. The smaller γ_{kl}, the more covariates parameters are expected to be inactive. As a result, a covariate can be active in some blocks while being inactive in others and varying in strength across the blocks in which it is active.

Last, since there is uncertainty about γ_{kl}, we place a prior on γ_{kl}. Likewise, we place a prior on the variance parameter τ^2.

4 Bayesian Variational GEM-MM algorithm

We explore the posterior by variational methods. While variational methods are scalable methods, implementing variational methods in high-dimensional settings can be challenging, as discussed in Sections 4.1 and 4.2. We address these problems by exploiting the MM principle [18, 19] to derive a Bayesian Variational GEM-MM algorithm which implements both the E-step and M-step of a Variational EM algorithm by MM algorithms.

We denote the parameters by $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{Z}, \boldsymbol{\delta}, \tau^2, \boldsymbol{\gamma}, \boldsymbol{I}, \boldsymbol{\beta})$. To keep the notation manageable, we denote the probability mass function of $\boldsymbol{y}$ by $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ and suppress the notational dependence on constants, including covariates and hyperparameters. Bayesian inference is based on the posterior density of $\boldsymbol{\theta}$ of the form

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \quad = \quad \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\boldsymbol{y})}, \tag{5}$$

where $p(\boldsymbol{\theta})$ is the prior density of $\boldsymbol{\theta}$ and $p(\boldsymbol{y})$ is the marginal likelihood of $\boldsymbol{y}$. The posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ is intractable, because the marginal likelihood $p(\boldsymbol{y})$ is intractable. A tractable lower bound on the intractable marginal likelihood $p(\boldsymbol{y})$ can be derived by using Jensen's inequality, giving

$$\log p(\boldsymbol{y}) \quad = \quad \log \int \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} q(\boldsymbol{\theta})\, \mathrm{d}\, \boldsymbol{\theta} \quad \geq \quad \int \left[\log \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right] q(\boldsymbol{\theta})\, \mathrm{d}\, \boldsymbol{\theta}, \tag{6}$$

where Q is an auxiliary distribution with density q with the same support as the posterior. We denote the lower bound LB_1 in (6) by

$$LB_1(q) \quad = \quad \mathbb{E}_Q \left[\log p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\right] - \mathbb{E}_Q \left[\log q(\boldsymbol{\theta})\right] \tag{7}$$

and assume that the auxiliary distribution belongs to a family of fully factorized distributions of the form

$$\begin{aligned} q(\boldsymbol{\theta} \mid \widetilde{\boldsymbol{\theta}}) &= q(\boldsymbol{\alpha} \mid \widetilde{\boldsymbol{\alpha}}) \prod_{i=1}^{n} q(\boldsymbol{Z}_i \mid \widetilde{\boldsymbol{Z}}_i) \prod_{k=1}^{K} q(\delta_k \mid \widetilde{\mu}_{\delta_k}, \widetilde{\sigma}^2_{\delta_k}) \; q((\tau^2)^{-1} \mid \widetilde{a}_{\tau^2}, \widetilde{b}_{\tau^2}) \\ & \prod_{k=1}^{K} \prod_{l=1}^{k} q(\gamma_{kl} \mid \widetilde{a}_{\gamma_{kl}}, \widetilde{b}_{\gamma_{kl}}) \prod_{k=1}^{K} \prod_{l=1}^{k} \prod_{s=1}^{p} q(I^s_{kl} \mid \widetilde{I}^s_{kl}) \prod_{k=1}^{K} \prod_{l=1}^{k} \prod_{s=1}^{p} q(\beta^s_{kl} \mid \widetilde{\mu}_{\beta^s_{kl}}, \widetilde{\sigma}^2_{\beta^s_{kl}}), \end{aligned}$$

where $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{Z}}, \widetilde{\boldsymbol{\mu}}_{\boldsymbol{\delta}}, \widetilde{\boldsymbol{\sigma}}^2_{\boldsymbol{\delta}}, \widetilde{\boldsymbol{a}}_{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{b}}_{\boldsymbol{\gamma}}, \widetilde{a}_{\tau^2}, \widetilde{b}_{\tau^2}, \widetilde{\boldsymbol{I}}, \widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}})$ denotes the parameters of the auxiliary distribution.

