A Theoretical Case Study of Structured Variational Inference for Community Detection

Mingzhang Yin
University of Texas at Austin

Y. X. Rachel Wang
University of Sydney

Purnamrita Sarkar
University of Texas at Austin

Abstract

Mean-field variational inference (MFVI) has been widely applied in large scale Bayesian inference. However, MFVI assumes independent distribution on the latent variables, which often leads to objective functions with many local optima, making optimization algorithms sensitive to initialization. In this paper, we study the advantage of structured variational inference in the context of a simple two-class Stochastic Blockmodel. To facilitate theoretical analysis, the variational distribution is constructed to have a simple pairwise dependency structure on the nodes of the network. We prove that, in a broad density regime and for general random initializations, unlike MFVI, the estimated class labels by structured VI converge to the ground truth with high probability, when the model parameters are known, estimated within a reasonable range or jointly optimized with the variational parameters. In addition, empirically we demonstrate structured VI is more robust compared with MFVI when the graph is sparse and the signal to noise ratio is low. The paper takes a first step towards quantifying the role of added dependency structure in variational inference for community detection.

1 Introduction

Variational inference (VI) is a widely used technique for approximating complex likelihood functions in Bayesian learning [1, 2, 3], and is known for its computational scalability. VI reduces an intractable posterior inference problem to an optimization framework by imposing simpler dependence structure and is considered a popular alternative to Markov chain Monte Carlo (MCMC) methods. Similar to the Expectation Maximization (EM) algorithm [4], VI works by the basic principle of constructing a tractable lower bound on the complete log-likelihood of a probabilistic model. One of the simplest forms of approximation is mean-field variational inference (MFVI), where the variational lower bound, also known as ELBO, is computed using the expectation with respect to a product distribution over the latent variables[2, 5, 6]. Though VI has achieved great empirical success in probabilistic models and deep learning, theoretical understanding of its convergence properties is still an open area of research.

Theoretical studies of variational methods (and similar algorithms that involve iteratively maximizing a lower bound) have drawn significant attention recently (see [7, 8, 9, 10, 11] for convergence properties of EM). For VI, the global optimizer of the variational lower bound is shown to be asymptotically consistent for a number of models including Latent Dirichlet Allocation (LDA) [2] and Gaussian mixture models [12]. In [13] the connection between VI estimates and profile M-estimation is explored and asymptotic consistency is established. In practice, however, it is well known the algorithm is not guaranteed to reach the global optimum and the performance of VI often suffers from local optima [14]. While in some models, convergence to the global optimum can be achieved with appropriate initializations [15, 16], understanding convergence with general initializations and the influence of local optima is less studied with a few exceptions [8, 17, 18].

In general, despite being computationally scalable, MFVI suffers from many stability issues including symmetry-breaking, multiple local optima, and sensitivity to initialization, which are consequences of the non-convexity of typical mean-field problems [19, 20]. The independence assumption on latent variables also leads to the underestimation of posterior uncertainty [14]. To address these problems, many studies suggest that modeling the latent dependency structure can expand the variational family under considera-

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

tion and lead to larger ELBO and more stable convergence [21, 22, 23, 24, 25, 26, 27, 28]. However, rigorous theoretical analysis with convergence guarantees in this setting remains largely underexplored.

In this paper, we aim to study the effect of added dependency structure in a MFVI framework. Since the behavior of the log-likelihood of MFVI is well understood for the very simple two class, equal sized Stochastic Blockmodel (SBM) [18, 29], we propose to add a simple pairwise link structure to MFVI in the context of inference for SBMs. We want to point out that our goal is not to come up with a new algorithm which has better theoretical guarantees or better empirical performance over the state of the art. In this paper, we study how added dependency structure can improve MFVI, which iteratively optimizes a notoriously nonconvex loss function. In particular, we focus on how random initializations behave for MFVI with added structure.

The stochastic blockmodel (SBM) [30] is a widely used network model for community detection in networks. There are a plethora of algorithms with theoretical guarantees for estimation for SBMs like Spectral methods [31, 32], semidefinite relaxation based methods [33, 34, 35], likelihood-based methods [36], modularity based methods [37, 38, 39]. Among these, likelihood-based methods remain important and relevant due to their flexibility in incorporating additional model structures. Examples include mixed membership SBM [40], networks with node covariates [41], and dynamic networks [42]. Among likelihood based methods, VI provides a tractable approximation to the log-likelihood and is a scalable alternative to more expensive methods like Profile Likelihood [39], or MCMC based methods [37, 38]. Computationally, VI was also shown to scale up well to very large graphs [43].

On the theoretical front, [44] proved that the global optimum of MFVI behaves optimally in the dense degree regime. In terms of algorithm convergence, [29] showed the batch coordinate ascent algorithm (BCAVI) for optimizing the mean-field objective has guaranteed convergence if the initialization is sufficiently close to the ground truth. [18] fully characterized the optimization landscape and convergence regions of BCAVI for a simple two-class SBM with random initializations. It is shown that uninformative initializations can indeed converge to suboptimal local optima, demonstrating the limitations of the MFVI objective function.

Coming back to structured variational inference, it is important to note that, if one added pairwise dependences between the posterior of each node, the natural approximate inference method is the belief propagation (BP) algorithm. Based on empirical evidence, it has been conjectured in [45] that BP is asymptotically optimal for a simple two-class SBM. In the sparse setting where phase transition occurs, [46] analyzed a local variant of BP and showed it is optimal given a specific initialization. Analysis for general multiple clusters and a linearized acyclic BP has been performed in [47]. In other parameter regions, rigorous theoretical understanding of BP, in particular, how adding dependence structure can improve convergence with general initializations is still an open problem.

Motivated by the above observations, in this paper, we present a theoretical case study of structured variational inference for SBM. We emphasize here that our primary contribution *does not* lie in proposing a new estimation algorithm that outperforms state-of-the-art methods; rather we use this algorithm as an example to understand the interplay between a non-convex objective function and an iterative optimization algorithm with respect to random initializations, and compare it with MFVI. We consider a two-class SBM with equal class size, an assumption commonly used in theoretical work [46, 18] where the analysis for the simplest case is nontrivial.

