On Variational Inference in Biclustering Models

Guanhua Fang, Ping Li
Cognitive Computing Lab
Baidu Research
10900 NE 8th St Bellevue WA 98004 USA
{guanhuafang, liping11}@baidu.com

Abstract

Biclustering structures exist ubiquitously in data matrices and the biclustering problem was first formalized by John Hartigan (1972) to cluster rows and columns simultaneously. In this paper, we develop a theory for the estimation of general biclustering models, where the data is assumed to follow certain statistical distribution with underlying biclustering structure. Due to the existence of latent variables, directly computing the maximal likelihood estimator is prohibitively difficult in practice and we instead consider the variational inference (VI) approach to solve the parameter estimation problem. Although variational inference method generally has good empirical performance, there are very few theoretical results around VI. In this paper, we obtain the precise estimation bound of variational estimator and show that it matches the minimax rate in terms of estimation error under mild assumptions in biclustering setting. Furthermore, we study the convergence property of the coordinate ascent variational inference algorithm, where both local and global convergence results have been provided. Numerical results validate our new theories.

1. Introduction

In a wide range of data analytic scenarios, we encounter two-mode matrices with biclustering structures (Hartigan, 1972) and we might be interested in modeling the effects of both rows and columns on the data matrix. Consider the following specific situations. In educational testing (Templin and Henson, 2010; Matechou et al., 2016), two modes could be test takers and question items. We expect that test takers with same skills will form a group and similar questions may also form into different groups. In gene expression studies (Prelić et al., 2006; Gu and Liu, 2008), one can organize the data matrix such that each row corresponds to a cancer patients and each column corresponds to transcript. Then the patients can form groups according to different cancer subtypes and the genes are also expected to exhibit clustering effect according to different pathways they belong to. In online e-commerce service (Dolnicar et al., 2012), researcher may wish to study user behaviors through their purchasing and navigation history. Users and items can be viewed as two modes. Users with similar shopping preferences can be clustered into a group and items with same functionality can also be grouped together.

To capture aforementioned group effects, statistical models are often used for modeling the structure of two-mode data matrix. Latent variables are introduced to capture underlying block effects (Govaert and Nadif, 2010). However, in two-mode block mixture models, the maximal likelihood estimator (MLE) is prohibitively hard to compute since computing marginal likelihood requires summation over exponentially many terms. Quite a few computational methods are developed to estimate the parameters in biclustering models (e.g., double k-means method (Maurizio, 2001), model-based expectation-maximization (EM) method (Pledger and Arnold, 2014), Gibbs-sampling based method (Meeds and Roweis, 2007; Gu and Liu, 2008), nonparametric Bayesian method (Niu et al., 2012)). Unfortunately, these methods have no corresponding theoretical results. We aim to fill this gap in the literature.

Variational inference (VI) approach (Jordan et al., 1999; Hoffman et al., 2013) is a powerful tool for parameter estimation in complicated hierarchical latent variable models. When the analytic form of posterior distribution of latent variables cannot be computed, VI seeks a good candidate distribution to approximate the true posterior and reduces computation complexity. VI method has also become popular in biclustering models in the recent literature (Guan et al., 2010; Vu and Aitkin, 2015) . In this paper, we start from a theoretical perspective and consider an estimation problem for biclustering models via variational inference and develop the corresponding theory. Specifically, we show both upper

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

and lower estimation bounds for varitaional estimator under biclustering setting, which bridges the gap in the literature of VI theory. Moreover, we study the coordinate ascent variational inference (CAVI) algorithm for parameter estimation and we establish the local convergence property of CAVI method. Furthermore, we also provide the global convergence results to discuss the situations under which CAVI may or may not return a consistent estimator. To the best of our knowledge, this is the first time that relatively complete theoretical results have been provided on variational inference estimation under the general biclustering settings.

The rest of paper is organized as follows. In Section 2, we provide a preliminary of biclustering models and variational inference methods. In Section 3, we study the theoretical properties of variational estimator and give detailed estimation bounds. In Section 4, we study the convergence property of CAVI algorithm. Both local and global convergence results are provided. Multiple numerical results are provided in Section 5 to support our theoretical findings. Finally, the concluding remarks are given in Section 6.