To make the lower bound as tight as possible—which is equivalent to minimizing the Kullback-Leibler divergence from $q(\boldsymbol{\theta} \mid \widetilde{\boldsymbol{\theta}})$ to $p(\boldsymbol{\theta} \mid \boldsymbol{y})$—we maximize the lower bound with respect to $\widetilde{\boldsymbol{\theta}}$.

To maximize the lower bound with respect to $\widetilde{\boldsymbol{\theta}}$, we need closed-form expressions of the expectations in (7) under Q. However, the expectation

$$\begin{aligned} \mathbb{E}_Q \left[\log \pi_{y_{ij};kl}(\boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{x}) \right] &= \left[\widetilde{\mu}_{\delta_k} + \widetilde{\mu}_{\delta_l} + \sum_{s=1}^{p} \widetilde{\mu}_{\beta^s_{kl}} f_s(x_i, x_j) \right] y_{ij} \\ &+ \mathbb{E}_Q \left[-\log \left(1 + \exp[\delta_k + \delta_l + \boldsymbol{\beta}^\top_{kl} \, \boldsymbol{f}(x_i, x_j)] \right) \right] \end{aligned}$$

is intractable, because the second expectation on the right-hand side is intractable. We replace the intractable expectation by a tractable lower bound obtained by Jensen's inequality, which implies that $\mathbb{E}_Q[-\log(.)] \geq -\log[\mathbb{E}_Q(.)]$ and therefore that $\mathbb{E}_Q[-\log(.)]$ is bounded below by $-\log\left[\Phi_{ij;kl}(.)\right]$, where

$$\Phi_{ij;kl}(\boldsymbol{x}) = 1 + \exp\left[\widetilde{\mu}_{\delta_k} + \frac{1}{2}\widetilde{\sigma}^2_{\delta_k} + \widetilde{\mu}_{\delta_l} + \frac{1}{2}\widetilde{\sigma}^2_{\delta_l} + \sum_{s=1}^{p} \widetilde{\mu}_{\beta^s_{kl}} f_s(x_i, x_j) + \frac{1}{2} \sum_{s=1}^{p} \widetilde{\sigma}^2_{\beta^s_{kl}} f^2_s(x_i, x_j) \right].$$

Maximizing the resulting lower bound, denoted by LB_2, is feasible, but the maximization problem is high-dimensional. We facilitate the high-dimensional maximization problem by exploiting the MM principle. The MM principle reduces the high-dimensional maximization problem to low-dimensional maximization problems which are straightforward to solve. The resulting Bayesian Variational GEM-MM algorithm increases the lower bound LB_2 at each iteration. We sketch the Bayesian Variational GEM-MM algorithm in Table 1 and provide details in Sections 4.1 and 4.2.

4.1 E-step: MM algorithm

The E-step maximizes LB_2 with respect to $\widetilde{\boldsymbol{Z}}_1, \ldots, \widetilde{\boldsymbol{Z}}_n$, where

$$LB_2(q) = \sum_{i=1}^{n} \sum_{k=1}^{K} \widetilde{Z}_{ik} \left[\Psi\left(\widetilde{\alpha}_k\right) - \Psi\left(\sum_{l=1}^{K} \widetilde{\alpha}_l \right) - \log \widetilde{Z}_{ik} \right] + \sum_{i<j}^{N} \sum_{k=1}^{K} \sum_{l=1}^{K} \widetilde{Z}_{ik} \, \widetilde{Z}_{jl} \, \Upsilon_{y_{ij};kl}(\boldsymbol{x}) + \text{const},$$

where $\Upsilon_{y_{ij};kl}(.)$ is the lower bound obtained of $\mathbb{E}_Q \left[\log \pi_{y_{ij};kl}(.) \right]$ discussed above.