We study structured VI by introducing a simple pairwise dependence structure between randomly paired nodes. By carefully bounding the mean field parameters and their logits in each iteration using a combination of concentration and Littlewood-Offord type anti-concentration arguments [48], we prove that in a broad density regime and under a fairly general random initialization scheme, the Variational Inference algorithm with Pairwise Structure (VIPS) can converge to the ground truth with probability tending to one, when the parameters are known, estimated within a reasonable range, or updated appropriately (Section 3). This is in contrast to MFVI, where convergence only happens for a narrower range of initializations. In addition, VIPS can escape from certain local optima that exist in the MFVI objective. These results highlight the theoretical advantage of the added dependence structure. Empirically, we demonstrate that VIPS is more robust compared to MFVI when the graph is sparse and the signal to noise ratio is low (Section 4). We observe similar trends hold in more general models with unbalanced class sizes and more than two classes. We hope that our analysis for the simple blockmodel setting can shed light on theoretical analysis of algorithms with more general dependence structure such as BP.

The paper is organized as follows. Section 2 contains the model definition and introduces VIPS. We present our theoretical results in Section 3. Finally in Section 4, we demonstrate the empirical performance of VIPS in contrast to MFVI and other algorithms. We conclude with a discussion on possible generalizations, accompa-

nied by promising empirical results in Section 5.

2 Preliminaries and Proposed Work

2.1 Preliminaries

The stochastic block model (SBM) is a generative network model with community structure. A K-community SBM for n nodes is generated as follows: each node is assigned to one of the communities in $\{1,\dots,K\}$ according to a Multinomial distribution with parameter π. These memberships are represented by $U \in \{0,1\}^{n\times K}$, where each row follows an independent Multinomial $(1;\pi)$ distribution. We have $U_{ik}=1$ if node i belongs to community k and $\sum_{k=1}^{K} U_{ik}=1$. Given the community memberships, links between pairs of nodes are generated according to the entries in a $K\times K$ connectivity matrix B. That is, if A denotes the $n\times n$ binary symmetric adjacency matrix, then, for $i\neq j$,

$$P(A_{ij}=1|U_{ik}=1,U_{j\ell}=1)=B_{k\ell}. \tag{1}$$

We consider undirected networks, where both B and A are symmetric. Given an observed A, the goal is to infer the latent community labels U and the model parameters (π,B). Since the data likelihood $P(A;B,\pi)$ requires summing over K^n possible labels, approximations such as MFVI are often needed to produce computationally tractable algorithms.

Throughout the rest of the paper, we will use $\mathbf{1}_n$ to denote the all-one vector of length n. When it is clear from the context, we will drop the subscript n. Let I be the identity matrix and $J=\mathbf{1}\mathbf{1}^T$. $\mathbb{1}_C$ denotes a vector where the i-th element is 1 if $i\in C$ and 0 otherwise, where C is some index set. Similar to [18], we consider a two-class SBM with equal class size, where $K=2$, $\pi=1/2$, and B takes the form $B_{11}=B_{22}=p$, $B_{12}=B_{21}=q$, with $p>q$. We denote the two true underlying communities by G_1 and G_2, where G_1,G_2 form a partition of $\{1,2,\dots,n\}$ and $|G_1|=|G_2|$. (For convenience, we assume n is even.) As will become clear, the full analysis of structured VI in this simple case is highly nontrivial.

2.2 Variational inference with pairwise structure (VIPS)

The well-known MFVI approximates the likelihood by assuming a product distribution over the latent variables. In other words, the posterior label distribution of the nodes is assumed to be independent in the variational distribution. To investigate how introducing dependence structure can help with the inference, we focus on a simple setting of linked pairs which are independent of each other. To be concrete, we randomly partition the n nodes into two sets: $P_1=\{z_1,\cdots,z_m\}$, $P_2=\{y_1,\cdots,y_m\}$, with $m=n/2$. Here $z_k,y_k\in\{1,\dots,n\}$ are the node indices. In our structured variational distribution, we label pairs of nodes (z_k,y_k) using index $k\in\{1,\dots,m\}$ and assume there is dependence within each pair. The corresponding membership matrices for P_1 and P_2 are denoted by Z and Y respectively, which are both $m\times 2$ sub-matrices of the full membership matrix U. More explicitly, the k^{th} row of matrix Z encodes the membership of node z_k in P_1, and similarly for Y. For convenience, we permute both the rows and columns of A based on the node ordering in P_1 followed by that in P_2 to create a partitioned matrix: $A=\left[\begin{array}{c|c} A^{zz} & A^{zy} \\ \hline A^{yz} & A^{yy}\end{array}\right]$, where each block is an $m\times m$ matrix. Given the latent membership variable (Z,Y), by Eq. (1) the likelihood of A is given by

$$\begin{aligned}
P(A^{zz}_{ij}|Z,B)&=\textstyle\prod_{a,b}[B_{ab}^{A^{zz}_{ij}}(1-B_{ab})^{1-A^{zz}_{ij}}]^{Z_{ia}Z_{jb}}\\
P(A^{zy}_{ij}|Y,Z,B)&=\textstyle\prod_{a,b}[B_{ab}^{A^{zy}_{ij}}(1-B_{ab})^{1-A^{zy}_{ij}}]^{Z_{ia}Y_{jb}}\\
P(A^{yy}_{ij}|Y,B)&=\textstyle\prod_{a,b}[B_{ab}^{A^{yy}_{ij}}(1-B_{ab})^{1-A^{yy}_{ij}}]^{Y_{ia}Y_{jb}}
\end{aligned} \tag{2}$$

where $a,b\in\{1,2\}$ and $A^{zy}=(A^{yz})^T$.

A simple illustration of the partition and how ordered pairs of nodes are linked to incorporate dependence is given in Figure 1, where the the true underlying communities G_1 and G_2 are shaded differently. After the partition, we have m pairs of linked nodes indexed from 1 to m. For convenience of analysis, we define the following sets for these pairs of linked nodes, as illustrated in Figure 1.

Define C_1, (C_1') as the set of indices i of pairs (z_i,y_i) with $z_i\in G_1$, $(y_i\in G_1)$. Similarly, C_2, (C_2') is the set of indices of pairs (z_i,y_i) with $z_i\in G_2$, $(y_i\in G_2)$. In Figure 1, G_1 is the set of shaded nodes. We will also make use of the sets $C_{ab}\coloneqq C_a\cap C_b'$, where $a,b\in\{1,2\}$. In Figure 1, these sets correspond to different combinations of shading, i.e. community memberships, of the linked pairs, e.g. C_{12} is the index set of pairs (z_i,y_i) with $z_i\in G_1, y_i\in G_2$.

