The Annals of Statistics
2020, Vol. 48, No. 5, 2575–2598
https://doi.org/10.1214/19-AOS1898
© Institute of Mathematical Statistics, 2020

THEORETICAL AND COMPUTATIONAL GUARANTEES OF MEAN FIELD VARIATIONAL INFERENCE FOR COMMUNITY DETECTION

BY ANDERSON Y. ZHANG[1] AND HARRISON H. ZHOU[2]

[1]*Department of Statistics, The Wharton School, University of Pennsylvania, ayz@wharton.upenn.edu*
[2]*Department of Statistics, Yale University, huibin.zhou@yale.edu*

The mean field variational Bayes method is becoming increasingly popular in statistics and machine learning. Its iterative coordinate ascent variational inference algorithm has been widely applied to large scale Bayesian inference. See Blei et al. (2017) for a recent comprehensive review. Despite the popularity of the mean field method, there exist remarkably little fundamental theoretical justifications. To the best of our knowledge, the iterative algorithm has never been investigated for any high-dimensional and complex model. In this paper, we study the mean field method for community detection under the stochastic block model. For an iterative batch coordinate ascent variational inference algorithm, we show that it has a linear convergence rate and converges to the minimax rate within $\log n$ iterations. This complements the results of Bickel et al. (2013) which studied the global minimum of the mean field variational Bayes and obtained asymptotic normal estimation of global model parameters. In addition, we obtain similar optimality results for Gibbs sampling and an iterative procedure to calculate maximum likelihood estimation, which can be of independent interest.

1. Introduction.

A major challenge of large scale Bayesian inference is the calculation of posterior distribution. For high-dimensional and complex models, the exact calculation of posterior distribution is often computationally intractable. To address this challenge, the mean field variational method [2, 19, 29] is used to approximate posterior distributions in a wide range of applications in many fields including natural language processing [8, 21], computational neuroscience [15, 25] and network science [1, 9, 17]. This method is different from Markov chain Monte Carlo (MCMC) [14, 27], another popular approximation algorithm. The variational inference approximation is deterministic for each iterative update, while MCMC is a randomized sampling algorithm, so that for large-scale data analysis, the mean field variational Bayes usually converges faster than MCMC [7], which is particularly attractive in the big data era.

In spite of a wide range of successful applications of the mean field variational Bayes, its fundamental theoretical properties are rarely investigated. The existing literature [3, 9, 30, 32, 33] is mostly on low-dimensional parameter estimation and on the global minimum of the variational Bayes method. For example, in a recent inspiring paper, Wang and Blei [31] studied the frequentist consistency of the variational method for a general class of latent variable models. They obtained consistency for low-dimensional global parameters and further showed asymptotic normality, assuming the global minimum of the variational Bayes method can be achieved. However, it is often computationally infeasible to attain the global minimum when the model is high-dimensional or complex. This motivates us to investigate the statistical properties of the mean field in high-dimensional settings, and more importantly, to understand the statistical and computational guarantees of the iterative variational inference algorithms.

Received December 2017; revised July 2019.
MSC2020 subject classifications. 60G05.
Key words and phrases. Mean field, variational inference, Bayesian, community detection, stochastic block model.

The success and the popularity of the mean field method in Bayesian inference mainly lies in the success of its iterative algorithm: Coordinate Ascent Variational Inference (CAVI) [7], which provides a computationally efficient way to approximate the posterior distribution. It is important to understand what statistical properties CAVI has and how do they compare to the optimal statistical accuracy. In addition, we want to investigate how fast CAVI converges for the purpose of implementation. With the ambition of establishing a universal theory of the mean field iterative algorithm for general models in mind, in this paper, we consider the community detection problem [1, 4, 12, 23, 24, 34] under the Stochastic Block Model (SBM) [4, 18, 20, 28] as our first step.

Community detection has been an active research area in recent years, with the SBM as a popular choice of model. The Bayesian framework and the variational inference for community detection are considered in [1, 3, 9, 13, 17, 26]. For high-dimensional settings, Celisse et al. [9] and Bickel et al. [3] are arguably the first to study the statistical properties of the mean field for SBMs. The authors built an interesting connection between full likelihood and variational likelihood, and then studied the closeness of maximum likelihood and maximum variational likelihood, from which they obtained consistency and asymptotic normality for global parameter estimation. From a personal communication with the authors of Bickel et al. [3], an implication of their results is that the variational method achieves exact community recovery under a strong signal-to-noise (SNR) ratio. Their analysis idea is fascinating, but it is not clear whether it is possible to extend the analysis to other SNR conditions under which exact recovery may never be possible. More importantly, it may not be computationally feasible to maximize the variational likelihood for the SBM, as seen from Theorem 2.1.

In this paper, we consider the statistical and computational guarantees of the iterative variational inference algorithm for community detection. The primary goal of community detection problem is to recover the community membership in a network. We measure the performance of the iterative variational inference algorithm by comparing its output with the ground truth. Denote the underlying ground truth by Z^*. For a network of n nodes and k communities, Z^* is an $n \times k$ matrix with each row a standard Euclidean basis in $\mathbb{R}^k$. The index of nonzero coordinate of each row $\{Z^*_{i,\cdot}\}_{i=1}^n$ gives the community assignment information for the corresponding node. We propose an iterative algorithm called Batch Coordinate Ascent Variational Inference (BCAVI), a slight modification of CAVI with batch updates, to make parallel and distributed computing possible. Let $\pi^{(s)}$ denote the output of the sth iteration, a $n \times k$ matrix with nonnegative entries. The summation of each row $\{\pi^{(s)}_{i,\cdot}\}_{i=1}^n$ is equal to 1, which is interpreted as an approximate posterior probability of assigning the corresponding node of each row into k communities. The performance of $\pi^{(s)}$ is measured by a ℓ_1 loss $\ell(\cdot,\cdot)$ compared with Z^*.

An informal statement of the main result: Let $\pi^{(s)}$ be the estimation of community membership from the iterative algorithm BCAVI after s iterations. Under weak regularity condition, for some $c_n = o_n(1)$, with high probability, we have for all $s \geq 0$,

$$\ell(\pi^{(s+1)}, Z^*) \leq \text{minimax rate} + c_n \ell(\pi^{(s)}, Z^*). \tag{1}$$

The main contribution of this paper is equation (1). The coefficient c_n is $o_n(1)$ and is independent of s, which implies $\ell(\pi^{(s)}, Z^*)$ decreases at a fast linear rate. In addition, we show that BCAVI converges to the statistical optimality [34]. It is worth mentioning that after $\log n$ iterations BCAVI attains the minimax rate, up to an error $o_n(n^{-a})$ for any constant $a > 0$. The conditions required for the analysis of BCAVI are relatively mild. We allow the number of communities to grow. The sizes of the communities are not assumed to be of the same order. The separation condition on global parameters covers a wide range of settings from consistent community detection to exact recovery.

To the best of our knowledge, this provides arguably the first theoretical justification for the iterative algorithm of the mean field variational method in a high-dimensional and complex setting. Though we focus on the problem of community detection in this paper, we hope the analysis would shed some light on analyzing other models, which may eventually lead to a general framework of understanding the mean field theory.

The techniques of analyzing the mean field can be extended to providing theoretical guarantees for other iterative algorithms, including Gibbs sampling and an iterative procedure for maximum likelihood estimation, which can be of independent interest. Results similar to equation (1) are obtained for both methods under the SBM.

Organization. The paper is organized as follows. In Section 2, we introduce the mean field theory and the implementation of BCAVI algorithm for community detection. All the theoretical justifications for the mean field method are in Section 3. Discussions on the convergence of the global minimizer and other iterative algorithms are presented in Section 4. The proofs of theorems are in Section 6. We include all the auxiliary lemmas and propositions and their corresponding proofs in the Supplementary Material [35].

Notation. Throughout this paper, for any matrix $X \in \mathbb{R}^{n\times m}$, its ℓ_1 norm is defined in analogous to that of a vector. That is, $\|X\|_1 = \sum_{i,j} |X_{i,j}|$. We use the notation $X_{i,\cdot}$ and $X_{\cdot,i}$ to indicate its ith row and column, respectively. For matrices X, Y of the same dimension, their inner product is defined as $\langle X, Y\rangle = \sum_{i,j} X_{i,j}Y_{i,j}$. For any set D, we use $|D|$ for its cardinality. We denote $\mathrm{Ber}(p)$ for a Bernoulli random variable with success probability p. For two positive sequences x_n and y_n, $x_n \lesssim y_n$ means $x_n \leq cy_n$ for some constant c not depending on n. We adopt the notation $x_n \asymp y_n$ if $x_n \lesssim y_n$ and $y_n \lesssim x_n$. To distinguish from the probabilities p, q, we use bold $\mathbf{p}$ and $\mathbf{q}$ to indicate distributions. The Kullback–Leibler divergence between two distributions is defined as $\mathrm{KL}(\mathbf{p}\|\mathbf{q}) = \mathbb{E}_{\mathbf{q}} \log(\mathbf{p}(x)/\mathbf{q}(x))$. We use $\psi(\cdot)$ for the digamma function, which is defined as the logarithmic derivative of Gamma function, that is, $\psi(x) = \frac{d}{dx}[\log\Gamma(x)]$. In any $\mathbb{R}^d$, we denote $\{e_a\}_{a=1}^d$ to be the standard Euclidean basis with $e_1 = (1, 0, 0, \ldots), e_2 = (0, 1, 0, \ldots, 0), \ldots, e_d = (0, 0, 0, \ldots, 1)$. We let 1_d be a vector of length d whose entries are all 1. We use $[d]$ to indicate the set $\{1, 2, \ldots, d\}$. Throughout this paper, the superscript "pri" (e.g., π^{pri}) indicates that this is a hyperparameter of priors.

2. Mean field method for community detection.

In this section, we first give a brief introduction to the variational inference method in Section 2.1. Then we introduce the community detection problem and the stochastic block model in Section 2.2. The Bayesian framework is presented in Section 2.3. Its mean field approximation and CAVI updates are given in Section 2.4 and Section 2.5, respectively. The BCAVI algorithm is introduced in Section 2.6.

2.1. *Mean field variational inference.*

We first present the mean field method in a general setting and then consider its application to the community detection problem. Let $\mathbf{p}(x|y)$ be an arbitrary posterior distribution for x, given observation y. Here, x can be a vector of latent variables, with coordinates $\{x_i\}$. It may be difficult to compute the posterior $\mathbf{p}(x|y)$ exactly. The variational Bayes ignores the dependence among $\{x_i\}$, by simply taking a product measure $\mathbf{q}(x) = \prod_i \mathbf{q}_i(x_i)$ to approximate it. Usually each $\mathbf{q}_i(x_i)$ is simple and easy to compute. The best approximation is obtained by minimizing the Kullback–Leibler divergence between $\mathbf{q}(x)$ and $\mathbf{p}(x|y)$:

$$\hat{\mathbf{q}}^{\mathrm{MF}} = \underset{\mathbf{q}\in\mathbf{Q}}{\arg\min}\, \mathrm{KL}(\mathbf{q}\|\mathbf{p}), \tag{2}$$

where $\mathbf{Q}$ is the space of all product measures. Despite the fact that every measure $\mathbf{q}$ has a simple product structure, the global minimizer $\hat{\mathbf{q}}^{\mathrm{MF}}$ remains computationally intractable.

To address this issue, an iterative Coordinate Ascent Variational Inference (CAVI) is widely used to approximate the global minimum. It is a greedy algorithm. The value of $\mathrm{KL}(\mathbf{q}\|\mathbf{p})$ decreases in each coordinate update:

$$\hat{\mathbf{q}}_i = \min_{\mathbf{q}_i \in \mathbf{Q}_i} \mathrm{KL}\Big[\mathbf{q}_i \prod_{j\neq i} \mathbf{q}_j \| \mathbf{p}\Big] \quad \forall i. \tag{3}$$

The coordinate update has an explicit formula

$$\hat{\mathbf{q}}_i(x_i) \propto \exp\big[\mathbb{E}_{\mathbf{q}_{-i}}\big[\log \mathbf{p}(x_i|x_{-i}, y)\big]\big], \tag{4}$$

where x_{-i} indicates all the coordinates in x except x_i, and the expectation is over $\mathbf{q}_{-i} = \prod_{j\neq i} \mathbf{q}_j(x_j)$. Equation (4) is usually easy to compute, which makes CAVI computationally attractive, although CAVI only guarantees to achieve a local minimum.

In summary, the mean field variational inference via CAVI can be represented in the following diagram:

$$\mathbf{p}(x|y) \overset{\text{approx.}}{\Longleftarrow} \hat{\mathbf{q}}^{\mathrm{MF}}(x) \overset{\text{approx.}}{\Longleftarrow} \hat{\mathbf{q}}^{\mathrm{CAVI}}(x),$$

where $\hat{\mathbf{q}}^{\mathrm{MF}}(x)$, the global minimum, serves mainly as an intermediate step in the mean field methodology. What is implemented in practice to approximate global minimum is an iterative algorithm like CAVI. This motivates us to consider directly the theoretical guarantees of the iterative algorithm in this paper.