A simple approach to maximizing LB_2 with respect to $\widetilde{\boldsymbol{Z}}_i$ is based on fixed-point (FP) updates of the form

$$\log \widetilde{Z}^{(t+1)}_{ik} \propto \Psi\left(\widetilde{\alpha}^{(t)}_k\right) - \Psi\left(\sum_{l=1}^{K} \widetilde{\alpha}^{(t)}_l \right) + \sum_{j \neq i}^{N} \sum_{l=1}^{K} \widetilde{Z}^{(t)}_{jl} \Upsilon^{(t)}_{y_{ij};kl}(\boldsymbol{x}),$$

where t denotes the iteration number. However, FP updates can be slow and are prone to be trapped at local maxima. Therefore, we leverage the MM principle [18, 19], which is less prone to be trapped at local maxima [7].

The MM principle exploits properties of the objective function to derive a simple-to-maximize minorizing function and then maximizes the minorizing function. Here, the objective function is the lower bound LB_2, which we want to maximize as a function of $\widetilde{\boldsymbol{Z}}_1, \ldots, \widetilde{\boldsymbol{Z}}_n$. By exploiting the log

Table 1: Sketch of Bayesian Variational GEM-MM algorithm

Minimize the Kullback-Leibler divergence from $q(\boldsymbol{\theta} \mid \widetilde{\boldsymbol{\theta}})$ to $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ by cycling through the components of $\widetilde{\boldsymbol{\theta}}$ as follows:

E-step:

E.1 Update $\widetilde{\boldsymbol{Z}}_1, \ldots, \widetilde{\boldsymbol{Z}}_n$ using the E-step MM algorithm described in Section 4.1.

E.2 Update $\widetilde{\alpha}_1, \ldots, \widetilde{\alpha}_K$:

$$\widetilde{\alpha}_k \quad = \quad \omega_k + \sum_{i=1}^{n} \widetilde{Z}_{ik}.$$

M-step:

M.1 Update $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\delta}}$, $\widetilde{\boldsymbol{\sigma}}^2_{\boldsymbol{\delta}}$, $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}$, $\widetilde{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}}$ using the M-step MM algorithm described in Section 4.2.

M.2 Update $\widetilde{a}_{\tau^2}, \widetilde{b}_{\tau^2}$:

$$\widetilde{a}_{\tau^2} \quad = \quad a_{\tau^2} + \frac{p\,K\,(K+1)}{4}$$

$$\widetilde{b}_{\tau^2} \quad = \quad b_{\tau^2} + \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{k}\sum_{s=1}^{p}(1-\widetilde{I}^s_{kl})\,(\widetilde{\mu}^2_{\beta^s_{kl}} + \widetilde{\sigma}^2_{\beta^s_{kl}}) + \frac{1}{2\,g}\sum_{k=1}^{K}\sum_{l=1}^{k}\sum_{s=1}^{p}\widetilde{I}^s_{kl}\,(\widetilde{\mu}^2_{\beta^s_{kl}} + \widetilde{\sigma}^2_{\beta^s_{kl}}).$$

M.3 Update $\widetilde{\boldsymbol{a}}_{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{b}}_{\boldsymbol{\gamma}}$:

$$\widetilde{a}_{\gamma_{kl}} = a_\gamma + \sum_{s=1}^{p}\widetilde{I}^s_{kl}$$

$$\widetilde{b}_{\gamma_{kl}} = b_\gamma + \sum_{s=1}^{p}(1-\widetilde{I}^s_{kl}).$$