We define the variational distribution for the latent membership matrix (Z,Y) as $Q(Z,Y)$, which we assume takes the form

$$Q(Z,Y)=\prod_{i=1}^{m}Q(Z_i,Y_i), \tag{3}$$

where Z_i denotes the i^{th} row of Z, and $Q(Z_i,Y_i)$ is a general categorical distribution with variational parameters defined as follows.

$$\psi_i^{cd}\coloneqq Q(Z_{i,c+1}=1,Y_{i,d+1}=1),$$

for $i \in \{1, \ldots, m\}, c, d \in \{0, 1\}$. This allows one to encode more dependence structure between the posteriors at different nodes than vanilla MFVI, since we allow for dependence within each linked pair of nodes while keeping independence between different pairs. We define the marginal probabilities as:

$$\begin{aligned} \phi_i &\coloneqq Q(Z_{i1} = 1) = \psi_i^{10} + \psi_i^{11} \\ \xi_i &\coloneqq Q(Y_{i1} = 1) = \psi_i^{01} + \psi_i^{11}. \end{aligned} \tag{4}$$

Next we derive the variational lower bound (also known

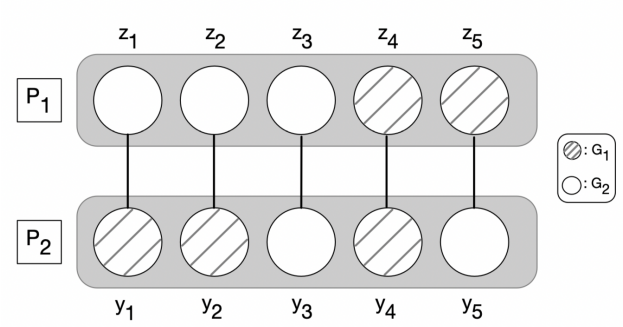

Figure 1: An illustration of a partition for $n = 10$. The shaded nodes belong to community G_1 and unshaded nodes belong to community G_2. The nodes are randomly partitioned into two sets P_1 and P_2, and pairs of nodes are linked from index 1 to m. Dependence structure within each linked pair is incorporated into the variational distribution $Q(Z, Y)$. For this partition and pair linking, $C_1 = \{4, 5\}$, $C_2 = \{1, 2, 3\}$, $C_1' = \{1, 2, 4\}$, $C_2' = \{3, 5\}$; $C_{11} = \{4\}$, $C_{12} = \{5\}$, $C_{21} = \{1, 2\}$, $C_{22} = \{3\}$.

as the evidence lower bound (ELBO)) on the data log-likelihood $\log P(A)$ using $Q(Z, Y)$. For pairwise structured variational inference (VIPS), ELBO takes the form

$$\begin{aligned} \mathcal{L}(Q; \pi, B) =& \mathbb{E}_{Z,Y \sim Q(Z,Y)} \log P(A|Z, Y) \\ & - \mathrm{KL}(Q(Z, Y)||P(Z, Y)), \end{aligned}$$

where $P(Z, Y)$ is the probability of community labels from SBM and follows independent Bernoulli (π) distribution, $\mathrm{KL}(\cdot||\cdot)$ denotes the usual Kullback–Leibler divergence between two distributions. Using the likelihood in Eq. (2), the ELBO becomes

$$\begin{aligned} \mathcal{L}(Q; \pi, B) =& \frac{1}{2}\mathbb{E}_Q \sum_{i \neq j, a, b} Z_{ia} Z_{jb} (A_{ij}^{zz} \alpha_{ab} + f(\alpha_{ab})) \\ & + \frac{1}{2}\mathbb{E}_Q \sum_{i \neq j, a, b} Y_{ia} Y_{jb} (A_{ij}^{yy} \alpha_{ab} + f(\alpha_{ab})) \\ & + \mathbb{E}_Q \sum_{i \neq j, a, b} Z_{ia} Y_{jb} (A_{ij}^{zy} \alpha_{ab} + f(\alpha_{ab})) \\ & + \mathbb{E}_Q \sum_{i, a, b} Z_{ia} Y_{ib} (A_{ii}^{zy} \alpha_{ab} + f(\alpha_{ab})) \\ & - \sum_{i=1}^{m} \mathrm{KL}(Q(z_i, y_i)||P(z_i)P(y_i)), \end{aligned} \tag{5}$$

where $\alpha_{ab} = \log(B_{ab}/(1 - B_{ab}))$ and $f(\alpha) = -\log(1 + e^{\alpha})$. The KL regularization term can be computed as

$$\begin{aligned} & \mathrm{KL}(Q(z_i, y_i)||P(z_i)P(y_i)) \\ =& \psi_i^{00} \log \tfrac{\psi_i^{00}}{(1-\pi)^2} + \psi_i^{01} \log \tfrac{\psi_i^{01}}{\pi(1-\pi)} \\ & + \psi_i^{10} \log \frac{\psi_i^{10}}{\pi(1-\pi)} + \psi_i^{11} \log \frac{\psi_i^{11}}{\pi^2} \\ =& \textstyle\sum_{0 \leq c, d \leq 1} \psi_i^{cd} \log(\psi_i^{cd})/(\pi^c \pi^d (1-\pi)^{1-c} (1-\pi)^{1-d}). \end{aligned} \tag{6}$$

Our goal is to maximize $\mathcal{L}(Q; \pi, B)$ with respect to the variational parameters ψ_i^{cd} for $1 \leq i \leq m$. Since $\sum_{c,d} \psi_i^{cd} = 1$ for each i, it suffices to consider ψ_i^{10}, ψ_i^{01} and ψ_i^{11}. By taking derivatives, we can derive a batch coordinate ascent algorithm for updating $\psi^{cd} = (\psi_1^{cd}, \ldots, \psi_m^{cd})$. Detailed calculation of the derivatives can be found in Section A of the Appendix. Recall that $\pi = \frac{1}{2}$. Also, define

$$t \coloneqq \frac{1}{2} \log \frac{p/(1-p)}{q/(1-q)} \qquad \lambda \coloneqq \frac{1}{2t} \log \frac{1-q}{1-p}, \tag{7}$$

$$\theta^{cd} \coloneqq \log \frac{\psi^{cd}}{1 - \psi^{01} - \psi^{10} - \psi^{11}}, \tag{8}$$

where θ^{cd} are logits, $c, d \in \{0, 1\}$ and all the operations are defined *element-wise.*