We refer the readers to a nice review and tutorial by Blei et al. [7] for more detail on the variational inference and CAVI. The derivation from equation (3) to equation (4) can be found in many variational inference literatures [6, 7]. We include it in Appendix D in the Supplementary Material for completeness.

2.2. *Community detection and stochastic block model.* The Stochastic Block Model (SBM) has been a popular model for community detection.

Consider an n-node network with its adjacency matrix denoted by A. It is an unweighted and undirected network without self-loops, with $A \in \{0,1\}^{n\times n}$, $A = A^T$ and $A_{i,i} = 0, \forall i \in [n]$. Each edge is an independent Bernoulli random variable with $\mathbb{E}A_{i,j} = P_{i,j}, \forall i < j$. In the SBM, the value of connectivity probability $P_{i,j}$ depends on the communities the two endpoints i and j belong to. We assume $P_{i,j} = p$ if both nodes come from the same community and $P_{i,j} = q$ otherwise. There are k communities in the network. We denote $z \in [k]^n$, as the assignment vector, with z_i indicating the index of community the ith node belongs to. Thus, the connectivity probability matrix P can be written as

$$P_{i,j} = B_{z_i,z_j},$$

where $B \in [0,1]^{k\times k}$ with diagonal entries as p and off-diagonal entries as q. That is, $B = q1_k1_k^T + (p-q)I_k$. Let $Z \in \Pi_0$ be the assignment matrix where

$$\Pi_0 = \big\{\pi \in \{0,1\}^{n\times k} : \|\pi_{i,\cdot}\|_0 = 1, \forall i \in [n]\big\}.$$

In each row $\{Z_{i,\cdot}\}_{i=1}^n$, there is only one 1 with all the other coordinates as 0, indicating the assignment of community for the corresponding node. Then P can be equivalently written as $P_{i,j} = Z_{i,\cdot}BZ_{j,\cdot}^T, \forall i < j$, or in a matrix form

$$P_{i,j} = \big(ZBZ^T\big)_{i,j} \quad \forall i < j.$$

The goal of community detection is to recover the assignment vector z, or equivalently, the assignment matrix Z. The equivalence can be seen by observing that there is a bijection r between $z \in [k]^n$ and $Z \in \Pi_0$ which is defined as follows:

$$r(z) = Z, \quad \text{where } Z_{i,a} = \mathbb{I}\{a = z_i\} \ \forall i \in [n], a \in [k]. \tag{5}$$

Since they are uniquely determined by each other, in our paper we may use z directly without explicitly defining $z = r^{-1}(Z)$ (or vice versa) when there is no ambiguity.

2.3. *A Bayesian framework.* Throughout the whole paper, we assume k, the number of communities, is known. We observe the adjacency matrix A. The global parameters p and q and the community assignment Z are unknown. From the description of the model in Section 2.2, we can write down the distribution of A as follows:

$$\mathbf{p}(A|Z, p, q) = \prod_{i<j} B_{z_i,z_j}^{A_{i,j}} (1 - B_{z_i,z_j})^{1-A_{i,j}}, \tag{6}$$

with $B = q 1_k 1_k^T + (p - q) I_k$ and $z = r^{-1}(Z)$. We are interested in Bayesian inference for estimating Z, with prior to be given on both p, q and Z.

We assume that $\{z_i\}_{i=1}^n$ have independent categorical (a.k.a. multinomial with size one) priors with hyperparameters $\{\pi_{i,\cdot}^{\text{pri}}\}_{i=1}^n$, where $\sum_{a=1}^k \pi_{i,a}^{\text{pri}} = 1, \forall i \in [n]$. In other words, $\{Z_{i,\cdot}\}_{i=1}^n$ are independently distributed by

$$\mathbb{P}(Z_{i,\cdot} = e_a^T) = \pi_{i,a}^{\text{pri}} \quad \forall a = 1, 2, \dots, k,$$

where $\{e_a\}_{a=1}^k$ are the coordinate vectors. Here, we allow the priors for $Z_{i,\cdot}$ to be different for different i. If additionally $\pi_{i,\cdot} = \pi_{j,\cdot}$ for all $i \neq j$ is assumed, and then this is reduced to the usual case of i.i.d. priors.

Since $\{A_{i,j}\}_{i<j}$ are Bernoulli, it is natural to consider a conjugate Beta prior for p and q. Let $p \sim \text{Beta}(\alpha_p^{\text{pri}}, \beta_p^{\text{pri}})$ and $q \sim \text{Beta}(\alpha_q^{\text{pri}}, \beta_q^{\text{pri}})$. Then the joint distribution is

$$\begin{aligned} \mathbf{p}(A, Z, p, q) = & \left[\prod_i \pi_{i,z_i}^{\text{pri}}\right]\left[\prod_{i<j} B_{z_i,z_j}^{A_{i,j}} (1 - B_{z_i,z_j})^{1-A_{i,j}}\right] \\ & \times \left[\frac{\Gamma(\alpha_p^{\text{pri}} + \beta_p^{\text{pri}})}{\Gamma(\alpha_p^{\text{pri}})\Gamma(\beta_p^{\text{pri}})} p^{\alpha_p^{\text{pri}}-1}(1-p)^{\beta_p^{\text{pri}}-1}\right] \\ & \times \left[\frac{\Gamma(\alpha_q^{\text{pri}} + \beta_q^{\text{pri}})}{\Gamma(\alpha_q^{\text{pri}})\Gamma(\beta_q^{\text{pri}})} q^{\alpha_q^{\text{pri}}-1}(1-q)^{\beta_q^{\text{pri}}-1}\right]. \end{aligned} \tag{7}$$

Our main interest is to infer Z, from the posterior distribution $\mathbf{p}(Z, p, q|A)$. However, the exact calculation of $\mathbf{p}(Z, p, q|A)$ is computationally intractable.

2.4. *Mean field approximation.* Since the posterior distribution $\mathbf{p}(Z, p, q|A)$ is computationally intractable, we apply the mean field approximation to approximate it by a product measure,

$$\mathbf{q}_{\pi,\alpha_p,\beta_p,\alpha_q,\beta_q}(Z, p, q) = \mathbf{q}_\pi(Z)\mathbf{q}_{\alpha_p,\beta_p}(p)\mathbf{q}_{\alpha_q,\beta_q}(q),$$

where $\{r^{-1}(Z_{i,\cdot})\}_{i=1}^n$ are independent categorical variables with parameters $\{\pi_{i,\cdot}\}_{i=1}^n$, that is, $\mathbf{q}_\pi(Z) = \prod_{i=1}^n \mathbf{q}_{\pi_{i,\cdot}}(Z_{i,\cdot})$ with

$$\mathbf{q}_{\pi_{i,\cdot}}(Z_{i,\cdot} = e_a) = \pi_{i,a} \quad \forall i \in [n], a \in [k],$$

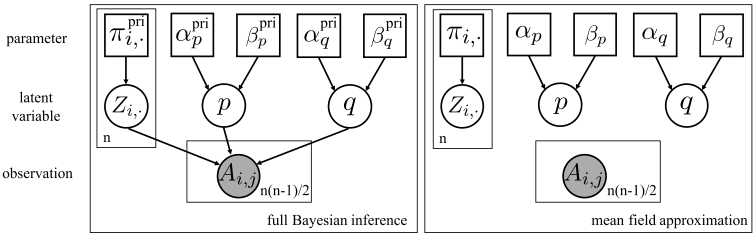

FIG. 1. *Graphical model presentations of full Bayesian inference* (*left panel*) *and the mean field approximation* (*right panel*) *for community detection. The edges show the dependence among variables.*

and $\mathbf{q}_{\alpha_p,\beta_p}(p)$ and $\mathbf{q}_{\alpha_q,\beta_q}(q)$ are Beta with parameters $\alpha_p, \beta_p, \alpha_q, \beta_q$ due to conjugacy. See Figure 1 for the graphical presentation of $\mathbf{q}_{\pi,\alpha_p,\beta_p,\alpha_q,\beta_q}(Z, p, q)$.

Note that the distribution class of $\mathbf{q}$ is fully captured by the parameters $(\pi, \alpha_p, \beta_p, \alpha_q, \beta_q)$, and then the optimization in equation (2) is equivalent to minimize over the parameters as

$$(\hat{\pi}^{\text{MF}}, \hat{\alpha}_p^{\text{MF}}, \hat{\beta}_p^{\text{MF}}, \hat{\alpha}_q^{\text{MF}}, \hat{\beta}_q^{\text{MF}}) = \mathop{\arg\min}_{\substack{\pi\in\Pi_1\\ \alpha_p,\beta_p,\alpha_q,\beta_q>0}} \text{KL}\big[\mathbf{q}_{\pi,\alpha_p,\beta_p,\alpha_q,\beta_q}(Z, p, q)\|\mathbf{p}(Z, p, q|A)\big], \tag{8}$$

where

$$\Pi_1 = \big\{\pi \in [0, 1]^{n\times k}, \|\pi_{i,\cdot}\|_1 = 1\big\}.$$

Here, Π_1 can be viewed as a relaxation of Π_0: it uses a ℓ_1 constraint on each row instead of the ℓ_0 constraint used in Π_0. The global minimizer $\mathbf{q}_{\hat{\pi}^{\text{MF}}}(Z)$ gives approximate probabilities to classify every node to each community. The optimization in equation (8) can be shown to be equivalent to a more explicit optimization as follows. Recall $\psi(\cdot)$ is the digamma function with $\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x}[\log\Gamma(x)]$.

THEOREM 2.1. *The mean field estimator* $(\hat{\pi}^{\text{MF}}, \hat{\alpha}_p^{\text{MF}}, \hat{\beta}_p^{\text{MF}}, \hat{\alpha}_q^{\text{MF}}, \hat{\beta}_q^{\text{MF}})$ *defined in equation* (8) *is equivalent to*

$$(\hat{\pi}^{\text{MF}}, \hat{\alpha}_p^{\text{MF}}, \hat{\beta}_p^{\text{MF}}, \hat{\alpha}_q^{\text{MF}}, \hat{\beta}_q^{\text{MF}}) = \mathop{\arg\min}_{\substack{\pi\in\Pi_1\\ \alpha_p,\beta_p,\alpha_q,\beta_q>0}} f(\pi, \alpha_p, \beta_p, \alpha_q, \beta_q; A),$$

where

$$\begin{aligned} f(\pi, \alpha_p, \beta_p, \alpha_q, \beta_q; A) = {} & t\langle A - \lambda 1_n 1_n^T + \lambda I_n, \pi\pi^T\rangle + \frac{1}{2}[\psi(\alpha_q) - \psi(\beta_q)]\|A\|_1 \\ & + \frac{n(n-1)}{2}[\psi(\beta_q) - \psi(\alpha_q + \beta_q)] \\ & - \sum_{i=1}^{n} \text{KL}\big[\text{Categorical}(\pi_{i,\cdot})\,\|\,\text{Categorical}\big(\pi^{\text{pri}}_{i,\cdot}\big)\big] \\ & - \text{KL}\big[\text{Beta}(\alpha_p, \beta_p)\,\|\,\text{Beta}\big(\alpha_p^{\text{pri}}, \beta_p^{\text{pri}}\big)\big] \\ & - \text{KL}\big[\text{Beta}(\alpha_q, \beta_q)\,\|\,\text{Beta}\big(\alpha_q^{\text{pri}}, \beta_q^{\text{pri}}\big)\big], \end{aligned}$$

and

$$t = \big[[\psi(\alpha_p) - \psi(\beta_p)] - [\psi(\alpha_q) - \psi(\beta_q)]\big]/2, \tag{9}$$

$$\lambda = \big[[\psi(\beta_q) - \psi(\alpha_q + \beta_q)] - [\psi(\beta_p) - \psi(\alpha_p + \beta_p)]\big]/(2t). \tag{10}$$

The explicit formulation in Theorem 2.1 is helpful to understand the global minimizer of the mean field method. However, the global minimizer $\hat{\pi}^{\mathrm{MF}}$ remains computationally infeasible as the objective function is not convex. Fortunately, there is a practically useful algorithm to approximate it.

2.5. *Coordinate ascent variational inference.* CAVI is possibly the most popular algorithm to approximate the global minimum of the mean field variational Bayes. It is an iterative algorithm. In equation (8), there are latent variables $\{Z_{i,\cdot}\}_{i=1}^n, p, q$. CAVI updates them one by one. Since the distribution class of $\mathbf{q}$ is uniquely determined by the parameters $\{\pi_{i,\cdot}\}_{i=1}^n, \alpha_p, \beta_p, \alpha_q, \beta_q$, equivalently we are updating those parameters iteratively. Theorem 2.2 gives explicit formulas for the coordinate updates.