M.4 Update $\widetilde{\boldsymbol{I}}$:

$$\widetilde{I}^s_{kl} = \left\{1 + \exp\left[\psi(\widetilde{b}_{\gamma_{kl}}) - \psi(\widetilde{a}_{\gamma_{kl}}) - \frac{1}{2}\frac{\widetilde{a}_{\tau^2}}{\widetilde{b}_{\tau^2}}\left(\widetilde{\mu}^2_{\beta^s_{kl}} + \widetilde{\sigma}^2_{\beta^s_{kl}}\right)\left(1-\frac{1}{g}\right) + \frac{1}{2}\log g\right]\right\}^{-1}.$$

concavity and arithmetic-geometric mean inequality [19], we construct a minorizing function for LB_2 of the form:

$$\begin{aligned} M_1(\widetilde{\boldsymbol{Z}}_1, \ldots, \widetilde{\boldsymbol{Z}}_n) \quad &= \quad \sum_{i=1}^{n}\left[\sum_{k=1}^{K}\widetilde{Z}^2_{ik}\left(-\frac{1}{\widetilde{Z}^{(t)}_{ik}} + \sum_{j\neq i}\sum_{l=1}^{K}\frac{\widetilde{Z}^{(t)}_{jl}}{2\widetilde{Z}^{(t)}_{ik}}\Upsilon^{(t)}_{y_{ij};kl}(.)\right)\right. \\ &+ \quad \left.\sum_{k=1}^{K}\widetilde{Z}_{ik}\left(\Psi(\widetilde{\alpha}^{(t)}_k) - \Psi(\sum_{k'}\widetilde{\alpha}^{(t)}_{k'}) - \log\widetilde{Z}^{(t)}_{ik} + 1\right)\right]. \end{aligned}$$

As a result, the high-dimensional maximization problem is replaced by$K$-dimensional quadratic programming problems which are straightforward to maximize [20].

4.2 M-step: MM algorithm

The M-step maximizes the lower bound LB_2 with respect to the parameters $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\delta}}$, $\widetilde{\boldsymbol{\sigma}}^2_{\boldsymbol{\delta}}$, $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}$, and $\widetilde{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}}$, where

$$\begin{aligned} LB_2(q) \quad &= \quad \sum_{k=1}^{K}\left(\frac{1}{2}\log\widetilde{\sigma}^2_{\delta_k} - \frac{\widetilde{\mu}^2_{\delta_k} + \widetilde{\sigma}^2_{\delta_k} - 2\widetilde{\mu}_{\delta_k}\mu_d}{2\sigma^2_d}\right) + \sum_{k\leq l}\sum_{s=1}^{p}\frac{1}{2}\log\widetilde{\sigma}^2_{\beta^s_{kl}} \\ &+ \quad \sum_{k\leq l}^{K}\sum_{s=1}^{p}\frac{1}{2}\frac{\widetilde{a}_{\tau^2}}{\widetilde{b}_{\tau^2}}(\widetilde{\mu}^2_{\beta^s_{kl}} + \widetilde{\sigma}^2_{\beta^s_{kl}})\left[\widetilde{I}^s_{kl}\left(1-\frac{1}{g}\right) - 1\right] + \sum_{k=1}^{K}\widetilde{\mu}_{\delta_k}\sum_{i=1}^{n}\widetilde{Z}_{ik}d_i(\boldsymbol{y}) \\ &+ \quad \sum_{1\leq i<j\leq N}\sum_{k=1}^{K}\sum_{l=1}^{K}\widetilde{Z}_{ik}\widetilde{Z}_{jl}\left(\sum_{s=1}^{p}\widetilde{\mu}_{\beta^s_{kl}}y_{ij}f_s(x_i,x_j) - \log\Phi_{ij;kl}(\boldsymbol{x})\right) + \text{const}, \end{aligned}$$

where $d_1(\boldsymbol{y}), \ldots, d_n(\boldsymbol{y})$ denote the degrees of nodes $1, \ldots, n$, respectively.