Given the model parameters p, q, the current values of ψ^{cd} and the marginals $\phi = \psi^{10} + \psi^{11}$, $\xi = \psi^{01} + \psi^{11}$ as defined in Eq. (4), the updates for θ^{cd} are given by:

$$\begin{aligned} \theta^{10} =& 4t[A^{zz} - \lambda(J - I)](\phi - \tfrac{1}{2}\mathbf{1}_m) \\ & + 4t[A^{zy} - \lambda(J - I) - \mathrm{diag}(A^{zy})](\xi - \frac{1}{2}\mathbf{1}_m) \\ & - 2t(\mathrm{diag}(A^{zy}) - \lambda I)\mathbf{1}_m, \end{aligned} \tag{9}$$

$$\begin{aligned} \theta^{01} =& 4t[A^{yy} - \lambda(J - I)](\xi - \tfrac{1}{2}\mathbf{1}_m) \\ & + 4t[A^{yz} - \lambda(J - I) - \mathrm{diag}(A^{yz})](\phi - \frac{1}{2}\mathbf{1}_m) \\ & - 2t(\mathrm{diag}(A^{yz}) - \lambda I)\mathbf{1}_m, \end{aligned} \tag{10}$$

$$\begin{aligned} \theta^{11} =& 4t[A^{zz} - \lambda(J - I)](\phi - \tfrac{1}{2}\mathbf{1}_m) \\ & + 4t[A^{zy} - \lambda(J - I) - \mathrm{diag}(A^{zy})](\xi - \frac{1}{2}\mathbf{1}_m) \\ & + 4t[A^{yy} - \lambda(J - I)](\xi - \tfrac{1}{2}\mathbf{1}_m) \\ & + 4t[A^{yz} - \lambda(J - I) - \mathrm{diag}(A^{yz})](\phi - \frac{1}{2}\mathbf{1}_m). \end{aligned} \tag{11}$$

Given θ^{cd}, we can update the current values of ψ^{cd} and the corresponding marginal probabilities ϕ, ξ using element-wise operations as follows:

$$\psi^{cd} = \frac{e^{\theta^{cd}}}{1+e^{\theta^{01}}+e^{\theta^{11}}+e^{\theta^{10}}}, \quad u := (\phi, \xi)$$
$$\phi = \frac{e^{\theta^{10}}+e^{\theta^{11}}}{1+e^{\theta^{10}}+e^{\theta^{01}}+e^{\theta^{11}}}, \quad \xi = \frac{e^{\theta^{01}}+e^{\theta^{11}}}{1+e^{\theta^{10}}+e^{\theta^{01}}+e^{\theta^{11}}}, \tag{12}$$

where $(c,d) = (1,0),(0,1),(1,1)$. The marginal probabilities are concatenated as $u = (\phi, \xi) \in [0,1]^n$. Thus u can be interpreted as the estimated posterior membership probability of all the nodes.

Since θ^{cd} determines ψ^{cd} in the categorical distribution and u represents the corresponding marginals, one can think of θ^{cd} and u as the local and global parameters respectively. It has been empirically shown that the structured variational methods can achieve better convergence property by iteratively updating the local and global parameters [2, 6, 22]. In the same spirit, in the full optimization algorithm, we update the parameters θ^{cd} and u iteratively by (9)–(12), following the order

$$\theta^{10} \to u \to \theta^{01} \to u \to \theta^{11} \to u \to \theta^{10} \cdots . \tag{13}$$

We call a full update of all the parameters $\theta^{10}, \theta^{01}, \theta^{11}, u$ in (13) as one *meta iteration* which consists of three inner iterations of u updates. We use $u_i^{(k)}$ $(i = 1,2,3)$ to denote the update in the i-th iteration of the k-th meta iteration, and u_0 to denote the initialization. Algorithm 1 gives the full algorithm when the model parameters are known.

Algorithm 1 Variational Inference with Pairwise Structure (VIPS)

input : Adjacency matrix $A \in \{0,1\}^{n\times n}$, model parameter $p, q, \pi = 1/2$.
output : The estimated node membership vector u.

Initialize the elements of u i.i.d. from an arbitrary distribution f_μ defined on $[0,1]$ with mean μ. Initialize $\theta^{10} = \theta^{01} = \theta^{11} = \mathbf{0}$;
Randomly select $n/2$ nodes as P_1 and the other $n/2$ nodes as P_2;
while *not converged* **do**
  Update θ^{10} by (9).
  Update $u = (\phi, \xi)$ by (12)
  Update θ^{01} by (10).
  Update $u = (\phi, \xi)$ by (12)
  Update θ^{11} by (11).
  Update $u = (\phi, \xi)$ by (12)
end

Remark 1. *So far we have derived the updates and described the optimization algorithm when the true parameters p, q are known. When they are unknown, they can be updated jointly with the variational parameters after each meta iteration as*

$$p = \frac{(\mathbf{1}_n - u)^T A(\mathbf{1}_n - u) + u^T A u + 2(\mathbf{1}_m - \psi^{10} - \psi^{01})^T diag(A^{zy})\mathbf{1}_m}{(\mathbf{1}_n - u)^T(J-I)(\mathbf{1}_n - u) + u^T(J-I)u + 2(\mathbf{1}_m - \psi^{10} - \psi^{01})^T \mathbf{1}_m}$$
$$q = \frac{(\mathbf{1}_n - u)^T A u + (\psi^{10} + \psi^{01})^T diag(A^{zy})\mathbf{1}_m}{(\mathbf{1}_n - u)^T(J-I)u_n + (\psi^{10} + \psi^{01})^T \mathbf{1}_m} \tag{14}$$

Although it is typical to update p, q and u jointly, as shown in [18], analyzing MFVI updates with known parameters can shed light on the convergence behavior of the algorithm. Initializing u randomly while jointly updating p, q always leads MFVI to an uninformative local optima. For this reason, in what follows we will analyze Algorithm 1 in the context of both fixed and updating parameters p, q.

3 Main results

In this section, we present theoretical analysis of the algorithm in three settings: (i) When the parameters are set to the true model parameters p, q; (ii) When the parameters are not too far from the true values, and are held fixed throughout the updates; (iii) Starting from some reasonable guesses of the parameters, they are jointly updated with latent membership estimates.

In the following analysis, we will frequently use the eigen-decomposition of the expected adjacency matrix $P = \mathbb{E}[A|U] = \frac{p+q}{2}\mathbf{1}_n\mathbf{1}_n^T + \frac{p-q}{2}v_2 v_2^T - pI$ where $v_2 = (v_{21}, v_{22})^T = (\mathbb{1}_{C_1} - \mathbb{1}_{C_2}, \mathbb{1}_{C_1'} - \mathbb{1}_{C_2'})^T$ is the second eigenvector. Since the second eigenvector is just a shifted and scaled version of the membership vector, the projection $|\langle u, v_2\rangle|$ is equivalent to the ℓ_1 error from true label z^* (up-to label permutation) by $\|u - z^*\|_1 = m - |\langle u, v_2\rangle|$. We consider the case where $p = c\rho_n, q = d\rho_n$ where the density $\rho_n \to 0$ at some rate and constant $c > d > 0$.