THEOREM 2.2. *Starting with some $\pi, \alpha_p, \beta_p, \alpha_q, \beta_q$, the CAVI update for each coordinate* (*i.e.*, *equation* (3) *and equation* (4)) *has an explicit expression as follows*:

- *Update on p*:

$$\alpha_p' = \alpha_p^{\mathrm{pri}} + \sum_{i<j}\sum_{a=1}^{k} \pi_{i,a}\pi_{j,a}A_{i,j} \quad \textit{and} \quad \beta_p' = \beta_p^{\mathrm{pri}} + \sum_{i<j}\sum_{a=1}^{k} \pi_{i,a}\pi_{j,a}(1-A_{i,j}).$$

- *Update on q*:

$$\alpha_q' = \alpha_q^{\mathrm{pri}} + \sum_{i<j}\sum_{a\neq b} \pi_{i,a}\pi_{j,b}A_{i,j} \quad \textit{and} \quad \beta_q' = \beta_q^{\mathrm{pri}} + \sum_{i<j}\sum_{a\neq b} \pi_{i,a}\pi_{j,b}(1-A_{i,j}).$$

- *Update on $Z_{i,\cdot}, \forall i = 1, 2, \dots, n$*:

$$\pi_{i,a}' \propto \pi_{i,a}^{\mathrm{pri}} \exp\left[2t \sum_{j\neq i} \pi_{j,a}(A_{i,j} - \lambda)\right] \quad \forall a = 1, 2, \dots, k,$$

where t and λ are defined in equation (9) *and equation* (10), *respectively*, *and the normalization satisfies $\sum_{a=1}^k \pi_{i,a}' = 1$.*

All coordinate updates in Theorem 2.2 have explicit formulas, which makes CAVI a computationally attractive way to approximate the global optimum $\hat{\mathbf{q}}^{\mathrm{MF}}$ for the community detection problem.

2.6. *Batch coordinate ascent variational inference.* The Batch Coordinate Ascent Variational Inference (BCAVI) is a batch version of CAVI. The difference lies in that CAVI updates the rows of π sequentially one by one, while BCAVI uses the value of π to update all rows $\{\pi_{i,\cdot}'\}$ according to Theorem 2.2. This makes BCAVI especially suitable for parallel and distributed computing, a nice feature for large scale network analysis.

We define a mapping $h : \Pi_1 \to \Pi_1$ as follows. For any $\pi \in \Pi_1$, we have

$$[h_{t,\lambda}(\pi)]_{i,a} \propto \pi_{i,a}^{\mathrm{pri}} \exp\left[2t \sum_{j\neq i} \pi_{ja}(A_{i,j} - \lambda)\right], \tag{11}$$

with parameters t and λ. For BCAVI, we update π by $\pi' = h_{t,\lambda}(\pi)$ in each batch iteration, with t, λ defined in equations (14) and (15). See Algorithm 1 for the detailed implementation of BCAVI algorithm.

REMARK 2.1. The definitions of $t^{(s)}$ and $\lambda^{(s)}$ in equations (14) and (15) involve the digamma function, which costs a nonnegligible computational resources each time called. Note that we have $\psi(x) \in (\log(x - \frac{1}{2}), \log x)$ for all $x > 1/2$. For the computational purpose,

Algorithm 1: Batch Coordinate Ascent Variational Inference (BCAVI)

```
Input: Adjacency matrix $A$, number of communities $k$, hyperparameters
       $\pi^{\mathrm{pri}}, \alpha_p^{\mathrm{pri}}, \beta_p^{\mathrm{pri}}, \alpha_q^{\mathrm{pri}}, \beta_q^{\mathrm{pri}}$, initializer $\pi^{(0)}$, number of iterations $S$.
Output: Mean variational Bayes approximation $\hat{\pi}, \hat{\alpha}_p, \hat{\beta}_p, \hat{\alpha}_q, \hat{\beta}_q$.
for $s = 1, 2, \ldots, S$ do
    Update $\alpha_p^{(s)}, \beta_p^{(s)}, \alpha_q^{(s)}, \beta_q^{(s)}$ by
    (12) $\alpha_p^{(s)} = \alpha_p^{\mathrm{pri}} + \sum_{a=1}^{k} \sum_{i<j} A_{i,j} \pi_{i,a}^{(s-1)} \pi_{j,a}^{(s-1)}$,
         $\beta_p^{(s)} = \beta_p^{\mathrm{pri}} + \sum_{a=1}^{k} \sum_{i<j} (1 - A_{i,j}) \pi_{i,a}^{(s-1)} \pi_{j,a}^{(s-1)}$,
    (13) $\alpha_q^{(s)} = \alpha_q^{\mathrm{pri}} + \sum_{a \neq b} \sum_{i<j} A_{i,j} \pi_{i,a}^{(s-1)} \pi_{j,b}^{(s-1)}$,
         $\beta_q^{(s)} = \beta_q^{\mathrm{pri}} + \sum_{a \neq b} \sum_{i<j} (1 - A_{i,j}) \pi_{i,a}^{(s-1)} \pi_{j,b}^{(s-1)}$.
    Define
    (14) $t^{(s)} = \frac{1}{2}[[\psi(\alpha_p^{(s)}) - \psi(\beta_p^{(s)})] - [\psi(\alpha_q^{(s)}) - \psi(\beta_q^{(s)})]]$,
    (15) $\lambda^{(s)} = \frac{1}{2t^{(s)}}[[\psi(\beta_q^{(s)}) - \psi(\alpha_q^{(s)} + \beta_q^{(s)})] - [\psi(\beta_p^{(s)}) - \psi(\alpha_p^{(s)} + \beta_p^{(s)})]]$,
    where $\psi(\cdot)$ is the digamma function. Then update $\pi^{(s)}$ with
         $\pi^{(s)} = h_{t^{(s)}, \lambda^{(s)}}(\pi^{(s-1)})$,
    where the mapping $h(\cdot)$ is defined as in equation (11).
end
We have $\hat{\pi} = \pi^{(S)}, \hat{\alpha}_p = \alpha_p^{(S)}, \hat{\beta}_p = \beta_p^{(S)}, \hat{\alpha}_q = \alpha_q^{(S)}, \hat{\beta}_q = \beta_q^{(S)}$.
```

we propose to use the logarithmic function instead of digamma function in Algorithm 1, that is, equations (14) and (15) are replaced by

$$t^{(s)} = \frac{1}{2} \log \frac{\alpha_p^{(s)} \beta_q^{(s)}}{\beta_p^{(s)} \alpha_q^{(s)}} \quad \text{and} \quad \lambda^{(s)} = \frac{1}{2t^{(s)}} \log \frac{\beta_q^{(s)}(\alpha_p^{(s)} + \beta_p^{(s)})}{(\alpha_q^{(s)} + \beta_q^{(s)})\beta_p^{(s)}}. \tag{16}$$

Later we show that $\alpha_p^{(s)}, \beta_p^{(s)}, \alpha_q^{(s)}, \beta_q^{(s)}$ are all at least in the order of np, which goes to infinity, and thus the error caused by using the logarithmic function to replace the digamma function is negligible. All theoretical guarantees obtained in Section 3 for Algorithm 1 (i.e., Theorem 3.1, Theorem 3.2) still hold if we use equation (16) to replace equations (14) and (15).

REMARK 2.2. The updating order of $\alpha_p, \beta_p, \alpha_q, \beta_q$ and π in Algorithm 1 can be exchanged. With $\alpha_p^{(0)}, \beta_p^{(0)}, \alpha_q^{(0)}, \beta_q^{(0)}$ and $\pi^{(0)}$ initialized, we can instead update π first, followed by the update on $\alpha_p, \beta_p, \alpha_q, \beta_q$. Theoretical guarantees including Theorem 3.1 and

Theorem 3.2 will still hold, under an additional consistency assumption on $\alpha_p^{(0)}$, $\beta_p^{(0)}$, $\alpha_q^{(0)}$, $\beta_q^{(0)}$, which can be met by simple methods, for instance, method of moments [5].

3. Theoretical justifications. In this section, we establish theoretical justifications for BCAVI for community detection under the stochastic block model. Though Z, p and q are all unknown, the main interest of community detection is on the recovery of the assignment matrix Z, while p and q are nuisance parameters. As a result, our main focus is on developing convergence rate of BCAVI for π.

3.1. *Loss function.* We use ℓ_1 norm to measure the performance of recovering Z. Let Φ be the set of all the bijections from $[k]$ to $[k]$. Then for any $Z, Z^* \in \Pi_1$, the loss function is defined as

$$\ell(Z, Z^*) = \inf_{\phi\in\Phi} \|Z - \phi \circ Z^*\|_1 = \inf_{\phi\in\Phi} \sum_{i,a} |Z_{i,a} - Z^*_{i,\phi(a)}|. \tag{17}$$

Note that the infimum over Φ addresses the issue of identifiability over the labels. For instance, in the case of $n = 4, k = 2$, the assignment vector $z = (1, 1, 2, 2)$ and $z' = (2, 2, 1, 1)$ give the same partition. In equation (17), two equivalent assignments give the same loss.

There are a few reasons for the choice of the ℓ_1 norm. When both $Z, Z' \in \Pi_0$, the ℓ_1 distance between Z and Z' is equal to the ℓ_0 norm, that is, the Hamming distance between the corresponding assignment vectors $r^{-1}(Z)$ and $r^{-1}(Z')$, which is the default metric used in community detection literature [12, 34]. The other reason is related to the interpretation of Π_1. Since each row of Π_1 corresponds to a categorical distribution, it is natural to use the ℓ_1 norm, the total variation distance, to measure their diffidence.

3.2. *Ground truth.* We use the superscript asterisk (*) to indicate the ground truth. The ground truth of connectivity matrix B^* is

$$B^* = q^* 1_k 1_k^T + (p^* - q^*) I_k,$$

where p^* is the within community connection probability and q^* is the between community connection probability. Throughout the paper, we assume $p^* > q^*$ such that the network satisfies the so-called "assortative" property, with the within-community connectivity probability larger than the between-community connectivity probability.

We further assume the network is generated by the true assignment matrix Z^* in the sense that $P_{i,j} = (Z^* B^* Z^{*T})_{i,j}$ for all $i \neq j$. We are interested in deriving a statistical guarantee of $\ell(\hat{\pi}^{(s)}, Z^*)$. Throughout this section, we consider cases $Z^* \in \Pi_0$ or $Z^* \in \Pi_0^{(\rho,\rho')}$, where $\Pi_0^{(\rho,\rho')}$ is defined to be a subset of Π_0 with all the community sizes bounded between $\rho n/k$ and $\rho' n/k$. That is,

$$\Pi_0^{(\rho,\rho')} = \{\pi \in \Pi_0 : \rho n/k \leq |\{i \in [n] : \pi_{i,a} = 1\}| \leq \rho' n/k, \forall a \in [k]\}.$$

It is worth mentioning that ρ, ρ' are not necessarily constants. We allow the community sizes not to be of the same order in the theoretical analysis.

3.3. *Theoretical justifications for BCAVI.* In Theorem 3.1, we present theoretic guarantees of the convergence rate of BCAVI when initialized properly. Define

$$w = \max_{i\in[n]} \max_{a,b\in[k]} \pi^{\text{pri}}_{i,a} / \pi^{\text{pri}}_{i,b} \quad \text{and} \quad \bar{n}_{\min} = \min_{a\neq b} [n_a + n_b]/2, \tag{18}$$

where $n_a = |\{i \in [n] : Z^*_{i,a} = 1\}|$ is the size of ath community for all $a \in [k]$. When $w = 1$, the priors for $\{r^{-1}(Z_{i,\cdot})\}_{i=1}^n$ are i.i.d. Categorical$(1/k, 1/k, \ldots, 1/k)$ and $\bar{n}_{\min} = n/2$ when there exist only two communities. The following quantity I plays a key role in the minimax theory [34]:

$$I = -2\log\left[\sqrt{p^*q^*} + \sqrt{(1-p^*)(1-q^*)}\right],$$

which is the Rényi divergence of order $1/2$ between two Bernoulli distributions: $\mathrm{Ber}(p^*)$ and $\mathrm{Ber}(q^*)$. The proof of Theorem 3.1 is deferred to Section 6.3.

THEOREM 3.1. *Let $Z^* \in \Pi_0$. Let $0 < c_0 < 1$ be any constant. Assume $0 < c_0 p^* < q^* < p^* = o_n(1)$,*

$$nI/\left[wk[n/\bar{n}_{\min}]^2\right] \to \infty \quad \text{and} \quad \alpha_p^{\mathrm{pri}}, \beta_p^{\mathrm{pri}}, \alpha_q^{\mathrm{pri}}, \beta_q^{\mathrm{pri}} = o_n\left((p^*-q^*)n^2/k\right). \tag{19}$$

Let c_{init} be some sufficiently small constant. For any initializer, $\pi^{(0)}$ satisfies $\ell(\pi^{(0)}, Z^*) \le c_{\mathrm{init}}\bar{n}_{\min}$ with probability at least $1-\epsilon$; there exist some constant $c > 0$ and some $\eta = o_n(1)$ such that in each iteration for the BCAVI algorithm, we have

$$\ell(\pi^{(s+1)}, Z^*) \le n\exp(-(1-\eta)\bar{n}_{\min}I) + \frac{\ell(\pi^{(s)}, Z^*)}{\sqrt{nI/[wk[n/\bar{n}_{\min}]^2]}} \quad \forall s \ge 0,$$

holds with probability at least $1 - \exp[-(\bar{n}_{\min}I)^{\frac{1}{2}}] - n^{-c} - \epsilon$.