Notation. For positive integer m, we use $[m]$ to denote set $\{1, \ldots, m\}$ and $[m_1] \times [m_2]$ to denote set $\{(i, j) : i \in [m_1], j \in [m_2]\}$. For two positive sequences $\{a_n\}$, $\{b_n\}$, $a_n \lesssim b_n$ means that $a_n \leq C b_n$ for some large constant C independent of n, and $a_n \asymp b_n$ means that $a_n \lesssim b_n$ and $b_n \lesssim a_n$. The symbols $\mathbb{E}$ and $P(\cdot)$ denote generic expectation and probability whose distribution may be determined from the context. Additionally, $\|x\|$ / $\|x\|_1$ is used to denote ℓ_2- / ℓ_1- norm of vector x and $\|X\|_F$ is used to denote Frobenius norm of matrix X. We use $x[i]$ to represent the i-th entry of vector x and use $X[i, j]$ to represent the entry of matrix X on i-th row and j-th column. We use ∇f to represent the derivative of function f with respect to θ.

2. Preliminary

In this paper, we consider the following general biclustering model. We assume

$$Y_{ij} \sim f_\theta(y|z_{1i}, z_{2j}), \;\; i \in [m_1], j \in [m_2],$$

where z_{1i}'s and z_{2j}'s are unobserved latent memberships with

$$z_{1i} \sim \pi_1 \text{ and } z_{2j} \sim \pi_2,$$

π_1 is a discrete probability distribution over J_1 latent classes and π_2 is a discrete probability distribution over J_2 latent classes. In other words, $P(z_{1i} = k) = \pi_1[k]$ $(k \in [J_1])$ and $P(z_{2j} = l) = \pi_2[l]$ $(l \in [J_2])$. Density function $f_\theta(y|z_1, z_2)$ are parameterized by θ. Furthermore, we assume $f_\theta(y|z_1, z_2)$ can be reduced to $f_{\theta_{z_1 z_2}}(y)$, that is, the observation only depends on latent class-specific parameter.

Given observations $(y_{ij}, i \in [m_1], j \in [m_2])$, we can write the likelihood function as

$$L(\theta) = \sum_{\mathbf{z}_1, \mathbf{z}_2} \prod_i \pi_1[z_{1i}] \prod_j \pi_2[z_{2j}] \{ \prod_{i,j} f_\theta(y_{ij}|z_{1i}, z_{2j}) \},$$

where $\mathbf{z}_1 = (z_{11}, \ldots, z_{1m_1})$ and $\mathbf{z}_2 = (z_{21}, \ldots, z_{2m_2})$.

Additionally, if only a subset $\Omega \subset [m_1] \times [m_2]$ can be observed, then the corresponding likelihood function can be written as

$$L(\theta) = \sum_{\mathbf{z}_1, \mathbf{z}_2} \prod_{i,j} \pi_1[z_{1i}] \pi_2[z_{2j}] \{ \prod_{(i,j) \in \Omega} f_\theta(y_{ij}|z_{1i}, z_{2j}) \}.$$

In particular, when Y_{ij} follows the Gaussian distribution with mean $\theta_{z_{1i} z_{2j}}$ and variance 1, then

$$f_\theta(y_{ij}|z_{1i}, z_{2j}) \propto \exp\{-(y_{ij} - \theta_{z_{1i} z_{2j}})^2/2\}.$$

When Y_{ij} is the binary response, that is, $Y_{ij} \sim$ Bernoulli$(\theta_{z_{1i} z_{2j}})$ with $0 < \theta_{z_{2i} z_{2j}} < 1$, then

$$f_\theta(y_{ij}|z_{1i}, z_{2j}) = \theta_{z_{1i} z_{2j}}^{y_{ij}} (1 - \theta_{z_{1i} z_{2j}})^{1 - y_{ij}}.$$

This is also known as the stochastic block model (SBM, Holland et al., 1983; Abbe, 2017; Zhou and Li, 2020).