When the true parameters p, q are known, it has been shown [49] that without dependency structure, MFVI with random initializations converges to the stationary points with non-negligible probability. When the variational distribution has a simple pairwise dependency structure as VIPS, we show a stronger result. To be concrete, in this setting, we establish that convergence happens with probability approaching 1. In addition, unlike MFVI, the convergence holds for general random initializations. We will first consider the situation when u_0 is initialized from a distribution centered at $\mu = \frac{1}{2}$ and show the results for $\mu \neq \frac{1}{2}$ in Corollary 1.

Theorem 1 (Sample behavior for known parameters)**.** *Assume $\theta^{10}, \theta^{01}, \theta^{11}$ are initialized as $\mathbf{0}$ and the elements of $u_0 = (\phi^{(0)}, \xi^{(0)})$ are initialized i.i.d. from*

Bernoulli($\frac{1}{2}$). *When* $p \asymp q \asymp \rho_n$, $p - q = \Omega(\rho_n)$, $\sqrt{n}\rho_n = \Omega(\log(n))$, *Algorithm 1 converges to the true labels asymptotically after the second meta iteration, in the sense that*

$$\|u_3^{(2)} - z^*\|_1 = n \exp(-\Omega_P(n\rho_n)))$$

z^* *are the true labels with* $z^* = \mathbb{1}_{G_1}$ *or* $\mathbb{1}_{G_2}$. *The same convergence holds for all the later iterations.*

Remark 2. *It is important to note that there are many algorithms (see [50] for a survey) which recover the memberships exactly in this regime. We do not compare our theoretical results with those or to well known thresholds for exact recovery [51], because our goal is not to design a new algorithm with an improved theoretical guarantees. Rather, we show that by introducing the simplest possible pairwise dependence structure, variational inference for a simple setting of a SBM improves over MFVI which has no such structure. The density regime simply makes the analysis somewhat easier.*

Proof. We provide a proof sketch here and defer the details to Section B of the Appendix. We assume for the first six iterations, we randomly partition A into six $A^{(i)}, i = 0, \ldots, 5$ by assigning each edge to one of the six subgraphs with equal probability. For the later iterations, we can use the whole graph A. Then $A^{(i)}$'s are independent with population matrix $P/6$. The graph splitting is a widely used technique for theoretical convenience [52, 53] and allows us to bound the noise in each iteration more easily. The main arguments involve lower bounding the size of the projection $|\langle u, v_2\rangle|$ in each iteration as it increases towards $n/2$, at which point the algorithm achieves strong consistency. For ease of exposition, we will scale everything by 6 so that p, q, λ correspond to the parameters for the full un-split matrix P. This does not affect the analysis in any way.

In each iteration, we decompose the intermediate $\theta^{10}, \theta^{01}, \theta^{11}$ into blockwise constant signal and random noise using the spectral property of the population matrix P. As an illustration, in the first meta iteration, we write the update in (9)–(11) as signal plus noise,

$$\begin{aligned}
\theta_i^{10} &= 4t(s_1 \mathbb{1}_{C_1} + s_2 \mathbb{1}_{C_2} + r_i^{(0)}) \\
\theta_i^{01} &= 4t(x_1 \mathbb{1}_{C_1'} + x_2 \mathbb{1}_{C_2'} + r_i^{(1)}) \\
\theta_i^{11} &= 4t(y_1 \mathbb{1}_{C_1} + y_2 \mathbb{1}_{C_2} + y_1 \mathbb{1}_{C_1'} + y_2 \mathbb{1}_{C_2'} + r_i^{(2)})
\end{aligned}$$

where t is a constant and the noise has the form

$$r^{(i)} = R^{(i)}(u_j^{(k)} - \frac{1}{2}\mathbf{1}) \tag{15}$$

for appropriate j, k, where $R^{(i)}$ arises from the sample noise in the adjacency matrix. We handle the noise from the first iteration $r^{(0)}$ with a Berry-Esseen bound conditional on u_0, and the later $r^{(i)}$ with a uniform bound. The blockwise constant signals s_1, x_1, y_1 are updated as $(\frac{p+q}{2} - \lambda)(\langle u, \mathbf{1}_n\rangle - m) + (\frac{p-q}{2})\langle u, v_2\rangle$ and s_2, x_2, y_2 are updated as $(\frac{p+q}{2} - \lambda)(\langle u, \mathbf{1}_n\rangle - m) - (\frac{p-q}{2})\langle u, v_2\rangle$. As $\langle u, v_2\rangle$ increases throughout the iterations, the signals become increasingly separated for the two communities. Using Littlewood-Offord type anti-concentration, we show in the first meta iteration,

$$\begin{aligned}
&\langle u_1, v_2\rangle = O_P(n\sqrt{\rho_n}), \quad \langle u_1, \mathbf{1}\rangle - m = 0 \\
&\langle u_2, v_2\rangle \geq \frac{n}{8} - o_P(n), \quad \langle u_2, \mathbf{1}\rangle - m = 0 \\
&\langle u_3, v_2\rangle \geq \frac{1}{4}n + o_P(n), \\
&-\frac{n}{8} - o_P(n) \leq \langle u_3, \mathbf{1}\rangle - m \leq \frac{n}{4} + o_P(n)
\end{aligned} \tag{16}$$

After the second meta iteration we have

$$\begin{aligned}
s_1^{(2)}, x_1^{(2)}, y_1^{(2)} &= \Omega_P(n\rho_n), \\
s_2^{(2)}, x_2^{(2)}, y_2^{(2)} &= -\Omega_P(n\rho_n); \\
2y_1^{(2)} - s_1^{(2)} &= \Omega_P(n\rho_n), \\
2y_1^{(2)} - x_1^{(2)} &= \Omega_P(n\rho_n); \\
s_1^{(2)} - (y_1^{(2)} + y_2^{(2)}) &= \Omega_P(n\rho_n), \\
x_1^{(2)} - (y_1^{(2)} + y_2^{(2)}) &= \Omega_P(n\rho_n);
\end{aligned} \tag{17}$$

Plugging (17) to (12), we have the desired convergence after the second meta iteration. □

The next corollary shows the same convergence holds when we use a general random initialization not centered at 1/2. In contrast, MFVI converges to stationary points $\mathbf{0}_n$ or $\mathbf{1}_n$ with such initializations.

Corollary 1. *Assume the elements of* u_0 *are i.i.d. sampled from a distribution with mean* $\mu \neq 0.5$. *When* $\sqrt{n}\rho_n = \Omega(\log(n))$, *applying Algorithm 1 with known* p, q, *we have* $\|u_1^{(3)} - z^*\|_1 = n \exp(-\Omega_P(n\rho_n)))$. *The same order holds for all the later iterations.*

The proof relies on showing after the first iteration, $u_1^{(1)}$ behaves like nearly independent Bernoulli($\frac{1}{2}$), the details of which can be found in Appendix B.