Theorem 3.1 establishes a linear convergence rate for BCAVI algorithm with allowable initialization. The with-high-probability result holds for all iterations and all allowed initializations. The coefficient $[nI/[wk[n/\bar{n}_{\min}]^2]]^{-1/2}$ is independence of s, and goes to 0 when n grows. The following theorem is an immediate consequence of Theorem 3.1.

THEOREM 3.2. *Under the same condition as in Theorem 3.1, for any $s \ge s_0 \triangleq [nI/k]/\log[nI/[wk[n/\bar{n}_{\min}]^2]]$, we have*

$$\ell(\pi^{(s)}, Z^*) \le n\exp(-(1-2\eta)\bar{n}_{\min}I), \tag{20}$$

with probability at least $1 - \exp[-(\bar{n}_{\min}I)^{\frac{1}{2}}] - n^{-c} - \epsilon$.

Theorem 3.2 shows that BCAVI provably attains the $n\exp(-(1-o(1))\bar{n}_{\min}I)$ rate after at most s_0 iterations. When the network is sparse, that is, p^* and q^* are at most in an order of $(\log n)/n$, the quantity s_0 can be shown to be $o(\log n)$, and then BCAVI converges to be minimax rate within $\log n$ iterations. When the network is dense, that is, p^* and q^* are far bigger than $(\log n)/n$, $\log n$ iterations are not enough to attain the minimax rate. However, $\ell(\pi^{(s)}, Z^*) = o(n^{-a})$ for any $a > 0$ when $s \ge \log n$, and thus all the nodes can be correctly clustered with high probability by clustering each note to a community with the highest assignment probability. Therefore, it is enough to pick the number of iterations to be $\log n$ in implementing BCAVI.

THEOREM 3.3. *Under the assumption $nI/(k\log k) \to \infty$, we have*

$$\inf_{\hat{\pi}} \sup_{Z^* \in \Pi_0^{(\rho,\rho')}} \mathbb{E}\ell(\hat{\pi}, Z^*) \ge \begin{cases} n\exp(-(1-o(1))\rho nI/k), & k \ge 3; \\ n\exp(-(1-o(1))nI/2), & k = 2. \end{cases}$$

Theorem 3.3 gives the minimax lower bound for community detection problems with respect to the $\ell(\cdot,\cdot)$ loss. In Theorem 3.2, under the additional assumption that $Z^* \in \Pi_0^{(\rho,\rho')}$, we have $\bar{n}_{\min} = n/2$ when $k = 2$ and $\bar{n}_{\min} \geq \rho n I/k$ when $k \geq 3$, leading to the RHS of equation (20) upper bounded by

$$\begin{cases} n\exp(-(1-o(1))\rho n I/k), & k \geq 3; \\ n\exp(-(1-o(1))nI/2), & k = 2. \end{cases}$$

This immediately reveals that BCAVI converges to the minimax rate after s_0 iterations. As a consequence, BCAVI is not only computationally efficient, but also achieves statistical optimality. The minimax lower bound in Theorem 3.3 is almost identical to the minimaxity established in [34]. The only difference is that [34] consider a ℓ_0 loss function. The proof of Theorem 3.3 is just a routine extension of that in [34]. Therefore, we omit the proof.

To help understand Theorem 3.1, we add a remark on conditions on model parameters and priors, and a remark on initialization.

REMARK 1 (Conditions on model parameters and priors). The community sizes are not necessarily of the same order in Theorem 3.1. If we further assume ρ, ρ' are constants, and the prior $\pi_{i,a}^{\mathrm{pri}} \asymp 1/k, \forall i \in [n], a \in [k]$ (for example, uniform prior), and then the first condition in equation (19) is equivalent to

$$nI/k^3 \to \infty,$$

noting that $n/\bar{n}_{\min} \asymp k$ and $w \asymp 1$. This condition is necessary for consistent community detection [34] when k is finite. The assumptions in equation (19) is slightly stronger than the assumption in [22], which is essentially $nI \geq Ck^2 \log k$ for a sufficient large constant C.

Under the assumption $nI/k^3 \to \infty$, since we have $I \asymp (p^* - q^*)^2/p^*$, it can be shown that p^*, q^* are far bigger than n^{-1}, and then the second part of equation (19) can also be easily satisfied. For instance, we can simply set $\alpha_p^{\mathrm{pri}}, \beta_p^{\mathrm{pri}}, \alpha_q^{\mathrm{pri}}, \beta_q^{\mathrm{pri}}$ all equals to 1, that is, consider noninformative priors.

REMARK 2 (Initialization). The requirement on the initializers for BCAVI in Theorem 3.1 is relatively weak. When k is a constant and the community sizes are of the same order, the condition needed is $\ell(\pi^{(0)}, Z^*) \leq cn$ for some small constant c. Many existing methodologies in community detection literature can be used. One popular choice is spectral clustering. Established in [10, 12, 20], the spectral clustering has a misclustering error bound as $\mathcal{O}(k^2/I)$. From equation (19), the error is $o(\bar{n}_{\min})$, and then the condition that Theorem 3.1 requires for initialization is satisfied. The semidefinite programming (SDP), another popular method for community detection, also enjoys satisfactory theoretical guarantees [11, 16], and is suitable as an initializer.

4. Discussion.

4.1. *Statistical guarantee of global minimizer.* Though it is often challenging to obtain the global minimizer of the mean field method, it is still interesting to understand the statistical property of the global minimizer $\hat{\pi}^{\mathrm{MF}}$. Assume that both p^* and q^* are known, the optimization problem stated in Theorem 2.1 can be further simplified. The posterior distribution becomes $\mathbf{p}(Z|A)$. We use a product measure $\mathbf{q}_\pi(Z) = \prod_i \mathbf{q}_i(\pi_{i,\cdot})$ for approximation, and then $\hat{\pi}^{\mathrm{MF}} = \arg\min_{\pi \in \Pi_1} \mathrm{KL}[\mathbf{q}_\pi(Z) \| \mathbf{p}(Z|A)]$. Theorem 4.1 reveals that $\hat{\pi}^{\mathrm{MF}}$ is rate-optimal, not surprisingly given the theoretical results obtained for BCAVI, an approximation of $\hat{\pi}^{\mathrm{MF}}$.

THEOREM 4.1. *Let $Z^* \in \Pi_0^{(\rho,\rho')}$. Assume p^* and q^* are known. Under the assumption $\rho n I/[wk^2[n/\bar{n}_{\min}]^2] \to \infty$, there exist some constant $c > 0$ and $\eta = o_n(1)$ such that*

$$\ell(\hat{\pi}^{\mathrm{MF}}, Z^*) \leq n\exp(-(1-\eta)\bar{n}_{\min} I),$$

with probability at least $1 - \exp[-(\bar{n}_{\min} I)^{\frac{1}{2}}] - n^{-c}$.

4.2. *Gibbs sampling.* In Section 3.3, we analyze an iterative algorithm, BCAVI, and establish its linear convergence toward statistical optimality. The framework and methodology we establish is not limited to BCAVI, but can be extended to other iterative algorithms, including Gibbs sampling.

As a popular Markov chain Monte Carlo (MCMC) algorithm, Gibbs sampling has been widely used in practice to approximate the posterior distribution. There is a strong tie between Gibbs sampling and the mean field variational inference: both implement coordinate updates using conditional distributions. Using the general notation introduced in Section 2.1, to approximate $\mathbf{p}(x|y)$, Gibbs sampling obtains the update on x_i by a random generation from the conditional distribution $\mathbf{p}(x_i|x_{-i}, y)$, while the variational inference updates in a deterministic way with $\exp[\mathbb{E}_{\mathbf{q}_{-i}} \log \mathbf{p}(x_i|x_{-i}, y)]$.

We present a batched version of Gibbs sampling for community detection. It involves iterative updates with:

- Generate $p^{(s)}$ by sampling from $\mathbf{p}(p|q^{(s-1)}, Z^{(s-1)}, A)$.
- Generate $q^{(s)}$ by sampling from $\mathbf{p}(q|p^{(s-1)}, Z^{(s-1)}, A)$.
- Generate $Z_{i,\cdot}^{(s)}$ independently by sampling from $\mathbf{p}(Z_{i,\cdot}^{(s-1)}|Z_{-i,\cdot}^{(s-1)}, p^{(s)}, q^{(s)}, A)$, for $i \in [n]$.

We include the detailed implementation as Algorithm 2 in the Supplementary Material (Section A.1). The similarity between Algorithm 1 and Algorithm 2 makes it possible for us to analyze the output of Gibbs sampling in a similar way as we did for the variational inference.

THEOREM 4.2. *Under the same condition as in Theorem 3.1, for any initializer $Z^{(0)}$ satisfies $\ell(Z^{(0)}, Z^*) \leq c_{\text{init}}\bar{n}_{\min}$ with probability at least $1 - \epsilon$, there exist some constant $c > 0$ and some $\eta, \eta' = o_n(1)$ that go to 0 slowly, such that for all $s \geq 0$ of the batched Gibbs sampling (Algorithm 2), we have*

$$\mathbb{E}_{Z^{(s+1)}}\big[\ell(Z^{(s+1)}, Z^*)|A, Z^{(0)}\big] \leq n \exp\big(-(1-\eta)\bar{n}_{\min} I\big) + c_n^s \ell(Z^{(0)}, Z^*) + (s+1)nb_n$$

holds with probability at least $1 - \exp[-(\bar{n}_{\min} I)^{\frac{1}{2}})] - n^{-c} - \epsilon$, where $b_n = \exp[-\eta'^2 \bar{n}_{\min}^2] + \exp[-\eta'^2 n^2 I]$ and $c_n = 1/\sqrt{nI/[wk[n/\bar{n}_{\min}]^2]}$. Consequently, for $s = [nI/k]/\log[nI/[wk[n/\bar{n}_{\min}]^2]]$, we have

$$\mathbb{E}_{Z^{(s+1)}}\big[\ell(Z^{(s+1)}, Z^*)|A, Z^{(0)}\big] \leq n \exp\big(-(1-2\eta)\bar{n}_{\min} I\big),$$

with probability at least $1 - \exp[-(\bar{n}_{\min} I)^{\frac{1}{2}}] - n^{-c} - \epsilon$.

Theorem 4.2 establishes theoretical justification for batched Gibbs sampling for community detection. Note that c_n is the same as in Theorem 3.1. When $s \leq e^n$, the additional term $(s+1)nb_n$ is dominated by $n \exp(-(1-\eta)\bar{n}_{\min} I)$, indicating that the batched Gibbs sampling has similar linear convergence as that in Theorem 3.1.

The additional term $(s+1)nb_n$ arises as we attempt to exclude extreme events occurring in each iteration caused by sampling. For instance, even under the assumption that we have obtained the true parameters p^*, q^*, Z^*, if we sample from the conditional distribution $\mathbf{p}(Z_{i,\cdot}|Z_{-i,\cdot}^*, p^*, q^*, A), \forall i \in [n]$, there still exists a nonzero probability such that the new Z generated behaves like a random guess. If this happens, the new Z will be a bad initialization for the upcoming iterations. Albeit these are extremely small probability events, they grow with s. When $s > e^n$, the term $(s+1)nb_n$ will become dominating. However, as we all know, the Gibbs sampling often converges to the posterior distribution, which indicates that the error rate should remain unchanged after sufficient iterations. We hope in the future the additional term $(s+1)nb_n$ can be removed with more advanced technical tools developed.

4.3. *An iterative algorithm for maximum likelihood estimation.* Maximum likelihood estimator (MLE) usually yields statistical optimality. However, the maximization of the likelihood $\mathbf{p}(A|Z, p, q)$ over Z, p, q is computationally infeasible. Inspired by the procedures proposed in Algorithm 1 and Algorithm 2, we may approach $\max \mathbf{p}(A|Z, p, q)$ by alternating maximization. We use a batched coordinate maximization:

- Maximize $\mathbf{p}(A|p, q^{(s-1)}, Z^{(s-1)})$ over p to obtain $p^{(s)}$.
- Maximize $\mathbf{p}(A|p^{(s-1)}, q, Z^{(s-1)})$ over q to obtain $q^{(s)}$.
- Maximize $\mathbf{p}(A|p^{(s-1)}, q^{(s-1)}, Z_{i,\cdot}, Z_{-i,\cdot}^{(s-1)})$ over $Z_{i,\cdot}$ to obtain $Z_{i,\cdot}^{(s)}$, for each $i \in [n]$.

We include its detailed implementation in Algorithm 3 in the Supplementary Material (Section A.2). We have the following theoretical guarantee of this iterative algorithm to approximate the MLE.