When Y_{ij} is the ordinal response, that is, $Y_{ij} \sim$ Multinom$(\theta_{z_{1i} z_{2j}}[1], \ldots, \theta_{z_{1i} z_{2j}}[K])$ (K is the number of categories), then

$$f_\theta(y_{ij}|z_{1i}, z_{2j}) = \prod_{k=1}^{K} (\theta_{z_{1i} z_{2j}}[k])^{\mathbf{1}\{y_{ij} = k\}}.$$

Furthermore, Y_{ij} can be an event process (i.e., $y_{ij} = (t_{ij,1}, \ldots, t_{ij,n})$ is a sequence of event times in $[0, T]$), which follows a certain counting process model with intensity function $\theta_{z_{1i} z_{2j}}(t)$. Then

$$f_\theta(y_{ij}|z_i, z_j) = \{ \prod_{l=1}^{n} \theta_{z_i z_j}(t_{ij,l}) \} \exp\{ - \int_{t=0}^{T} \theta_{z_{1i} z_{2j}}(t) dt \}.$$

Especially, when $\theta_{ij}(t) \equiv \theta_{z_{1i} z_{2j}}$, it reduces to the homogeneous Poisson point process. Therefore,

$$f_\theta(y_{ij}|z_{1i}, z_{2j}) = \theta_{z_{1i} z_{2j}}^n \exp\{-T \theta_{z_{1i} z_{2j}}\}.$$

Approximation via varitaional inference. By a close look at the formula of $L(\theta)$, it is difficult to compute the likelihood directly, which requires summation over exponentially many terms. An alternative approach is to utilize variational inference (Jordan et al., 1999; Hoffman et al., 2013) methods to optimize the evidence lower bound (ELBO) instead of the log likelihood. In general, the ELBO function is defined as

$$\text{ELBO} = \mathbb{E}_{z \sim q(z)} l(\theta, z) - \mathbb{E}_{z \sim q(z)} \log q(z),$$

where the expectation is taken with respect to latent variables z and $q(z)$ is an approximate distribution function for posterior of z. Under our setting, $z = (\mathbf{z}_1, \mathbf{z}_2)$,

$$l(\theta, z) = \sum_i \log \pi_1[z_{1j}] + \sum_j \log \pi_2[z_{2j}] + \sum_{i,j} \log f_\theta(y_{ij}|z_{1i}, z_{2j}).$$

For computational feasibility, we consider a mean-field family (Blei et al., 2017) for the choice of $q(z)$. More precisely, we take $q(z) := \prod_i q_i(z_{1i}) \prod_j q_j(z_{2j})$, $q_i(z_{1i}) = \text{multinom}(\phi_{1i})$ with $\phi_{1i} = (\phi_{1i}[1], \ldots, \phi_{1i}[J_1])$ and $q_j(z_{2j}) = \text{multinom}(\phi_{2j})$ with $\phi_{2j} = (\phi_{2j}[1], \ldots, \phi_{2j}[J_z])$. Here $\text{multinom}(\phi)$ represents a multinomial distribution with parameter ϕ. By calculation, the ELBO can be obtained,

$$\begin{aligned}\text{ELBO} = & \sum_{(i,j)} \sum_{k,l} \phi_{1i}[k]\phi_{2j}[l] \log f_\theta(y_{ij}|k,l) \\ & + \sum_i \sum_k \phi_{1i}[k] \log(\pi_1[k]/\phi_{1i}[k]) \\ & + \sum_j \sum_l \phi_{2j}[l] \log(\pi_2[l]/\phi_{2j}[l]). \end{aligned} \quad (1)$$

Although variational inference is a powerful tool in optimization for complex statistical models, the VI theory has not been fully explored yet in the literature.

3. Estimation Bound of Variational Estimator

It is known that the ELBO function is a lower bound of the log-likelihood function, that is,

$$\log L(\theta) - \text{ELBO} = KL(q(z)\|p_\theta(z|y)).$$

Since KL divergence is always non-negative, therefore the maximizer of ELBO may not be an unbiased estimator of true parameter. Under the biclustering model, we have

$$\begin{aligned} & KL(q(z)\|p_\theta(z|y)) \\ = \; & KL(\prod_i q(z_{1i}) \prod_j q(z_{2j}) \| p_\theta(\mathbf{z}_1, \mathbf{z}_2|y)) \\ = \; & \sum_{\mathbf{z}_1, \mathbf{z}_2} \prod_i \phi_{1i}[z_{1i}] \prod_j \phi_{2j}[z_{2j}] \cdot \\ & \{\log(\prod_i \phi_{1i}[z_{1i}] \prod_j \phi_{2j}[z_{2j}]) - \log p_\theta(\mathbf{z}_1, \mathbf{z}_2|y)\}. \end{aligned}$$

As sample sizes m_1 and m_2 go to infinity, we are able to show that $KL(q(z)\|p_\theta(z|y)$ goes to 0 in probability. This can guarantee that the VI estimator is asymptotically consistent. Before going to the main results, we first introduce some additional notations and assumptions.