To implement the M-step, we exploit the MM principle, because the large number of parameters makes gradient ascent and Newton's method problematic: The matrix inversion required by Newton's methods is expensive, while coordinate ascent methods may neither be stable nor straightforward to implement [19]. In contrast, MM updates do not require matrix inversion, because all parameters are separated, and are robust and straightforward to implement. A second reason is that increasing rather than maximizing objective functions can speed up convergence [21, 19],

By using log concavity and exponential convexity inequality [19], we construct a minorizing function for LB_2 of the form:

$$\begin{aligned} M_2(\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\delta}}, \widetilde{\boldsymbol{\sigma}}^2_{\boldsymbol{\delta}}, \widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}}) &= \sum_{k=1}^{K} M_{2,\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\delta}}}(\widetilde{\mu}_{\delta_k}) + \sum_{k=1}^{K} M_{2,\widetilde{\boldsymbol{\sigma}}^2_{\boldsymbol{\delta}}}(\widetilde{\sigma}^2_{\delta_k}) \\ &+ \sum_{k\le l}^{K}\sum_{s=1}^{p} M_{2,\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}}(\widetilde{\mu}_{\beta^s_{kl}}) + \sum_{k\le l}^{K}\sum_{s=1}^{p} M_{2,\widetilde{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}}}(\widetilde{\sigma}^2_{\beta^s_{kl}}) + \text{const}, \end{aligned}$$

where

$$\begin{aligned} Q_{\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\delta}}}(\widetilde{\mu}_{\delta_k}) &= -\frac{\widetilde{\mu}^2_{\delta_k}}{2\sigma^2_d} + \widetilde{\mu}_{\delta_k}\left(\frac{\mu_d}{\sigma^2_d} + \sum_{i=1}^{n} \widetilde{Z}_{ik}\, d_i(\boldsymbol{y})\right) \\ &\quad - \sum_{i=1}^{n}\sum_{j\neq i}\sum_{l=1}^{K} \widetilde{Z}_{ik}\,\widetilde{Z}_{jl}\,\frac{\Phi^{(t)}_{ij;kl}(\boldsymbol{x}) - 1}{(2\,p+4)\,\Phi^{(t)}_{ij;kl}(\boldsymbol{x})} \exp\left[(2\,p+4)(\widetilde{\mu}_{\delta_k} - \widetilde{\mu}^{(t)}_{\delta_k})\right] \\ Q_{\widetilde{\boldsymbol{\sigma}}^2_{\boldsymbol{\delta}}}(\widetilde{\sigma}^2_{\delta_k}) &= \frac{1}{2}\log\widetilde{\sigma}^2_{\delta_k} - \frac{\widetilde{\sigma}^2_{\delta_k}}{2\sigma^2_d} \\ &\quad - \sum_{i=1}^{n}\sum_{j\neq i}\sum_{l=1}^{K} \widetilde{Z}_{ik}\,\widetilde{Z}_{jl}\,\frac{\Phi^{(t)}_{ij;kl}(\boldsymbol{x}) - 1}{(2\,p+4)\,\Phi^{(t)}_{ij;kl}(\boldsymbol{x})} \exp\left[(2\,p+4)\,\frac{1}{2}\,(\widetilde{\sigma}^2_{\delta_k} - \widetilde{\sigma}^{2\ (t)}_{\delta_k})\right] \\ Q_{\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}}(\widetilde{\mu}_{\beta^s_{kl}}) &= \frac{1}{2}\,\frac{\widetilde{a}_{\tau^2}}{\widetilde{b}_{\tau^2}}\left[\widetilde{I}^s_{kl}\left(1-\frac{1}{g}\right) - 1\right]\widetilde{\mu}^2_{\beta^s_{kl}} + \widetilde{\mu}_{\beta^s_{kl}}\sum_{i=1}^{n}\sum_{j\neq i}\widetilde{Z}_{ik}\,\widetilde{Z}_{jl}\,y_{ij}\,f_s(x_i,x_j) \\ &\quad - \sum_{i=1}^{n}\sum_{j\neq i}\widetilde{Z}_{ik}\,\widetilde{Z}_{jl}\,\frac{\Phi^{(t)}_{ij;kl}(\boldsymbol{x}) - 1}{(2\,p+4)\Phi^{(t)}_{ij;kl}(\boldsymbol{x})} \exp\left[(2\,p+4)\,f_s(x_i,x_j)(\widetilde{\mu}_{\beta^s_{kl}} - \widetilde{\mu}^{(t)}_{\beta^s_{kl}})\right] \\ Q_{\widetilde{\boldsymbol{\sigma}}^2_{\boldsymbol{\beta}}}(\widetilde{\sigma}^2_{\beta^s_{kl}}) &= \frac{1}{2}\log\widetilde{\sigma}^2_{\beta^s_{kl}} + \frac{1}{2}\,\frac{\widetilde{a}_{\tau^2}}{\widetilde{b}_{\tau^2}}\left[\widetilde{I}^s_{kl}\left(1-\frac{1}{g}\right) - 1\right]\widetilde{\sigma}^2_{\beta^s_{kl}} \\ &\quad - \sum_{i=1}^{n}\sum_{j\neq i}\widetilde{Z}_{ik}\,\widetilde{Z}_{jl}\frac{\Phi^{(t)}_{ij;kl}(\boldsymbol{x}) - 1}{(2\,p+4)\,\Phi^{(t)}_{ij;kl}(\boldsymbol{x})} \exp\left[(2\,p+4)\,\frac{1}{2}\,f^2_s(x_i,x_j)\,(\widetilde{\sigma}^2_{\beta^s_{kl}} - \widetilde{\sigma}^{2\ (t)}_{\beta^s_{kl}})\right]. \end{aligned}$$