THEOREM 4.3. *Under the same condition as in Theorem* 3.1, *for any initializer* $Z^{(0)}$ *satisfies* $\ell(Z^{(0)}, Z^*) \le c_{\text{init}}\bar{n}_{\min}$ *with probability at least* $1-\epsilon$, *there exist some constant* $c > 0$ *and some* $\eta = o_n(1)$, *such that for all* $s \ge 0$ *of the iterative algorithm of MLE* (*Algorithm* 3), *we have*

$$\ell\big(Z^{(s+1)}, Z^*\big) \le n \exp\big(-(1-\eta)\bar{n}_{\min} I\big) + \frac{\ell(Z^{(s)}, Z^*)}{\sqrt{nI/[wk[n/\bar{n}_{\min}]^2]}}$$

holds with probability at least $1 - \exp[-(\bar{n}_{\min} I)^{\frac{1}{2}}] - n^{-c} - \epsilon$.

Algorithm 3 is essentially the same with the procedure proposed in [12]. However, [12] can only analyze the performance of one single iteration from $Z^{(0)}$ (i.e., $\ell(Z^{(1)}, Z^*)$), and it requires extra data splitting steps. Theorem 4.3 provides a stronger and cleaner result compared with that of [12].

4.4. *Extension to more general SBMs.* In the same way as many other community detection literatures, we assume the network is assortative throughout this paper. However, the relative value of p^*, q^* plays a minimal role in the proofs of Theorem 3.1 and others. What really matters is the key quantity I which captures the difference between p^* and q^*. All the theoretical results established in this paper hold if we instead assume the network is disassortative, that is, $p^* < q^*$.

The SBM studied in this paper is homogeneous, in the sense that all the within-community connection probabilities are equally p and all the between-community ones are equally q. A less restricted model is to allow heterogeneousness. Let $p_{a,b}$ be the probability of connection between two communities $a, b \in [k]$. On the algorithmic side, we can have priors on $\{p_{a,b}\}_{a\le b}$ and have them updated analogously to equations (12) and (13). Updates on π will be more complicated than that in equation (11) as it involves $\{t_{a,b}\}_{a,b}$ and $\{\lambda_{a,b}\}_{a,b}$. On the theoretical side, we may still assume the network has sort of assortative or disassortative structure, for instance, $\max_{a<b} p_{a,b} \le q^* < p^* \le \min_a p_{a,a}$, so that the quantity I is still one possible way to measure the difficulty of distinguishing different communities. Further investigation is beyond the scope of this paper.

5. Numeric studies. In this section, we present numeric performances of BCAVI (Algorithm 1), the batched Gibbs sampling (Algorithm 2) and the iterative algorithm for MLE (Algorithm 3) on synthetic data. We set $n = 2000$ and $k = 10$. The sizes of communities are equal (thus, $n_{\min} = 200$). The within-community connection probability $p^* = 0.17$ and the between-community one $q^* = 0.08$. We initialize by spectral clustering, and then execute the

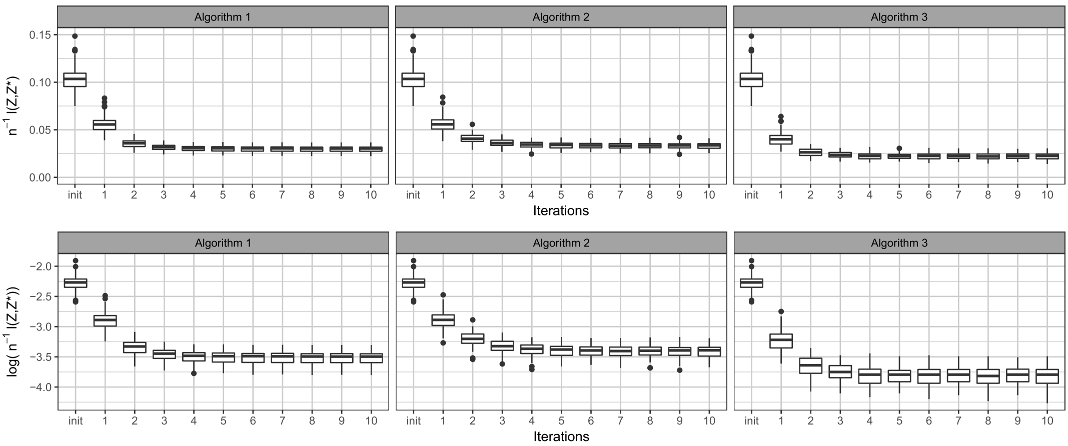

FIG. 2. *Numeric performances of Algorithms* 1, 2 *and* 3 *on synthetic data. Top*: *the y-axis is the misclassification proportation measured by $n^{-1}\ell(Z, Z^*)$ where $\ell(\cdot,\cdot)$ is defined as in equation* (17). *Bottom*: *after a logarithm transformation*, *that is*, $\log(n^{-1}\ell(Z, Z^*))$. *The "init" in the x-axis indicates errors of the initializer* (*spectral clustering*).

aforementioned algorithms separately, each with 10 iterations. The results are reported in Figure 2, based on 100 independent draws from the underlying SBM. We run the batched Gibbs sampling 10 times and report the mean value, as we are interested in its average performance.

In Figure 2, all three algorithms have similar behaviors. Initialized by spectral clustering, their error rates decrease linearly before convergence. It is interesting to observe that Algorithm 3 has a slightly smaller error rate than Algorithm 1, which is slightly better than Algorithm 2. Nevertheless, they all achieve the optimal rate which is around $\exp(-n_{\min}I) \approx 0.022$, up to some $\exp(-o(1)n_{\min}I)$ factor. Overall, the numeric results presented in Figure 2 are consistent with the theoretical justifications established in Theorems 3.1, 4.2 and 4.3.

6. Proofs of main theorems. In this section, we give proofs of the theorems in Section 2 and Section 3. We first present the proof of Theorem 2.1 in Section 6.1. Then we give the proof Theorem 2.2 in Section 6.2. The proof of Theorem 3.1 is given in Section 6.3.

6.1. *Proof of Theorem* 2.1. From equation (8), by some algebra (see equation (56) in Appendix D for detailed derivation) we have

$$\begin{aligned}&(\hat{\pi}^{\mathrm{MF}}, \hat{\alpha}_p^{\mathrm{MF}}, \hat{\beta}_p^{\mathrm{MF}}, \hat{\alpha}_q^{\mathrm{MF}}, \hat{\beta}_q^{\mathrm{MF}}) \\ &\quad= \mathop{\arg\min}_{\substack{\pi\in\Pi_1\\ \alpha_p,\beta_p,\alpha_q,\beta_q>0}} \mathbb{E}_{\mathbf{q}}[\log \mathbf{p}(A|Z,p,q)] - \mathrm{KL}(\mathbf{q}(Z,p,q)\|\mathbf{p}(Z,p,q)),\end{aligned} \tag{21}$$

where we use $\mathbf{q}$ instead of $\mathbf{q}_{\pi,\alpha_p,\beta_p,\alpha_q,\beta_q}$ for simplicity. From the conditional distribution in equation (6), the log-likelihood function can be simplified as

$$\log \mathbf{p}(A|Z,p,q) = \sum_{a,b}\sum_{i<j} Z_{ia}Z_{jb}\left[A_{i,j}\log\frac{B_{ab}}{1-B_{ab}} + \log(1-B_{ab})\right].$$

Due to the independence of Z and p, q under $\mathbf{q}$, we have

$$\begin{aligned}\mathbb{E}_{\mathbf{q}}[\log \mathbf{p}(A|Z,p,q)] &= \mathbb{E}_{\mathbf{q}(p,q)}\left[\mathbb{E}_{\mathbf{q}(Z)}\left[\sum_{a,b}\sum_{i<j} Z_{i,a}Z_{j,b}\left[A_{i,j}\log\frac{B_{ab}}{1-B_{ab}} + \log(1-B_{ab})\right]\right]\right] \\ &= \mathbb{E}_{\mathbf{q}(p,q)}\left[\sum_{a,b}\sum_{i<j} \pi_{i,a}\pi_{j,b}\left[A_{i,j}\log\frac{B_{ab}}{1-B_{ab}} + \log(1-B_{ab})\right]\right].\end{aligned}$$

Since $B_{a,a} = p, \forall a \in [k]$ and $B_{a,b} = q, \forall a \neq b$, we have

$$\begin{aligned}\mathbb{E}_{\mathbf{q}}[\log \mathbf{p}(A|Z, p, q)] &= \mathbb{E}_{\mathbf{q}(p,q)}\biggl[\sum_{a}\sum_{i<j} \pi_{i,a}\pi_{j,a}\biggl[A_{i,j}\log\frac{p(1-q)}{q(1-p)} + \log\frac{1-p}{1-q}\biggr]\biggr] \\ &\quad + \mathbb{E}_{\mathbf{q}(p,q)}\biggl[\sum_{a,b}\sum_{i<j} \pi_{i,a}\pi_{j,b}\biggl[A_{i,j}\log\frac{q}{1-q} + \log(1-q)\biggr]\biggr].\end{aligned} \tag{22}$$

By properties of Beta distribution, we obtain

$$\begin{aligned}\mathbb{E}_{\mathbf{q}(p,q)}\log\frac{p(1-q)}{q(1-p)} &= \mathbb{E}_{\mathbf{q}(p)}[\log p - \log(1-p)] - \mathbb{E}_{\mathbf{q}(q)}[\log q - \log(1-q)] \\ &= [\psi(\alpha_p) - \psi(\beta_p)] - [\psi(\alpha_q) - \psi(\beta_q)]\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_{\mathbf{q}(p,q)}\log\frac{1-q}{1-p} &= \mathbb{E}_{\mathbf{q}(q)}\log(1-q) - \mathbb{E}_{\mathbf{q}(p)}\log(1-p) \\ &= [\psi(\beta_q) - \psi(\alpha_q + \beta_q)] - [\psi(\beta_p) - \psi(\alpha_p + \beta_p)].\end{aligned}$$

This leads to

$$\begin{aligned}&\mathbb{E}_{\mathbf{q}(p,q)}\biggl[\sum_{a}\sum_{i<j} \pi_{i,a}\pi_{j,a}\biggl[A_{i,j}\log\frac{p(1-q)}{q(1-p)} + \log\frac{1-p}{1-q}\biggr]\biggr] \\ &\quad = 2t\biggl[\sum_{a}\sum_{i<j}\pi_{i,a}\pi_{j,a}(A_{i,j} - \lambda)\biggr] \\ &\quad = t\langle A - \lambda 1_n 1_n^T + \lambda I_n, \pi\pi^T\rangle.\end{aligned} \tag{23}$$

Similarly, we can obtain

$$\begin{aligned}&\mathbb{E}_{\mathbf{q}(p,q)}\biggl[\sum_{a,b}\sum_{i<j}\pi_{i,a}\pi_{j,b}\biggl[A_{i,j}\log\frac{q}{1-q} + \log(1-q)\biggr]\biggr] \\ &\quad = \biggl[\mathbb{E}_{\mathbf{q}(q)}\log\frac{q}{1-q}\biggr]\sum_{i<j}A_{i,j}\sum_{a,b}\pi_{i,a}\pi_{j,b} + [\mathbb{E}_{\mathbf{q}(q)}\log(1-q)]\sum_{i<j}\sum_{a,b}\pi_{i,a}\pi_{j,b} \\ &\quad = \frac{1}{2}[\psi(\alpha_q) - \psi(\beta_q)]\|A\|_1 + \frac{n(n-1)}{2}[\psi(\beta_q) - \psi(\alpha_q + \beta_q)],\end{aligned} \tag{24}$$

where we use the fact that $(\sum_a \pi_{i,a}) = \|\pi_{i,\cdot}\|_1 = 1, \forall i \in [n]$. Now consider the Kullback–Leibler divergence between $\mathbf{q}(Z, p, q)$ and $\mathbf{p}(Z, p, q)$. Due to the independence of p, q and $\{Z_{i,\cdot}\}_{i=1}^n$ in both distributions, we have

$$\begin{aligned}\mathrm{KL}(\mathbf{q}(Z, p, q)\|\mathbf{p}(Z, p, q)) &= \mathrm{KL}(\mathbf{q}(Z)\|\mathbf{p}(Z)) + \mathrm{KL}(\mathbf{q}(p)\|\mathbf{p}(p)) + \mathrm{KL}(\mathbf{q}(q)\|\mathbf{p}(q)) \\ &= \sum_{i=1}^{n}\mathrm{KL}[\mathrm{Categorical}(\pi_{i,\cdot})\,\|\,\mathrm{Categorical}(\pi_{i,\cdot}^{\mathrm{pri}})] \\ &\quad + \mathrm{KL}[\mathrm{Beta}(\alpha_p, \beta_p)\,\|\,\mathrm{Beta}(\alpha_p^{\mathrm{pri}}, \beta_p^{\mathrm{pri}})] \\ &\quad + \mathrm{KL}[\mathrm{Beta}(\alpha_q, \beta_q)\,\|\,\mathrm{Beta}(\alpha_q^{\mathrm{pri}}, \beta_q^{\mathrm{pri}})].\end{aligned} \tag{25}$$

By equations (21)–(25), we conclude with the desired result.