Assumption A1: For every $k \neq k' \in [J_1]$, there exists $l \in [J_2]$ such that

$$\theta_{kl} \neq \theta_{k'l}. \quad (2)$$

For every $l \neq l' \in [J_2]$, there exists $k \in [J_1]$ such that

$$\theta_{kl} \neq \theta_{kl'}. \quad (3)$$

Assumption A1 is an identifiability assumption that the matrix θ cannot have two same columns or two same rows. This constraint ensures that the individuals from different classes should have different structural properties. For example, when $J_1 = J_2 = 2$ and $\theta = \begin{pmatrix} \theta_a & \theta_a \\ \theta_b & \theta_b \end{pmatrix}$ $(\theta_a \neq \theta_b)$, we can easily differentiate objects from difference classes for the first mode, while we fail to classify the objects for the second mode. This can lead to mis-classification of z_{2j}'s.

Assumption A2: There exists a bounded and compact set $\mathcal{B}$ such that

$$\theta_{kl} \in \mathcal{B}, \text{ for } k \in [J_1], l \in [J_2].$$

Assumption A2 is a compactness assumption to make sure the objective function has nice continuity property. For example, when Y_{ij} is binary and has distribution Bernoulli$(\theta_{z_i z_j})$, we require $\theta_{kl} \in [\xi, 1-\xi]$ for positive constant ξ. This is also a usual assumption in SBM literature (Celisse et al., 2012).

Assumption A3: There exist positive constants γ_1 and γ_2 such that

$$\pi_1[k] \in [\gamma_1, 1-\gamma_1], \text{ for } k \in [J_1]$$

and

$$\pi_2[l] \in [\gamma_2, 1-\gamma_2], \text{ for } l \in [J_2].$$

The above assumption implies that no class is drained. In this paper, γ_1 and γ_2 are assumed to be free of sample sizes, m_1 and m_2. Moreover, z_{1i}'s and z_{2j}'s are the realizations of multinomial random variables with parameter π_1 and π_2 respectively. Define empirical latent class probability $\tilde{\pi}_1[k] := \frac{N_1[k]}{m_1}$ and $\tilde{\pi}_2[l] := \frac{N_2[l]}{m_2}$, where $N_1[k] := |\{i \in [J_1] : z_{1i} = k\}|$ and $N_2[l] := |\{j \in [J_2] : z_{2j} = l\}\}|$. Therefore, by large of large number, we know that $\tilde{\pi}_1 \in [\gamma_1/2, 1-\gamma_1/2]$ and $\tilde{\pi}_2 \in [\gamma_2/2, 1-\gamma_2/2]$. But unfortunately, we do not have access to z_{1i}'s and z_{2j}'s. They have to be estimated as well.

Assumption A4: It is assumed that $f_\theta(y|z_1, z_2) \equiv f_{\theta_{z_1,z_2}}(y)$ for $z_1 \in [J_1]$ and $z_2 \in [J_2]$ and is log-concave twice differentiable function. We define

$$h_{kl}(\theta) = \mathbb{E}_{y \sim f_{\theta_{kl}}} \log f_\theta(y) \quad (4)$$

and assume $h_{kl}(\theta)$ is a strictly concave function of θ over $\mathcal{B}$ for any $k \in [J_1], l \in [J_2]$.

Assumption A4 puts the requirement on the density of biclustering model. Smooth function $f_\theta(y|z_1, z_2)$ is assumed to depend on θ_{kl} only. Function $h_{kl}(\theta)$ is the expectation of log density function with respect to true distribution. In most case, $h_{kl}(\theta)$ is naturally a strictly concave function. For example, when $f_\theta(y)$ is the density of Gaussian distribution with variance 1, we have

$$h_{kl}(\theta) = -\frac{1}{2}(\theta_{kl} - \theta)^2,$$

which is obviously strictly concave. When $f_\theta(y)$ is the density of Bernoulli(θ), then

$$h_{kl}(\theta) = \theta_{kl} \log \theta + (1 - \theta_{kl}) \log(1 - \theta),$$

which is also strictly concave on its domain.