The next proposition focuses on the behavior of special points in the optimization space for u. In particular, we show that Algorithm 1 enables us to move away from the stationary points $\mathbf{0}_n$ and $\mathbf{1}_n$, whereas in MFVI, the optimization algorithm gets trapped in these stationary points [18].

Proposition 1 (Escaping from stationary points)**.**

(i) $(\psi^{00}, \psi^{01}, \psi^{10}, \psi^{11}) = (\mathbf{1}, \mathbf{0}, \mathbf{0}, \mathbf{0}), \ (\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{1})$ *(these vectors are m-dimensional) are the stationary points of the pairwise structured ELBO when* p, q *are known, which maps to* $u = \mathbf{0}_n$ *and* $\mathbf{1}_n$ *respectively.*

(ii) With the updates in Algorithm 1, when $u_0 = \mathbf{0}_n, \mathbf{1}_n$, *VIPS converges to the true labels with* $\|u_1^{(3)} - z^*\|_1 = n\exp(-\Omega_P(n\rho_n)))$.

The above results requires knowing the true p and q. The next corollary shows that, even if we do not have access to the true parameters, as long as some reasonable estimates can be obtained, the same convergence as in Theorem 1 holds thus demonstrating robustness to misspecified parameters. Here we hold the parameters fixed and only update u as in Algorithm 1.

Proposition 2 (Parameter robustness)**.** *If we replace true* p, q *with some estimation* $\hat{p}, \hat{q}$ *in Algorithm 1, the same conclusion as in Theorem 1 holds if*

$$1.\ \tfrac{p+q}{2} > \hat{\lambda}, \quad 2.\ \hat{\lambda} - q = \Omega(\rho_n), \quad 3.\ \hat{t} = \Omega(1).$$

where $\hat{t} = \frac{1}{2}\log\frac{\hat{p}/(1-\hat{p})}{\hat{q}/(1-\hat{q})}$, $\hat{\lambda} = \frac{1}{2\hat{t}}\log\frac{1-\hat{q}}{1-\hat{p}}$.

When $\hat{p}, \hat{q} \asymp \rho_n$, we need $\hat{p} - \hat{q} = \Omega(\rho_n)$ and $\hat{p}, \hat{q}$ not too far from the true values to achieve convergence. The proof is deferred to the Appendix.

Finally, we consider updating the parameters jointly with u (as explained in Remark 1) by first initializing the algorithm with some reasonable $p^{(0)}, q^{(0)}$.

Theorem 2 (Updating parameters and u simultaneously)**.** *Suppose we initialize with some estimates of true* (p, q) *as* $\hat{p} = p^{(0)}$, $\hat{q} = q^{(0)}$ *satisfying the conditions in Proposition 2 and apply two meta iterations in Algorithm 1 to update* u *before updating* $\hat{p} = p^{(1)}, \hat{q} = q^{(1)}$. *After this, we alternate between updating* u *and the parameters after each meta iteration. Then*

$$p^{(1)} = p + O_P(\sqrt{\rho_n}/n), \quad q^{(1)} = q + O_P(\sqrt{\rho_n}/n),$$
$$\|u_3^{(2)} - z^*\|_1 = n\exp(-\Omega(n\rho_n)),$$

and the same holds for all the later iterations.

4 Experiments

In this section, we present some numerical results. In Figures 2 to 4 we show the effectiveness of VIPS in our theoretical setting of two equal sized communities. In Figures 5 (a) and (b) we show that empirically the advantage of VIPS holds even for unbalanced community sizes and $K > 2$. Our goal is two-fold: (i) we demonstrate that the empirical convergence behavior of VIPS coincides well with our theoretical analysis in Section 3; (ii) in practice VIPS has superior performance over MFVI in both the simple setting we have analyzed and more general settings, thus confirming the advantage of the added dependence structure. For the sake of completeness, we also include comparisons with other popular algorithms, even though it is not our goal to show VIPS outperforms these methods.

In Figure 2, we compare the convergence property of VIPS with MFVI for initialization from independent Bernoulli's with means $\mu = 0.1, 0.5$, and 0.9. We randomly generate a graph with $n = 3000$ nodes with parameters $p_0 = 0.2, q_0 = 0.01$ and show results from 20 random trials. We plot $\min(\|u - z^*\|_1, \|u - (\mathbf{1} - z^*)\|_1)$, or the ℓ_1 distance of the estimated label u to the ground truth z^* on the Y axis versus the iteration number on the X axis. In this experiments, both VIPS and MFVI were run with the true p_0, q_0 values. As shown in Figure 2, when $\mu = \frac{1}{2}$, VIPS converges to z^* after two meta iterations (6 iterations) for all the random initializations. In contrast, for MFVI, a fraction of the random initializations converge to $\mathbf{0}_n$ and $\mathbf{1}_n$. When $\mu \neq \frac{1}{2}$, VIPS converges to the ground truth after three meta iterations, whereas MFVI stays at the stationary points $\mathbf{0}_n$ and $\mathbf{1}_n$. This is consistent with our theoretical results in Theorem 1 and Corollary 1, and those in [18].

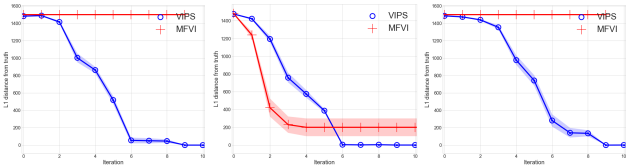

Figure 2: ℓ_1 distance from ground truth (Y axis) vs. number of iterations (X axis). The line is the mean of 20 random trials and the shaded area shows the standard deviation. u is initialized from i.i.d. Bernoulli with mean $\mu = 0.1, 0.5, 0.9$ from the left to right.

In Figure 3, we show when the true p, q are unknown, the dependence structure makes the algorithm more robust to estimation errors in $\hat{p}, \hat{q}$. The heatmap represents the normalized mutual information (NMI) [54] between u and z^*, with $\hat{p}$ on the X axis and $\hat{q}$ on the Y axis. We only examine pairs with $\hat{p} > \hat{q}$. Both VIPS and MFVI were run with $\hat{p}$ and $\hat{q}$, which were held fixed and differ from the true values to varying extent. The dashed line represents the true p, q used to generate the graph. For each $\hat{p}, \hat{q}$ pair, the mean NMI for 20 random initializations from i.i.d Bernoulli($\frac{1}{2}$) is shown. VIPS recovers the ground truth in a wider range of $\hat{p}, \hat{q}$ values than MFVI. We show in Section D of the Appendix that similar results also hold for $K = 2$ with unbalanced community sizes.