6.2. *Proof of Theorem* 2.2. Note that

$$B_{z_i,z_j} = \Bigg[\sum_{a=1}^{k} Z_{i,a} Z_{j,a}\Bigg] p + \Bigg[\sum_{a\neq b} Z_{i,a} Z_{j,b}\Bigg] q.$$

We rewrite the joint distribution $\mathbf{p}(p, q, z, A)$ in equation (7) as follows:

$$\begin{aligned}
\mathbf{p}(p, q, Z, A) = {}& \Bigg[\prod_{i=1}^{n} \pi_{i,z_i}^{\mathrm{pri}}\Bigg]\Bigg[\prod_{i<j}\big[p^{A_{i,j}}(1-p)^{1-A_{i,j}}\big]^{\sum_{a=1}^{k} Z_{i,a}Z_{j,a}}\Bigg] \\
& \times \Bigg[\prod_{i<j}\big[q^{A_{i,j}}(1-q)^{1-A_{i,j}}\big]^{\sum_{a\neq b}^{k} Z_{i,a}Z_{j,b}}\Bigg] \\
& \times \Bigg[\frac{\Gamma(\alpha_p^{\mathrm{pri}}+\beta_p^{\mathrm{pri}})}{\Gamma(\alpha_p^{\mathrm{pri}})\Gamma(\beta_p^{\mathrm{pri}})} p^{\alpha_p^{\mathrm{pri}}-1}(1-p)^{\beta_p^{\mathrm{pri}}-1}\Bigg] \\
& \times \Bigg[\frac{\Gamma(\alpha_q^{\mathrm{pri}}+\beta_q^{\mathrm{pri}})}{\Gamma(\alpha_q^{\mathrm{pri}})\Gamma(\beta_q^{\mathrm{pri}})} q^{\alpha_q^{\mathrm{pri}}-1}(1-q)^{\beta_q^{\mathrm{pri}}-1}\Bigg].
\end{aligned} \tag{26}$$

Updates on p and q. From equation (26), p has conditional probability as

$$\mathbf{p}(p|q, Z, A) \propto \Bigg[\prod_{i<j}\big[p^{A_{i,j}}(1-p)^{1-A_{i,j}}\big]^{\sum_{a=1}^{k} Z_{i,a}Z_{j,a}}\Bigg]\Bigg[\frac{\Gamma(\alpha_p^{\mathrm{pri}}+\beta_p^{\mathrm{pri}})}{\Gamma(\alpha_p^{\mathrm{pri}})\Gamma(\beta_p^{\mathrm{pri}})} p^{\alpha_p^{\mathrm{pri}}-1}(1-p)^{\beta_p^{\mathrm{pri}}-1}\Bigg].$$

Then the CAVI update in equation (4) leads to

$$\begin{aligned}
\hat{\mathbf{q}}(p) &\propto \exp\big[\mathbb{E}_{\mathbf{q}(q,Z)} \log \mathbf{p}(p|q, Z, A)\big] \\
&\propto \exp\Bigg[\mathbb{E}_{\mathbf{q}(Z)} \sum_{i<j}\sum_{a=1}^{k} Z_{i,a}Z_{j,a} \log\big[p^{A_{i,j}}(1-p)^{1-A_{i,j}}\big]\Bigg] \\
&\quad \times \Bigg[\frac{\Gamma(\alpha_p^{\mathrm{pri}}+\beta_p^{\mathrm{pri}})}{\Gamma(\alpha_p^{\mathrm{pri}})\Gamma(\beta_p^{\mathrm{pri}})} p^{\alpha_p^{\mathrm{pri}}-1}(1-p)^{\beta_p^{\mathrm{pri}}-1}\Bigg] \\
&= \exp\Bigg[\sum_{i<j}\sum_{a=1}^{k} \pi_{i,a}\pi_{j,a} \log\big[p^{A_{i,j}}(1-p)^{1-A_{i,j}}\big]\Bigg]\Bigg[\frac{\Gamma(\alpha_p^{\mathrm{pri}}+\beta_p^{\mathrm{pri}})}{\Gamma(\alpha_p^{\mathrm{pri}})\Gamma(\beta_p^{\mathrm{pri}})} p^{\alpha_p^{\mathrm{pri}}-1}(1-p)^{\beta_p^{\mathrm{pri}}-1}\Bigg].
\end{aligned}$$

It can be written as

$$\begin{aligned}
\hat{\mathbf{q}}(p) \propto {}& \big[p^{\sum_{i<j}\sum_{a=1}^{k}\pi_{i,a}\pi_{j,a}A_{i,j}}(1-p)^{\sum_{i<j}\sum_{a=1}^{k}\pi_{i,a}\pi_{j,a}(1-A_{i,j})}\big] \\
& \times \Bigg[\frac{\Gamma(\alpha_p^{\mathrm{pri}}+\beta_p^{\mathrm{pri}})}{\Gamma(\alpha_p^{\mathrm{pri}})\Gamma(\beta_p^{\mathrm{pri}})} p^{\alpha_p^{\mathrm{pri}}-1}(1-p)^{\beta_p^{\mathrm{pri}}-1}\Bigg].
\end{aligned}$$

The distribution of p is still Beta $p \sim \mathrm{Beta}(\alpha'_p, \beta'_p)$, with

$$\alpha'_p = \alpha_p^{\mathrm{pri}} + \sum_{i<j}\sum_{a=1}^{k} \pi_{i,a}\pi_{j,a}A_{i,j} \quad \text{and} \quad \beta'_p = \beta_p^{\mathrm{pri}} + \sum_{i<j}\sum_{a=1}^{k} \pi_{i,a}\pi_{j,a}(1-A_{i,j}).$$

Similar analysis on q yields updates on α'_q and β'_q. Hence, its proof is omitted.

Updates on $\{Z_{i,\cdot}\}_{i=1}^n$. From equation (26), the conditional distribution on $Z_{i,\cdot}$ is

$$\mathbf{p}(Z_{i,\cdot}|Z_{-i,\cdot}, p, q, A) \propto \pi_{i,z_i}^{\text{pri}}\left[\prod_{j\neq i} B_{z_i,z_j}^{A_{i,j}}(1-B_{z_i,z_j})^{1-A_{i,j}}\right].$$

Consequently, up to a constant not depending on i, we have

$$\begin{aligned}
&\log \mathbb{P}(Z_{i,a}=1|Z_{-i,\cdot}, p, q, A)\\
&\quad= \log \pi_{i,a}^{\text{pri}} + \left[\sum_{j\neq i} Z_{j,a}\left[A_{i,j}\log\frac{p}{1-p}+\log(1-p)\right]\right.\\
&\qquad\left.+\sum_{j\neq i}\sum_{b\neq a} Z_{j,b}\left[A_{i,j}\log\frac{q}{1-q}+\log(1-q)\right]\right]\\
&\quad= \log \pi_{i,a}^{\text{pri}} + \left[\sum_{j\neq i} Z_{j,a}\left[A_{i,j}\log\frac{p(1-q)}{q(1-p)}-\log\frac{1-q}{1-p}\right]\right.\\
&\qquad\left.+\sum_{j\neq i}\left[A_{i,j}\log\frac{q}{1-q}+\log(1-q)\right]\right].
\end{aligned}$$

Then the CAVI update from equation (4) leads to

$$\begin{aligned}
\pi'_{i,a} &= \hat{\mathbf{q}}_{Z_{i,\cdot}}(Z_{i,a}=1)\\
&\propto \exp\left[\mathbb{E}_{\mathbf{q}(p,q,z_{-i})}\log\mathbb{P}(Z_{i,a}=1|Z_{-i,\cdot}, p, q, A)\right]\\
&= \exp\left[\mathbb{E}_{\mathbf{q}(p)}\mathbb{E}_{\mathbf{q}(q)}\mathbb{E}_{\mathbf{q}(Z_{-i,\cdot})}\log\mathbb{P}(Z_{i,}=1|Z_{-i,\cdot}, p, q, A)\right]\\
&\propto \pi_{i,a}^{\text{pri}}\exp\left[\mathbb{E}_{\mathbf{q}(p)}\mathbb{E}_{\mathbf{q}(q)}\sum_{j\neq i}\pi_{j,a}\left[A_{i,j}\log\frac{p(1-q)}{q(1-p)}-\log\frac{1-q}{1-p}\right]\right],
\end{aligned}\tag{27}$$

where we use the property that p, q, Z are all independent of each other under $\mathbf{q}$. Recall that $p \sim \text{Beta}(\alpha_p, \beta_p)$ and $q \sim \text{Beta}(\alpha_q, \beta_q)$. It can be shown that

$$\mathbb{E}_{\mathbf{q}(p)}\log\frac{p}{1-p} = \psi(\alpha_p)-\psi(\beta_p) \quad\text{and}\quad \mathbb{E}_{\mathbf{q}(p)}\log(1-p) = \psi(\beta_p)-\psi(\alpha_p+\beta_p),$$

where $\psi(\cdot)$ is digamma function. Similar results hold for $\mathbb{E}_{\mathbf{q}(q)}\log(q/(1-q))$ and $\mathbb{E}_{\mathbf{q}(q)}\log(1-q)$. Plug in these expectations to equation (27), we have

$$\pi'_{i,a} \propto \pi_{i,a}^{\text{pri}}\exp\left[2t\sum_{j\neq i}\pi_{j,a}(A_{i,j}-\lambda)\right].$$

6.3. *Proof of Theorem* 3.1. Theorem 3.1 gives a theoretical justification for all iterations in the BCAVI algorithm. Due to the limit of pages, in this section we assume $\ell(\pi^{(0)}, Z^*) = o(\bar{n}_{\min})$. The proof of the case $\ell(\pi^{(0)}, Z^*)$ in a constant order of $\bar{n}_{\min}$ is essentially the same with slight modification, and we defer it to Section B.2 in the Supplementary Material.

PROOF. Let $\gamma = o(1)$ be any sequence that goes to zero when n grows. Define t^* and λ^* as the true counterparts of t and λ, by

$$t^* = \frac{1}{2}\log\frac{p^*(1-q^*)}{q^*(1-p^*)} \quad\text{and}\quad \lambda^* = \frac{1}{2t^*}\log\frac{1-q^*}{1-p^*}.$$

The key to prove Theorem 3.1 is to develop a uniform analysis for the mapping $h(\cdot)$ (cf. equation (11)). In Theorem 6.1, we will show $\ell(\cdot, Z^*)$ decreases in a desired way for a *single* implementation of $h(\cdot)$, *uniformly* for all π, t, λ that are in some neighborhoods of their true values.

THEOREM 6.1 (Uniform analysis for $h(\cdot)$). *Consider any $\pi \in \Pi_1$ such that $\|\pi - Z^*\|_1 \leq \gamma \bar{n}_{\min}$. Let η' be any sequence such that $\eta' = o(1)$. Consider any t and λ with $|t - t^*| \leq \eta'(p^* - q^*)/p^*$ and $|\lambda - \lambda^*| \leq \eta'(p^* - q^*)$. We define $\mathcal{F}$ to be the event, that after applying the mapping $h_{t,\lambda}(\cdot)$, there exists some $\eta = o(1)$ such that*

$$\|h_{t,\lambda}(\pi) - Z^*\|_1 \leq n \exp\bigl(-(1-\eta)\bar{n}_{\min} I\bigr) + \frac{\|\pi - Z^*\|_1}{\sqrt{nI/[wk[n/\bar{n}_{\min}]^2]}},$$

holds uniformly over all the eligible π, t and λ. Under the same assumption as Theorem 3.1, we have

$$\mathbb{P}(\mathcal{F}) \geq 1 - \exp\bigl[-(\bar{n}_{\min} I)^{\frac{1}{2}})\bigr] - n^{-r}$$

for some constant $r > 0$.

Once Theorem 6.1 is established, we can extend it to develop a uniform one-step analysis for BCAVI iterations. For any $\pi \in \Pi_1$, define

$$\alpha_p = \alpha_p^{\text{pri}} + \sum_{a=1}^{k} \sum_{i<j} A_{i,j} \pi_{i,a} \pi_{j,a}, \qquad \beta_p = \beta_p^{\text{pri}} + \sum_{a=1}^{k} \sum_{i<j} (1 - A_{i,j}) \pi_{i,a} \pi_{j,a} \tag{28}$$

and

$$\alpha_q = \alpha_q^{\text{pri}} + \sum_{a \neq b} \sum_{i<j} A_{i,j} \pi_{i,a} \pi_{j,b}, \qquad \beta_q = \beta_q^{\text{pri}} + \sum_{a \neq b} \sum_{i<j} (1 - A_{i,j}) \pi_{i,a} \pi_{j,b}, \tag{29}$$

and consequently,

$$t = \frac{1}{2}\bigl[[\psi(\alpha_p) - \psi(\beta_p)] - [\psi(\alpha_q) - \psi(\beta_q)]\bigr], \tag{30}$$

$$\lambda = \frac{1}{2t}\bigl[[\psi(\beta_q) - \psi(\alpha_q + \beta_q)] - [\psi(\beta_p) - \psi(\alpha_p + \beta_p)]\bigr]. \tag{31}$$

From Lemma C.1, we have a concentration of t, λ toward t^*, λ^*. That is, there exists some $\eta' = o(1)$, such that with probability at least $1 - e^3 5^{-n}$, the following inequalities hold:

$$|t - t^*| \leq \eta'(p^* - q^*)/p^* \quad \text{and} \quad |\lambda - \lambda^*| \leq \eta'(p^* - q^*),$$

uniformly over all the eligible π such that $\|\pi - Z^*\|_1 \leq \gamma \bar{n}_{\min}$. Therefore, from Theorem 6.1 and Lemma C.1, we have Theorem 6.2 for a single BCAVI iteration.