Next, we define the variational estimator $\hat{\theta}_{kl}$'s, $\hat{\pi}_1$, $\hat{\pi}_2$, $\hat{\phi}_{1i}$'s and $\hat{\phi}_{2j}$'s as

$$(\hat{\theta}, \hat{\pi}_1, \hat{\pi}_2, \hat{\phi}_{1i}, \hat{\phi}_{2j}) := \arg \max_{\theta, \pi_1, \pi_2, \phi_{1i}, \phi_{2j}} \text{ELBO}.$$

The domain for original parameter is $\theta_{kl} \in \mathcal{B}$, $\pi_1[k] \in [\gamma_1, 1 - \gamma_1]$, $\pi_2[l] \in [\gamma_2, 1 - \gamma_2]$. The domain for variational parameter is $\phi_{1i} \in \mathcal{S}_1$ and $\phi_{2j} \in \mathcal{S}_2$, where $\mathcal{S}_1 = \{\sum_{k=1}^{J_1} \phi[k] = 1 \text{ and } 0 \le \phi[k] \le 1\}$ and $\mathcal{S}_2 = \{\sum_{l=1}^{J_2} \phi[l] = 1 \text{ and } 0 \le \phi[l] \le 1\}$. In a summary, the total number of parameters is $J_1 J_2 + (J_1 - 1)m_1 + (J_2 - 1)m_2 + J_1 + J_2 - 2$.

3.1. Theoretical Results

Next, we provide the theoretical estimation bounds for the variational estimator, including classification consistency, population consistency and parameter consistency. We use superscript "$*$" to denote the true values in the rest of paper.

Classification consistency. We use δ_{1k} to denote the probability mass function on discrete sets $\{1, \ldots, J_1\}$ that assigns the total probability at $k \in [J_1]$ and use δ_{2l} to denote the probability mass function on discrete sets $\{1, \ldots, J_2\}$ that puts the total probability at $l \in [J_2]$. Let $\hat{q}_{1i} = \text{multinom}(\hat{\phi}_{1i})$ and $\hat{q}_{2j} = \text{multinom}(\hat{\phi}_{2j})$ be the estimated approximate posterior distribution for z_{1i} and z_{2j}, respectively. Note that the model is invariant in regard to class label permutation. Without loss of generality, we always assume that the estimated class labels can be permuted to match the true labels when the estimator is consistent. We then can show that the $\hat{q}_{1i}$ converges to $\delta_{1z^*_{1i}}$ and $\hat{q}_{2j}$ converges to $\delta_{2z^*_{2j}}$. The result is stated in Theorem 1.

Theorem 1 *Under Assumptions A1 - A4, there exist constants c_0 and C such that*

$$d_{TV}(\hat{q}_{1i}, \delta_{1z^*_{1i}}) \le J_1 \exp\{-c_0 m_2\}$$

and

$$d_{TV}(\hat{q}_{2j}, \delta_{1z^*_{2j}}) \le J_2 \exp\{-c_0 m_1\}$$

hold with probability at least $1 - (m_1 + m_2) \max\{J_1, J_2\} \exp\{-C \min\{m_1, m_2\}\}$ *for* $i \in [J_1]$ *and* $j \in [J_2]$.

Here, d_{TV} is the total variation distance between two distribution. Theorem 1 tells us that we can classify z_{1i}'s and z_{2j}'s into correct latent classes with high probability. To be more specific, we define estimated label as $\hat{z}_{1i} := \arg\max_{k \in [J_1]} \hat{q}_{1i}[k]$, $\hat{z}_{2j} := \arg\max_{l \in [J_2]} \hat{q}_{2j}[l]$, and $\hat{\mathbf{z}}_1 = (\hat{z}_{11}, \ldots, \hat{z}_{1,m_1})$, $\hat{\mathbf{z}}_2 = (\hat{z}_{21}, \ldots, \hat{z}_{2m_2})$. We then have $P(\hat{\mathbf{z}}_1 = \mathbf{z}^*_1, \hat{\mathbf{z}}_2 = \mathbf{z}^*_2) \to 1$, which is known as *strong consistency* in the SBM literature (Abbe et al., 2015; Mossel et al., 2014; Gao et al., 2017). Moreover, mis-classification errors decrease exponentially fast in terms of sample sizes.