In Figure 4, we compare VIPS with MFVI under different network sparsities and signal-to-noise ratios (SNR) as defined by $r_0 = p_0/q_0$. For the sake of completeness, we also include two other popular algorithms, Belief Propagation (BP) [55] and Spectral Clustering [31]. We plot the mean and standard deviation of NMI for

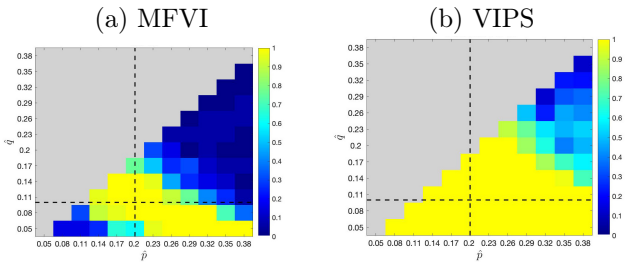

Figure 3: NMI averaged over 20 random initializations for each $\hat{p}, \hat{q}$ ($\hat{p} > \hat{q}$). The true parameters are $(p_0, q_0) = (0.2, 0.1)$, $\pi = 0.5$ and $n = 2000$. The dashed lines indicate the true parameter values.

20 random trials in each setting. In each trial, to meet the conditions in Theorem 2, we started VIPS with $\hat{p}$ equal to the average degree of A, and $\hat{q} = \hat{p}/r_0$. $\hat{p}$ and $\hat{q}$ were updated alternatingly with u according to Eq. (14) after three meta iterations in Algorithm 1, a setting similar to that of Theorem 2.

In Figure 4-(a), the average expected degree is fixed as the SNR p_0/q_0 increases on the X axis, whereas in Figure 4-(b), the SNR is fixed and we vary the average expected degree on the X axis. The results show that VIPS consistently outperforms MFVI, indicating the advantage of the added dependence structure. Note that we plot BP with the model parameters initialized at true (p_0, q_0) , since it is sensitive to initialization setting, and behaves poorly with mis-specified ones. Despite this, VIPS is largely comparable to BP and Spectral Clustering. For average degree 20 (Figure 4-(b)), BP outperforms all other methods, because of the correct parameter setting. This NMI value becomes 0.4 with high variance, if we provide initial $\hat{p}, \hat{q}$ values to match the average degree but $\hat{p}/\hat{q} = 10$. In contrast, VIPS is much more robust to the initial choice of $\hat{p}, \hat{q}$, which we show in Section C of the Appendix.

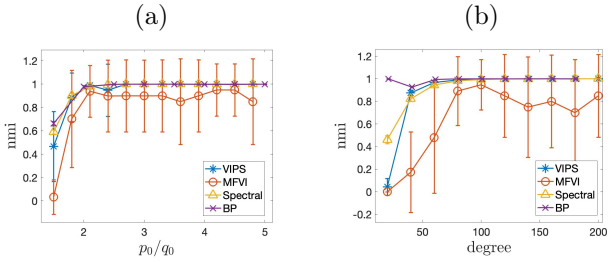

Figure 4: Comparison of NMI under different SNR p_0/q_0 and network degrees. The lines and error bars are means and standard deviations from 20 random trials. (a) Vary p_0/q_0 with degree fixed at 70. (b) Vary the degree with $p_0/q_0 = 2$. In both figures $n = 2000$.

Additional experiments (Appendix, section D) show that VIPS with fixed mis-specified parameters (within reasonable deviation from the truth), fixed true parameters and parameters updated with Eq. (14) converge to the truth when initialized by independent Bernoulli's.

5 Discussion and Generalizations

In this paper, we propose a simple Variational Inference algorithm with Pairwise Structure (VIPS) in a SBM with two equal sized communities. VI has been extensively applied in the latent variable models mainly due to their scalability and flexibility for incorporating changes in model structure. However, theoretical understanding of the convergence properties is limited and mostly restricted to the mean field setting with fully factorized variational distributions (MFVI). Theoretically we prove that in a SBM with two equal sized communities, VIPS can converge to the ground truth with probability tending to one for different random initialization schemes and a range of graph densities. In contrast, MFVI only converges for a constant fraction of Bernoulli(1/2) random initializations. We consider settings where the model parameters are known, estimated or appropriately updated as part of the iterative algorithm.

Though our main results are for $K = 2, \pi = 0.5$, we conclude with a discussion on generalizations to unbalanced clusters and SBMs with $K > 2$ equal communities. To apply VIPS for general $K > 2$ clusters, we will have $K^2 - 1$ categorical distribution parameters ψ^{cd} for $c, d \in \{1, 2, \ldots, K\}$ and marginal likelihood $\phi_1, \ldots, \phi_{K-1}, \xi_1, \ldots, \xi_{K-1}$. The updates are similar to Eq. (11) and Eq. (12) and are deferred to the Appendix (section C). Similar to the $K = 2$ case, we update the local and global parameters iteratively. As for the unbalanced case (see Appendix Section C), the updates involve an additional term which is the logit of π. We assume that π is known and fixed.

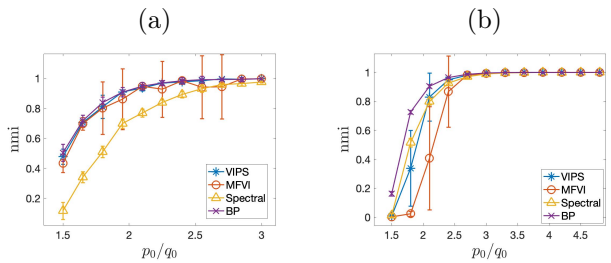

Figure 5: Comparison of VIPS, MFVI, Spectral and BP using error-bars from 20 random trials for $n = 2000$, average degree 50, p_0/q_0 is changed on X axis. (a) $\pi = 0.3$ (b) K =3, $B = (p - q)I + qJ$. For BP, MFVI and VIPS, we use true parameters.

In Figure 5-(a), we show results for unbalanced SBM

with $\pi = 0.3$, which is assumed to be known. In Figure 5-(b), similar to the setting in [18], we consider a SBM with three equal-sized communities. The parameters are set as $n = 2000$, average degree 50, p_0 and q_0 are changed to get different SNR values and the random initialization is from Dirichlet$(1, 1, 1)$. For a fair comparison of VIPS, MFVI and BP, we use the true p_0, q_0 values in all three algorithms; robustness to parameter specification of VIPS is included in the Appendix C. We see that for the unbalanced setting (Figure 5-(a)) VIPS performs as well as BP and better than Spectral Clustering. For the $K = 3$ setting (Figure 5-(b)) VIPS performs worse than BP and Spectral for very low SNR values, whereas for higher SNR it performs comparably to Spectral and BP, and better than MFVI, which has much higher variance.