THEOREM 6.2 (Uniform one-step analysis for BCAVI). *Consider any $\pi \in \Pi_1$ such that $\|\pi - Z^*\|_1 \leq \gamma \bar{n}_{\min}$. Define $\alpha_p, \beta_p, \alpha_q, \beta_q, t, \lambda$ as functions of π by equations (28)–(31). We define $\mathcal{F}$ to be the event, that after applying the mapping $h_{t,\lambda}(\cdot)$, there exists some $\eta = o(1)$ such that*

$$\|h_{t,\lambda}(\pi) - Z^*\|_1 \leq n \exp\bigl(-(1-\eta)\bar{n}_{\min} I\bigr) + \frac{\|\pi - Z^*\|_1}{\sqrt{nI/[wk[n/\bar{n}_{\min}]^2]}},$$

holds uniformly over all the eligible π. Under the same assumption as Theorem 3.1, we have

$$\mathbb{P}(\mathcal{F}) \geq 1 - \exp\bigl[-(\bar{n}_{\min} I)^{\frac{1}{2}})\bigr] - n^{-r}$$

for some constant $r > 0$.

Theorem 6.2 is sufficient to prove Theorem 3.1. Once the event $\mathcal{F}$ defined in Theorem 6.2 holds, we can apply Theorem 6.2 iteratively on BCAVI iterations, and the loss will decrease in the desired way. □

The only thing left unproved is Theorem 6.1. We provide a proof sketch in Section 6.3.1 and a detailed proof in Section 6.3.2.

6.3.1. *Proof sketch of Theorem* 6.1. The estimation $[h_{t,\lambda}(\pi)]_{i,\cdot}$ is a function of π and $A_{i,\cdot}$. Its error comes from two different sources. The first source is more fundamental. That is, even if we implement the mapping $h(\cdot)$ on the true value Z^*, we will still make some mistakes, as we only observe a noisy A. This contributes to the minimax rate. The second source comes from the fact that we have π instead of Z^*. The magnitude of this error is related to the deviation $\|\pi - Z^*\|_1$. Since these two errors behave differently, we decompose $[h_{t,\lambda}(\pi)]_{i,\cdot} - Z^*_{i,\cdot}$ accordingly into two parts, using generic notation $f_{i,1}$, $f_{i,2}$:

$$\|[h_{t,\lambda}(\pi)]_{i,\cdot} - Z^*_{i,\cdot}\|_1 \le f_{i,1}(Z^*, A_{i,\cdot}) + f_{i,2}(\pi - Z^*, A_{i,\cdot}).$$

Consequently,

$$\|h_{t,\lambda}(\pi) - Z^*\|_1 \le \underbrace{\sum_{i=1}^{n} f_{i,1}(Z^*, A_{i,\cdot})}_{\text{involves } Z^*} + \underbrace{\sum_{i=1}^{n} f_{i,2}(\pi - Z^*, A_{i,\cdot})}_{\text{involves } \pi - Z^*}. \tag{32}$$

The first term on the RHS of equation (32) leads to the minimax rate $n\exp(-(1-\eta)\bar{n}_{\min}I)$. The second term is upper bounded by

$$\sum_{i=1}^{n} f_{i,2}(\pi - Z^*, A_{i,\cdot}) \le s\|A - \mathbb{E}A\|_{\mathrm{op}}^2 \|\pi - Z^*\|_1,$$

involving the spectral norm of $A - \mathbb{E}A$ and the deviation between π and Z^*, where s is not dependent on π, Z^* or A. The coefficient $s\|A - \mathbb{E}A\|_{\mathrm{op}}^2$ will be shown to be smaller than 1.

6.3.2. *Proof of Theorem* 6.1. Denote $z = r^{-1}(Z^*)$. By the definition of $h_{t,\lambda}(\cdot)$ in equation (11), we have

$$\begin{aligned}\|[h_{t,\lambda}(\pi)]_{i,\cdot} - Z^*_{i,\cdot}\|_1 &\le \frac{2\sum_{a\neq z_i} \pi^{\mathrm{pri}}_{i,a} \exp[2t\sum_{j\neq i}\pi_{j,a}(A_{i,j} - \lambda)]}{\sum_a \pi^{\mathrm{pri}}_{i,a}\exp[2t\sum_{j\neq i}\pi_{j,a}(A_{i,j}-\lambda)]} \\ &\le 2w\sum_{a\neq z_i} 1 \wedge \exp\Big[2t\sum_{j\neq i}(\pi_{j,a} - \pi_{j,z_i})(A_{i,j} - \lambda)\Big],\end{aligned}$$

where the last inequality is due to the fact $x/(x+y) \le \frac{x/y}{x/y+1} \le \min\{1, x/y\}$, $\forall x, y > 0$ and that $|\pi^{\mathrm{pri}}_{i,a}/\pi^{\mathrm{pri}}_{i,z_i}| \le w$, $\forall a$ by the definition of w in equation (18).

We are going to upper bound $f(x) = 1 \wedge \exp(-x)$ by a step function. Let $x_0 < 0$ and m be some integer, whose values will be determined later. The interval $(x_0, 0]$ can be divided into m equal-length and disjoint segments. For each $l = 0, 1, \dots, (m-1)$, the value of f on $((l+1)x_0/m, lx_0/m]$ is upper bounded by $\exp(lx_0/m)$, due to monotonicity. Hence, we have

$$f(x) \le \exp(x_0) + \sum_{l=0}^{m-1} \exp\Big[\frac{lx_0}{m}\Big]\mathbb{I}\Big[x \ge \frac{(l+1)x_0}{m}\Big].$$

By taking $x_0 = -(n_a + n_{z_i})I/2$ and letting $x = 2t\sum_{j\neq i}(\pi_{j,a} - \pi_{j,z_i})(A_{i,j} - \lambda)$, we have

$$\begin{aligned}\|[h_{t,\lambda}(\pi)]_{i,\cdot} - Z^*_{i,\cdot}\|_1 \le{}& 2w\sum_{a\neq z_i}\exp\Big[-\frac{(n_a + n_{z_i})I}{2}\Big] + 2w\sum_{a\neq z_i}\sum_{l=0}^{m-1}\Big[\exp\Big[-\frac{l(n_a+n_{z_i})I}{2m}\Big] \\ &\times \mathbb{I}\Big[2t\sum_{j\neq i}(\pi_{j,a} - \pi_{j,z_i})(A_{i,j} - \lambda) \ge -\frac{(l+1)(n_a + n_{z_i})I}{2m}\Big]\Big].\end{aligned}$$

Summing over $i \in [n]$, we obtain

$$\begin{aligned}
\|h_{t,\lambda}(\pi) - Z^*\|_1 &= \sum_{i=1}^{n} \|[h_{t,\lambda}(\pi)]_{i,\cdot} - Z^*_{i,\cdot}\|_1 = \sum_{b=1}^{k} \sum_{i:z_i=b} \|[h_{t,\lambda}(\pi)]_{i,\cdot} - Z^*_{i,\cdot}\|_1 \\
&\le 2w \sum_{b=1}^{k} \sum_{i:z_i=b} \sum_{a\neq z_i} \exp\biggl[-\frac{(n_a+n_b)I}{2}\biggr] \\
&\quad + 2w \sum_{b=1}^{k} \sum_{i:z_i=b} \sum_{a\neq b} \sum_{l=0}^{m-1} \biggl[\exp\biggl[-\frac{l(n_a+n_b)I}{2m}\biggr] \\
&\quad \times \mathbb{I}\biggl[2t \sum_{j\neq i} (\pi_{j,a} - \pi_{j,b})(A_{i,j} - \lambda) \ge -\frac{(l+1)(n_a+n_b)I}{2m}\biggr]\biggr].
\end{aligned}$$

Using the fact that $\min_{a\neq b}(n_a + b_b)/2 \ge \bar{n}_{\min}$, we have

$$\begin{aligned}
\|h_{t,\lambda}(\pi) - Z^*\|_1 &\le 2w \sum_{i=1}^{n} \sum_{a\neq z_i} \exp(-\bar{n}_{\min} I) + 2w \sum_{b=1}^{k} \sum_{a\neq b} \sum_{l=0}^{m-1} \biggl[\exp\biggl[-\frac{l(n_a+n_b)I}{2m}\biggr] \\
&\quad \times \sum_{i:z_i=b} \mathbb{I}\biggl[2t \sum_{j\neq i} (\pi_{j,a} - \pi_{j,b})(A_{i,j} - \lambda) \ge -\frac{(l+1)(n_a+n_b)I}{2m}\biggr]\biggr] \\
&\le 2wnk \exp(-\bar{n}_{\min} I) + 2w \sum_{l=0}^{m-1} \sum_{a=1}^{k} \sum_{b\neq a} \biggl[\exp\biggl[-\frac{l(n_a+n_b)I}{2m}\biggr] \\
&\quad \times \sum_{i:z_i=b} \mathbb{I}\biggl[\sum_{j\neq i} (\pi_{j,a} - \pi_{j,b})(A_{i,j} - \lambda) \ge -\frac{(l+1)(n_a+n_b)I}{4mt}\biggr]\biggr].
\end{aligned} \tag{33}$$

Throughout the proof, we choose some $m \to \infty$ slowly such that

$$m = o(\bar{n}_{\min} I) \quad \text{and} \quad m = o\bigl(wnI/\bigl[k[n/\bar{n}_{\min}]^2\bigr]^{1/4}\bigr). \tag{34}$$

The key to the rest of the analysis is to understand equation (33) through the decomposition of the critical quantity $\sum_{j\neq i}(\pi_{j,a} - \pi_{j,b})(A_{i,j} - \lambda)$ in the way as we describe in Section 6.3.1. We will show for any pair of $a, b \in [k]$ such that $a \neq b$, and any $i \in [n]$ such that $z_i = b$, it is equal to a summation of two terms: one only involves the ground truth Z^*, and the other involves the deviation $\pi - Z^*$. The former remains steady along iterations and contributes to the minimax rate, while the latter is related to $\|\pi - Z^*\|_1$.

Let $\theta_{a,b}$ be a vector of length n such that

$$[\theta_{a,b}]_j = \pi_{j,a} - Z^*_{j,a} + Z^*_{j,b} - \pi_{j,b} \quad \forall j \in [n]. \tag{35}$$

Then we have

$$\begin{aligned}
\sum_{j\neq i} (\pi_{j,a} - \pi_{j,b})(A_{i,j} - \lambda) &= \sum_{j\neq i} (Z^*_{j,a} - Z^*_{j,b})(A_{i,j} - \lambda) \\
&\quad + \sum_{j\neq i} \bigl(\pi_{j,a} - Z^*_{j,a} + Z^*_{j,b} - \pi_{j,b}\bigr)(A_{i,j} - \lambda) \\
&= \sum_{j\neq i} (Z^*_{j,a} - Z^*_{j,b})(A_{i,j} - \lambda) + \sum_{j\neq i} (A_{i,j} - \lambda)[\theta_{a,b}]_j
\end{aligned}$$

$$
\begin{aligned}
&= \underbrace{\sum_{j\neq i}(Z^*_{j,a}-Z^*_{j,b})(A_{i,j}-\lambda)}_{\text{involves } Z^*} + \underbrace{(A_{i,\cdot}-\mathbb{E}A_{i,\cdot})\theta_{a,b}}_{\text{involves } \theta_{a,b}\text{, that is, } \pi-Z^*} \\
&\quad + \underbrace{\sum_{j\neq i}(\mathbb{E}A_{i,j}-\lambda)[\theta_{a,b}]_j}_{\text{relatively small}} \\
&:= S^{(1)}_{i,a,b}+S^{(2)}_{i,a,b}+S^{(3)}_{i,a,b}.
\end{aligned}
$$