These one-dimensional minorizing functions can be maximized by one-step Newton updates [19].

5 Experiments

We use two blog networks, summarized in Table 2.

The first blog network is the Political Blogs network of [22], where a directed edge from blogger i to blogger j indicates that blogger i cited blogger j. Since the blog network is directed, we convert it into an undirected network by specifying an undirected edge between two blogs i and j if there is an edge from i to j or j to i or both. We use two covariates: The first one indicates whether both blogs are liberal, while the second one indicates whether both blogs are conservative.

The second blog network is the Catalog Blogs 3 network of [23], where an undirected edge between two bloggers i and j indicates friendship. We use observed group memberships as covariates. We construct 39 indicators, one for each of the 39 groups: The indicator of group s is 1 if both blogs are members of group s otherwise 0. Note that although the number of original predictors is small, the number of block-dependent parameters is $O(K^2)$. In all of the experiments, we use the constants $\omega_1 = \cdots = \omega_K = 1$, $\mu_d = 0$, $\sigma_d^2 = 1$, $a_\gamma = 1$, $b_\gamma = 1$, $a_{\tau^2} = 1$, $b_{\tau^2} = 1$, and g =10,000.

Data	Nodes	Edge variables	Edges	Covariates	Block-dependent predictors	
					$K = 5$	$K = 20$
P. Blogs	1,490	1,109,305	16,783	2	70	880
C. Blogs	10,312	53,163,516	333,983	39	1,180	16,420

Table 2: Summary statistics of two blog networks.

To demonstrate the advantage of the Bayesian Variational GEM-MM algorithm with MM updates over the Bayesian Variational EM algorithm with FP updates, we compare FP and MM updates using both blog networks with $K = 5$ and $K = 20$ blocks. In both cases, we use MM algorithms in the M-step and averaged the results over 50 runs with random starting values.