Population consistency. By the definition of variational estimator, we can obtain the relation between $\hat{\pi}_1, \hat{\pi}_2$ and $\hat{\phi}_{1i}$'s, $\hat{\phi}_{2i}$'s, which is

$$\begin{aligned} \hat{\pi}_1[k] &= \frac{\sum_{i \in [m_1]} \hat{\phi}_{1i}[k]}{m_1} \text{ for } k \in [J_1] \\ \hat{\pi}_2[l] &= \frac{\sum_{j \in [m_2]} \hat{\phi}_{2j}[l]}{m_1} \text{ for } l \in [J_2] \end{aligned}$$

by simplifying the optimality conditions. By Theorem 1 and law of large number, we can obtain the consistency of $\hat{\pi}_1$ and $\hat{\pi}_2$.

Theorem 2 *Under Assumptions A1 - A4, there exist constants $c_0 - c_2$ and C such that*

$$|\hat{\pi}_1[k] - \pi^*_1[k]| \le J_1 \exp\{-c_0 m_2\} + \frac{c_1}{\sqrt{m_1}}$$

and

$$|\hat{\pi}_2[l] - \pi^*_2[l]| \le J_2 \exp\{-c_0 m_1\} + \frac{c_2}{\sqrt{m_2}}$$

hold with probability at least $1 - 2(m_1 + m_2) \max\{J_1, J_2\} \exp\{-C \min\{m_1, m_2\}\}$ *for* $k \in [J_1]$ *and* $l \in [J_2]$.

Here the estimation errors of π_1 and π_2 come from two sources, variational approximation and sampling noise (i.e., the deviation between empirical distribution of z_{1i}'s / z_{2i}'s and true prior π_1 / π_2.). When both m_1, m_2 go to infinity, then the sampling noise will become the dominated term. Hence the estimation errors of $\hat{\pi}_1$ and $\hat{\pi}_2$ achieve the optimal rate, i.e., $O(\frac{1}{\sqrt{m_1}})$ and $O(\frac{1}{\sqrt{m_2}})$.

Parameter consistency. Next we move onto the estimation of θ which is the key parameter to differentiate between different classes. The upper bound is given in the following theorem.

Theorem 3 *Under Assumptions $A1-A4$, there exist constants C_1-C_3 such that it holds*

$$\frac{\|\hat{\theta}-\theta^*\|_F^2}{J_1J_2} \leq \frac{C_1(J_1J_2+m_1\log J_1+m_2\log J_2)}{m_1m_2} + \max\{J_1,J_2\}\exp\{-C_2\min\{m_1,m_2\}\} \quad (5)$$

with probability at least

$$1-\exp\{-C_3(J_1J_2+m_1\log J_1+m_2\log J_2)\}- (m_1+m_2)\max\{J_1,J_2\}\exp\{-C_3\min\{m_1,m_2\}\}.$$

Again, we can see that the upper bound for $\frac{1}{J_1J_2}\|\hat{\theta}-\theta\|_F^2$ consists of two main parts (sampling noises and variational approximation errors). When both m_1 and m_2 go to infinity at the same order, the first term on the right hand side of (5) will denominate the second term. In fact, $\frac{(J_1J_2+m_1\log J_1+m_2\log J_2)}{m_1m_2}$ is the optimal error rate for any estimator in the estimation of biclustering model, see explanations in the next part. Moreover, the part involving J_1J_2 reflects the number of parameters in θ, while the part involving $(m_1\log J_1+m_2\log J_2)$ comes from the complexity of estimating the latent memberships for both modes. It is the price we need to pay when we do not know the true clustering information.

Lower bound of parameter estimation. To show that our upper bound is tight, we first reformulate our parameter setting. We define an augmented parameter space,

$$\mathcal{B}_\Theta := \{\Theta\in\mathbb{R}^{m_1\times m_2}:\Theta_{ij}=\theta_{z_{1i}z_{2j}}, z_{1i}\in[J_1], z_{2j}\in[J_2], \theta\in\mathcal{B}\}. \quad (6)$$

In other words, the parameter Θ is constructed based on θ by letting $\Theta_{ij}=\theta_{z_{1i}z_{2j}}$ through the latent memberships. Let $\hat{z}_{1i}=\arg\max_k\hat{\phi}_{1i}[k]$ and $\hat{z}_{2j}=\arg\max_j\hat{\phi}_{2j}[k]$ be the estimated latent memberships. We then can construct $\hat{\Theta}$ as $\hat{\Theta}_{ij}=\hat{\theta}_{\hat{z}_{1i}\hat{z}_{2j}}$. Similarly, we can define Θ^* in the same way. We then have

$$\frac{1}{m_1m_2}\|\hat{\Theta}-\Theta^*\|_F^2 \asymp \frac{1}{J_1J_2}\|\hat{\theta}-\theta^*\|_F^2.$$

Therefore, we only need to work on the lower bound of $\frac{1}{m_1m_2}\|\hat{\Theta}-\Theta^*\|_F^2$.