With the help of the above decomposition, the indicator function in equation (33) can be decomposed accordingly:

$$
\begin{aligned}
&\mathbb{I}\left[\sum_{j\neq i}(\pi_{j,a}-\pi_{j,b})(A_{i,j}-\lambda)\geq -\frac{(l+1)(n_a+n_b)I}{4mt}\right] \\
&\quad = \mathbb{I}\left[S^{(1)}_{i,a,b}+S^{(2)}_{i,a,b}+S^{(3)}_{i,a,b}\geq -\frac{(l+3/2)(n_a+n_b)I}{4mt}+\frac{(n_a+n_b)I/2}{4mt}\right] \\
&\quad \leq \mathbb{I}\left[S^{(1)}_{i,a,b}\geq -\frac{(l+3/2)(n_a+n_b)I}{4mt}-S^{(3)}_{i,a,b}\right]+\mathbb{I}\left[S^{(2)}_{i,a,b}\geq \frac{(n_a+n_b)I/2}{4mt}\right] \\
&\quad \leq \mathbb{I}\left[S^{(1)}_{i,a,b}\geq -\frac{(l+3/2)(n_a+n_b)I}{4mt}-S^{(3)}_{i,a,b}\right]+\mathbb{I}\left[S^{(2)}_{i,a,b}\geq \frac{\bar{n}_{\min}I}{4mt}\right],
\end{aligned}
$$

where in the last equality we use the fact again that $\min_{a\neq b}(n_a+b_b)/2\geq \bar{n}_{\min}$. As a result, equation (33) can be written as

$$
\begin{aligned}
\|h_{t,\lambda}(\pi)-Z^*\|_1 &\leq 2wnk\exp(-\bar{n}_{\min}I) \\
&\quad + 2w\sum_{l=0}^{m-1}\sum_{a=1}^{k}\sum_{b\neq a}\left[\exp\left[-\frac{l(n_a+n_b)I}{2m}\right]\right. \\
&\quad \times \left.\sum_{i:z_i=b}\mathbb{I}\left[S^{(1)}_{i,a,b}\geq -\frac{(l+3/2)(n_a+n_b)I}{4mt}-S^{(3)}_{i,a,b}\right]\right] \\
&\quad + 2w\sum_{a=1}^{k}\sum_{b\neq a}\left[\left[\sum_{l=0}^{m-1}\exp\left[-\frac{l(n_a+n_b)I}{2m}\right]\right]\times\sum_{i:z_i=b}\mathbb{I}\left[S^{(2)}_{i,a,b}\geq\frac{\bar{n}_{\min}I}{4mt}\right]\right].
\end{aligned}
$$

We further simplify equation (33) by using two quantities L^{sum}_1, L^{sum}_2 defined as follows:

$$
L^{\text{sum}}_1 \triangleq \sum_{l=0}^{m-1}\sum_{a=1}^{k}\sum_{b\neq a}\left[\exp\left[-\frac{l(n_a+n_b)I}{2m}\right]\sum_{i:z_i=b}\mathbb{I}\left[S^{(1)}_{i,a,b}\geq -\frac{(l+3/2)(n_a+n_b)I}{4mt}-S^{(3)}_{i,a,b}\right]\right]
$$

and

$$
L^{\text{sum}}_2 \triangleq \sum_{a=1}^{k}\sum_{b\neq a}\sum_{i:z_i=b}\mathbb{I}\left[S^{(2)}_{i,a,b}\geq\frac{\bar{n}_{\min}I}{4mt}\right].
$$

By equations (19) and (34), the exponent in $\exp[-l(n_a+n_b)I/(2m)]$ goes to infinity for all $l\geq 1$, which implies $\sum_{l=0}^{m-1}\exp[-l(n_a+n_b)I/(2m)]\leq 2$. Thus, we have

$$
\|h_{t,\lambda}(\pi)-Z^*\|_1\leq 2wnk\exp(-\bar{n}_{\min}I)+\underbrace{2wL^{\text{sum}}_1}_{\text{involves } Z^*}+\underbrace{4wL^{\text{sum}}_2}_{\text{involves } \pi-Z^*}.
$$

The two quantities L_1^{sum} and L_2^{sum} correspond to the two sources of errors we describe in Section 6.3.1. The former one leads to the optimal rate, while the latter one is related to the deviation between π and Z^*. Upper bounds on L_1^{sum} and L_2^{sum} are established in Section B.1, summarized as follows:

- For L_1^{sum}, there exists a sequence $\eta'' = o(1)$ such that with probability at least $1 - \exp[-2(\bar{n}_{\min} I)^{\frac{1}{2}}]$, we have

$$L_1^{\text{sum}} \leq nmk \exp\bigl[-(1-2\eta'')\bar{n}_{\min} I\bigr]. \tag{36}$$

- For L_2^{sum}, there exist constants c and r such that with probability at least $1 - n^{-r} - \exp(-5np^*)$, we have

$$L_2^{\text{sum}} \leq \frac{cknp^* \|\pi - Z^*\|_1}{(\bar{n}_{\min} I/(mt^*))^2} + \frac{cn^2kp^* \exp(-5np^*)}{\bar{n}_{\min} I/(mt^*)}. \tag{37}$$

Plugging upper bounds on L_1^{sum} and L_2^{sum}, we have

$$\begin{aligned}\|h_{t,\lambda}(\pi) - Z^*\|_1 \leq{}& 2wnk \exp(-\bar{n}_{\min} I) + 2wnmk \exp\bigl[-(1-2\eta'')\bar{n}_{\min} I\bigr] \\ &+ \frac{4cwknp^* \|\pi - Z^*\|_1}{(\bar{n}_{\min} I/(mt^*))^2} + \frac{4cwkn^2 p^* \exp(-5np^*)}{\bar{n}_{\min} I/(mt^*)},\end{aligned}$$

with probability at least $1 - \exp[-2(\bar{n}_{\min} I)^{\frac{1}{2}}] - n^{-r} - \exp(-5np^*)$.

The last thing to do to complete the proof is to simplify the above with-high-probability result. By Propositions C.2 and C.3, we have $p^* t^{*2} \asymp I$, which leads to

$$\frac{wknp^*}{(\bar{n}_{\min} I/(mt^*))^2} = wm^2 \frac{p^* t^{*2}}{I} \frac{n^2}{\bar{n}_{\min}^2} \frac{k}{nI} \asymp wm^2 \frac{n^2}{\bar{n}_{\min}^2} \frac{k}{nI} = o\left[\frac{1}{\sqrt{nI/[wk[n/\bar{n}_{\min}]^2]}}\right],$$

where the last inequality is due to the relative value of m defined in equation (34). Similarly, we have

$$\begin{aligned}\frac{wkn^2 p^* \exp(-5np^*)}{\bar{n}_{\min} I/(mt^*)} &= wmk \sqrt{\frac{p^* t^{*2}}{I}} \frac{\sqrt{np^*}}{\sqrt{nI}} \frac{n}{\bar{n}_{\min}} n \exp(-5np^*) \\ &\asymp wmk \frac{\sqrt{np^*}}{\sqrt{nI}} \frac{n}{\bar{n}_{\min}} n \exp(-5np^*) \\ &\leq n \exp(-5\bar{n}_{\min} I).\end{aligned}$$

Thus, with probability at least $1 - \exp[-(\bar{n}_{\min} I)^{\frac{1}{2}}] - n^{-r}$, there exists some $\eta = o(1)$, such that

$$\|h_{t,\lambda}(\pi) - Z^*\|_1 \leq n \exp\bigl(-(1-\eta)\bar{n}_{\min} I\bigr) + \frac{\|\pi - Z^*\|_1}{\sqrt{nI/[wk[n/\bar{n}_{\min}]^2]}}.$$

This completes the proof of Theorem 6.1. Due to the limit of space, we refer readers to Section B.1 in the Supplementary Material for the establishment of upper bounds on L_1^{sum} and L_2^{sum}.

SUPPLEMENTARY MATERIAL

Supplement to "Theoretical and computational guarantees of mean field variational inference for community detection" (DOI: 10.1214/19-AOS1898SUPP; .pdf). In the Supplementary Material [35], we provide the detailed implementations of the batched Gibbs sampling and the iterative algorithm for MLE in Algorithm 2 and Algorithm 3, respectively. We include proofs of Theorem 4.1, Theorem 4.2 and Theorem 4.3. Besides, we establish the upper bounds on L_1^{sum} and L_2^{sum} which are used in the proof of Theorem 6.1, and study the case where $\ell(\pi^{(0)}, \pi^*)$ is in a constant order of $\bar{n}_{\min}$ to complement the proof of Theorem 3.1. In addition, all the auxiliary propositions and lemmas are also included in the supplement.

REFERENCES

[1] AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

[2] BEAL, M. J. (2003). *Variational Algorithms for Approximate Bayesian Inference*. Univ. London, London.

[3] BICKEL, P., CHOI, D., CHANG, X. and ZHANG, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Statist.* **41** 1922–1943. MR3127853 https://doi.org/10.1214/13-AOS1124

[4] BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.

[5] BICKEL, P. J., CHEN, A. and LEVINA, E. (2011). The method of moments and degree distributions for network models. *Ann. Statist.* **39** 2280–2301. MR2906868 https://doi.org/10.1214/11-AOS904

[6] BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning*. *Information Science and Statistics*. Springer, New York. MR2247587 https://doi.org/10.1007/978-0-387-45528-0

[7] BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. MR3671776 https://doi.org/10.1080/01621459.2017.1285773

[8] BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.

[9] CELISSE, A., DAUDIN, J.-J. and PIERRE, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Stat.* **6** 1847–1899. MR2988467 https://doi.org/10.1214/12-EJS729

[10] CHIN, P., RAO, A. and VU, V. (2015). Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *COLT* 391–423.

[11] FEI, Y. and CHEN, Y. (2019). Exponential error rates of SDP for block models: Beyond Grothendieck's inequality. *IEEE Trans. Inform. Theory* **65** 551–571. MR3901009 https://doi.org/10.1109/TIT.2018.2839677

[12] GAO, C., MA, Z., ZHANG, A. Y. and ZHOU, H. H. (2017). Achieving optimal misclassification proportion in stochastic block models. *J. Mach. Learn. Res.* **18** Art. ID 60. MR3687603

[13] GAO, C., VAN DER VAART, A. W. and ZHOU, H. H. (2015). A general framework for Bayes structured linear models. Preprint. Available at arXiv:1506.02174.

[14] GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409. MR1141740

[15] GRABSKA-BARWIŃSKA, A., BARTHELMÉ, S., BECK, J., MAINEN, Z. F., POUGET, A. and LATHAM, P. E. (2017). A probabilistic approach to demixing odors. *Nat. Neurosci.* **20** 98–106. https://doi.org/10.1038/nn.4444

[16] GUÉDON, O. and VERSHYNIN, R. (2016). Community detection in sparse networks via Grothendieck's inequality. *Probab. Theory Related Fields* **165** 1025–1049. MR3520025 https://doi.org/10.1007/s00440-015-0659-z

[17] HOFMAN, J. M. and WIGGINS, C. H. (2008). Bayesian approach to network modularity. *Phys. Rev. Lett.* **100** Art. ID 258701. https://doi.org/10.1103/PhysRevLett.100.258701

[18] HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088 https://doi.org/10.1016/0378-8733(83)90021-7

[19] JORDAN, M. I., GHAHRAMANI, Z., JAAKKOLA, T. S. and SAUL, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.* **37** 183–233.

[20] LEI, J. and RINALDO, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43** 215–237. MR3285605 https://doi.org/10.1214/14-AOS1274

[21] LIANG, P., PETROV, S., JORDAN, M. I. and KLEIN, D. (2007). The infinite PCFG using hierarchical Dirichlet processes. In *EMNLP-CoNLL* 688–697.

[22] LU, Y. and ZHOU, H. H. (2016). Statistical and computational guarantees of Lloyd's algorithm and its variants. Preprint. Available at arXiv:1612.02099.

[23] MOSSEL, E., NEEMAN, J. and SLY, A. (2012). Stochastic block models and reconstruction. Preprint. Available at arXiv:1202.1499.

[24] NEWMAN, M. E. J. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **103** 8577–8582. https://doi.org/10.1073/pnas.0601602103

[25] PENNY, W. D., TRUJILLO-BARRETO, N. J. and FRISTON, K. J. (2005). Bayesian fMRI time series analysis with spatial priors. *NeuroImage* **24** 350–362.

[26] RAZAEE, Z. S., AMINI, A. A. and LI, J. J. (2019). Matched bipartite block model with covariates. *J. Mach. Learn. Res.* **20** Art. ID 34. MR3948074

[27] ROBERT, C. P. (2004). *Monte Carlo Methods*. Wiley, New York.

[28] ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856 https://doi.org/10.1214/11-AOS887

[29] WAINWRIGHT, M. J. and JORDAN, M. I. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1** 1–305.

[30] WANG, B. and TITTERINGTON, D. M. (2006). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Anal.* **1** 625–649. MR2221291 https://doi.org/10.1214/06-BA121

[31] WANG, Y. and BLEI, D. M. (2019). Frequentist consistency of variational Bayes. *J. Amer. Statist. Assoc.* **114** 1147–1161. MR4011769 https://doi.org/10.1080/01621459.2018.1473776

[32] WESTLING, T. and MCCORMICK, T. H. (2015). Establishing consistency and improving uncertainty estimates of variational inference through M-estimation. Preprint. Available at arXiv:1510.08151.

[33] YOU, C., ORMEROD, J. T. and MÜLLER, S. (2014). On variational Bayes estimation and variational information criteria for linear regression models. *Aust. N. Z. J. Stat.* **56** 73–87. MR3200293 https://doi.org/10.1111/anzs.12063

[34] ZHANG, A. Y. and ZHOU, H. H. (2016). Minimax rates of community detection in stochastic block models. *Ann. Statist.* **44** 2252–2280. MR3546450 https://doi.org/10.1214/15-AOS1428

[35] ZHANG, A. Y. and ZHOU, H. H. (2020). Supplement to “Theoretical and computational guarantees of mean field variational inference for community detection”. https://doi.org/10.1214/19-AOS1898SUPP