Figure 1 compares lower bounds of two update methods based FP and MM. Overall, the MM method is far superior to the FP method on both small (i.e., Political Blogs) and large (i.e., Catalog Blogs) data sets under $K = 5$ as well as $K = 20$. We speculate that the superior nature of MM updates stems from the fact that the MM principle separates $\widetilde{\boldsymbol{Z}}_1, \ldots, \widetilde{\boldsymbol{Z}}_n$. The separation in turn weakens the dependence of updates of $\widetilde{\boldsymbol{Z}}_1, \ldots, \widetilde{\boldsymbol{Z}}_n$.

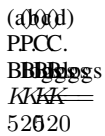

Figure 1: Lower bounds based on FP and MM implementation of E-step with $K = 5$ and $K = 20$.

We select K based on the lower bound on the logarithm of the marginal likelihood $p(\boldsymbol{y})$, on which Bayesian model selection is based. For both data sets, we vary K from 1 to 20 and select the K corresponding to the best lower bound on $\log p(\boldsymbol{y})$. For the Political Blogs, the best lower bound is achieved at $K = 12$ as shown in Figure 2(a). Figure 2(b) shows a histogram of parameter means $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}$ of all predictors and demonstrates that predictors are relevant across blocks. Figures 2(c) and 2(d) show that the covariate terms vary in strength across blocks.

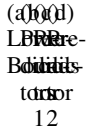

Figure 2: Political Blogs: (a) Lower bounds under $K = 1, \ldots, 20$. The best lower bound is achieved at $K = 12$. (b) Histogram of parameter means $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}$ of predictors. (c) and (d) Predictors 1 and 2 are active in most blocks.

In the case of the Catalog Blogs data set, the best lower bound is achieved at $K = 11$ (Figure 3(a)). In contrast to the Polical Blogs data set, there is evidence of sparsity (Figure 3(b)): The high peak at 0 indicates that a substantial proportion of covariate terms in inactive in some blocks. Figure 3(d) shows parameter estimates of predictor 12 across all blocks. Almost 50% of them are close to 0 which demonstrates that predictor 12 is inactive in most blocks. Observe that a predictor can act have both positive and negative effects on edge probabilities depending on the blocks involved, as shown in Figure 3(c).

Figure 3: Catalog Blogs: (a) Lower bounds under $K = 1, \ldots, 20$. The best lower bound is achieved at $K = 11$. (b) Histogram of parameter means $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}$ of predictors. The peak at 0 indicates that many block-dependent covariate parameters are inactive. (c) and (d) Predictor 2 is active in most blocks while predictor 12 is inactive in most.

To compare the predictive power of models, we construct a test data set by sampling 10% zero and 10% non-zero edges at random, then run models on the training data set where all test edges are set to zero plus other edges that are not in the test data set. We compare three models: the stochastic block model without predictors, the stochastic block model with block-invariant predictors, i.e., $\boldsymbol{\beta}_{kl} = \boldsymbol{\beta}$, and the stochastic block model with block-dependent predictors. The cross-validation procedure is repeated 10 times to obtain standard errors. We set $K = 10$ motivated by the results above. We use the predictive probabilities of all zero and non-zero test edges to compare three methods:

$$p(y_{ij} \mid \boldsymbol{y}) = \sum_{k=1}^{K} \sum_{l=1}^{K} p(y_{ij} \mid Z_{ik} = 1, Z_{jl} = 1, \widetilde{\boldsymbol{\mu}}_{\boldsymbol{\delta}}, \widetilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}})\, p(Z_{ik} = 1 \mid \widetilde{\boldsymbol{Z}}_i)\, p(Z_{jl} = 1 \mid \widetilde{\boldsymbol{Z}}_j). \quad (8)$$

Table 3 shows predictive probabilities of three models averaged over 10 training and test data sets. It is evident that incorporating predictors increases predictive power and that block-dependent predictors are best in terms of predictive power.