References

[1] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, November 1999.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[3] Tommi S. Jaakkola and Michael I. Jordon. Improving the mean field approximation via the use of mixture distributions. In *Learning in Graphical Models*, pages 163–173. MIT Press, Cambridge, MA, USA, 1999.

[4] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[5] David M Blei, Michael I Jordan, et al. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143, 2006.

[6] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[7] Sivaraman Balakrishnan, Martin J Wainwright, Bin Yu, et al. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

[8] Ji Xu, Daniel J Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two gaussians. In *Advances in Neural Information Processing Systems*, pages 2676–2684, 2016.

[9] Bowei Yan, Mingzhang Yin, and Purnamrita Sarkar. Convergence of gradient em on multi-component mixture of gaussians. In *Advances in Neural Information Processing Systems*, pages 6956–6966, 2017.

[10] Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. In *International Conference on Machine Learning*, pages 613–621, 2014.

[11] Jeongyeol Kwon and Constantine Caramanis. Global convergence of em algorithm for mixtures of two component linear regression. *arXiv preprint arXiv:1810.05752*, 2018.

[12] Debdeep Pati, Anirban Bhattacharya, and Yun Yang. On statistical optimality of variational bayes. *arXiv preprint arXiv:1712.08983*, 2017.

[13] Ted Westling and Tyler H. McCormick. Beyond prediction: A framework for inference with variational approximations in mixture models. *arXiv preprint arXiv:1510.08151*, 2015.

[14] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[15] Bo Wang, DM Titterington, et al. Convergence properties of a general algorithm for calculating variational bayesian estimates for a normal mixture model. *Bayesian Analysis*, 1(3):625–650, 2006.

[16] Pranjal Awasthi and Andrej Risteski. On some provably correct cases of variational inference for topic models. In *Advances in Neural Information Processing Systems*, pages 2098–2106, 2015.

[17] Behrooz Ghorbani, Hamid Javadi, and Andrea Montanari. An instability in variational inference for topic models. *arXiv preprint arXiv:1802.00568*, 2018.

[18] Soumendu Sundar Mukherjee, Purnamrita Sarkar, YX Rachel Wang, and Bowei Yan. Mean field for the stochastic blockmodel: Optimization landscape and convergence issues. In *Advances in Neural Information Processing Systems*, pages 10694–10704, 2018.

[19] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[20] T Jaakkola. Tutorial on variational approximation methods. *Advanced mean field methods: theory and practice*, page 129, 2001.

[21] Eric P Xing, Michael I Jordan, and Stuart Russell. A generalized mean field algorithm for variational inference in exponential families. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 583–591. Morgan Kaufmann Publishers Inc., 2002.

[22] Matthew D Hoffman and David M Blei. Structured stochastic variational inference. In *Artificial Intelligence and Statistics*, 2015.

[23] Ryan J Giordano, Tamara Broderick, and Michael I Jordan. Linear response methods for accurate covariance estimates from mean field variational bayes. In *NIPS*, pages 1441–1449, 2015.

[24] Dustin Tran, David Blei, and Edo M Airoldi. Copula variational inference. In *NIPS*, pages 3564–3572, 2015.

[25] Rajesh Ranganath, Dustin Tran, and David Blei. Hierarchical variational models. In *ICML*, pages 324–333, 2016.

[26] Mingzhang Yin and Mingyuan Zhou. Semi-implicit variational inference. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5660–5669. PMLR, 2018.

[27] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *ICML*, pages 1530–1538, 2015.

[28] Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-free variational inference. In *NIPS*, pages 5529–5539, 2017.

[29] Anderson Y Zhang and Harrison H Zhou. Theoretical and computational guarantees of mean field variational inference for community detection. *arXiv preprint arXiv:1710.11268*, 2017.

[30] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[31] Karl Rohe, Sourav Chatterjee, Bin Yu, et al. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

[32] Amin Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19(2):227–284, 2010.

[33] Olivier Guédon and Roman Vershynin. Community detection in sparse networks via grothendieck's inequality. *Probability Theory and Related Fields*, 165(3-4):1025–1049, 2016.

[34] Amelia Perry and Alexander S Wein. A semidefinite program for unbalanced multisection in the stochastic block model. In *2017 International Conference on Sampling Theory and Applications (SampTA)*, pages 64–67. IEEE, 2017.

[35] Arash A Amini, Elizaveta Levina, et al. On semidefinite relaxations for the block model. *The Annals of Statistics*, 46(1):149–179, 2018.

[36] A. A. Amini, A. Chen, P. J. Bickel, and E. Levina. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122, 2013.

[37] T. A. B. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100, 1997.

[38] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[39] P. J. Bickel and A. Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

[40] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of machine learning research*, 9(Sep):1981–2014, 2008.

[41] Zahra S Razaee, Arash A Amini, and Jingyi Jessica Li. Matched bipartite block model with covariates. *Journal of Machine Learning Research*, 20(34):1–44, 2019.

[42] Catherine Matias and Vincent Miele. Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1119–1141, 2017.

[43] Prem K Gopalan and David M Blei. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences*, 110(36):14534–14539, 2013.

[44] Peter Bickel, David Choi, Xiangyu Chang, Hai Zhang, et al. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, 41(4):1922–1943, 2013.

[45] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

[46] Elchanan Mossel, Joe Neeman, Allan Sly, et al. Belief propagation, robust reconstruction and optimal recovery of block models. *The Annals of Applied Probability*, 26(4):2211–2256, 2016.

[47] Emmanuel Abbe and Colin Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap. *arXiv preprint arXiv:1512.09080*, 2015.

[48] Paul Erdös. On a lemma of littlewood and offord. *Bulletin of the American Mathematical Society*, 51(12):898–902, 1945.

[49] Purnamrita Sarkar, Y. X. Rachel Wang, and Soumendu Sunder Mukherjee. When random initializations help: a study of variational inference for community detection. *arXiv e-prints*, May 2019.

[50] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

[51] Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2015.

[52] Frank McSherry. Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 529–537. IEEE, 2001.

[53] Kamalika Chaudhuri, Fan Chung Graham, and Alexander Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. In *COLT*, volume 23 of *JMLR Proceedings*, pages 35.1–35.23. JMLR.org, 2012.

[54] Simone Romano, James Bailey, Vinh Nguyen, and Karin Verspoor. Standardized mutual information for clustering comparisons: one step further in adjustment for chance. In *International Conference on Machine Learning*, pages 1143–1151, 2014.

[55] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E*, 84:066106, Dec 2011.