Data	without predictors	block-invariant predictors	block-dependent predictors
Political Blogs	-5,693 (5)	-4,925 (5)	**-4,788** (6)
Catalog Blogs	-122,509 (42)	-121,321 (41)	**-120,906** (41)

Table 3: Predictive probabilities of three models on test data sets. Mean test log likelihoods over 10 runs with random initializations are reported together with standard errors in parentheses.

6 Discussion

We believe that the MM principle is a promising approach to deal with the large number of local maxima encountered in applications of variational methods to statistical learning. We note that the convergence of variational GEM-MM algorithms can be sped up by (a) exploiting the separation of the parameters of the maximization problem and implementing parallel computing and by (b) using quasi-Newton methods.

References

[1] P. J. Bickel and A. Chen. A nonparametric view of network models and Newman-Girvan and other modularities. In *PNAS*, volume 106, pages 21068–21073, 2009.

[2] T. A. B. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14:75–100, 1997.

[3] K. Nowicki and T. A. B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

[4] J. J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Statistics and Computing*, 18(2):173–183, 2008.

[5] H. Zanghi, F. Picard, V. Miele, and C. Ambroise. Strategies for online inference of model-based clustering in large and growing networks. *Annals of Applied Statistics*, 4:687–714, 2010.

[6] P. Gopalan, D. Mimno, S. M. Gerrish, M. J. Freedman, and D. M. Blei. Scalable inference of overlapping communities. In *Neural Information Processing Systems*, 2012.

[7] D. Q. Vu, D. R. Hunter, and M. Schweinberger. Model-based clustering of large networks. *Annals of Applied Statistics*, 7:1010–1039, 2013.

[8] K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic block model. *Annals of Statistics*, 39:1878–1915, 2011.

[9] C. E. Priebe, D. L. Sussman, M. Tang, and J. T. Vogelstein. Statistical inference on errofully observed graphs. *Journal of the American Statistical Association*, 107:1119–1128, 2012.

[10] P. Sarkar and P. Bickel. The role of normalization in spectral clustering of stochastic blockmodels. Technical report, 2013. arXiv preprint arxiv:1310.0532.

[11] J. Lei and A. Rinaldo. Consistency of spectral clustering in sparse stochastic block models. Technical report, 2013. arXiv preprint arxiv:1312.2050.

[12] A. Celisse, J. J. Daudin, and L. Pierre. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6:1847–1899, 2012.

[13] P. J. Bickel, D. Choi, X. Chang, and H. Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Annals of Statistics*, 41:1922–1943, 2013.

[14] C. Tallberg. A Bayesian approach to modeling stochastic blockstructures with covariates. *Journal of Mathematical Sociology*, 29:1–23, 2005.

[15] P. D. Hoff. Random effects models for network data. In *Dynamic Social Network Modeling and Analysis*, pages 303–312. National Academies Press, 2003.

[16] M. A. J. van Duijn, T. A. B. Snijders, and B. J. H. Zijlstra. P2: a random effects model with covariates for directed graphs. *Statistica Neerlandica*, 58:234–254, 2004.

[17] P. N. Krivitsky, M. S. Handcock, A. E. Raftery, and P. D. Hoff. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks*, 31:204–213, 2009.

[18] D. R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–38, 2004.

[19] K. Lange. *Numerical Analysis for Statisticians*. Springer, New York, 2010.

[20] S. M. Stefanov. Convex quadratic minimization subject to a linear constraint and box constraints. *Appl Math Res Express*, 2004(1):17–42, 2004.

[21] R. M. Neal and G. E. Hinton. A new view of the EM algorithm that justifies incremental and other variants. In *Learning in Graphical Models*, pages 355–368, 1993.

[22] Lada A. Adamic and Natalie Glance. The political blogosphere and the 2004 U.S. election: Divided they blog. In *Proceedings of LinkKDD 2005*, pages 36–43. ACM, 2005.

[23] R. Zafarani and H. Liu. Social computing data repository at ASU, 2009.