Theorem 4 *Under Assumptions $A1-A4$, there exist some constants $C,c>0$ such that*

$$\inf_{\check{\Theta}}\sup_{\Theta^*\in\mathcal{B}_\Theta} P\left(\frac{1}{m_1m_2}\|\check{\Theta}-\Theta^*\|_F^2 > \frac{C(J_1J_2+m_1\log J_1+m_2\log J_2)}{m_1m_2}\right)>c.$$

By Theorem 3 and Theorem 4, we know that the variational estimator could achieve the minimax rate when sample sizes m_1 and m_2 tend to infinity at the same order. Compared with the error bound of MLE (Gao et al., 2016), we can see that the bias induced by variational approximation is only of order $\exp\{-C\min\{m_1,m_2\}\}$ which is negligible when both m_1 and m_2 go to infinity.

Partial observation case. Furthermore, we consider the situation that the data may be only partially observed. That is, we observe $(y_{ij}:(i,j)\in\Omega)$ instead of $(y_{ij}:(i,j)\in[m_1]\times[m_2])$, where Ω is the subset of $[m_1]\times[m_2]$. In addition, we assume that each pair (i,j) is observed completely at random with fixed observation rate p $(0<p<1)$. Thus, the size of Ω is approximately pm_1m_n. Under this setting, we can generalize our theoretical results as follows.

Theorem 5 *Under the partial observation setting and Assumptions $A1-A4$, there exist $C_0'-C_6'$ such that*

$$\begin{aligned} d_{TV}(\hat{q}_{1i},\delta_{1z_{1i}^*}) &\leq J_1\exp\{-C_0'pm_2\},\\ d_{TV}(\hat{q}_{2j},\delta_{1z_{2j}^*}) &\leq J_2\exp\{-C_0'pm_1\},\end{aligned}$$

$$\begin{aligned} |\hat{\pi}_1[k]-\pi_1^*[k]| &\leq J_1\exp\{-C_0'pm_2\}+\frac{C_1'}{\sqrt{m_1}},\\ |\hat{\pi}_2[l]-\pi_2^*[l]| &\leq J_2\exp\{-C_0'pm_1\}+\frac{C_2'}{\sqrt{m_2}},\end{aligned}$$

and

$$\frac{1}{J_1J_2}\|\hat{\theta}-\theta^*\|_F^2 \leq \frac{C_3'(J_1J_2+m_1\log J_1+m_2\log J_2)}{pm_1m_2} + \max\{J_1,J_2\}\exp\{-C_4'p\min\{m_1,m_2\}\}$$

hold with probability at least $1-\max\{J_1,J_2\}(m_1+m_2)\exp\{-C_6'p(1-p)\min\{m_1,m_2\}\}-\exp\{-C_5'(J_1J_2+m_1\log J_1+m_2\log J_2)\}$.

To ensure the estimation consistency, we need that p should not be too small. From the bounds in Theorem 5, it is required that

$$pm_1m_2/\max\{J_1J_2,m_1\log J_1,m_2\log J_2\}\to\infty$$

and

$$\max\{J_1,J_2\}(m_1+m_2)\exp\{-C_6'p(1-p)\min\{m_1,m_2\}\}\to 0$$

as $m_1,m_2\to\infty$. After simplification, we know that the observation rate p should be at least of order

$$\frac{\max\{\log m_1,\log m_2\}}{\min\{m_1,m_2\}}.$$

Such requirement for p is nearly optimal since

$$p=\Omega(\max\{\frac{\log J_1}{m_2},\frac{\log J_2}{m_1}\})$$

is required for the consistency of maximum likeilhood estimator (Gao et al., 2016).

Connection to the literature. The theoretical properties of maximal likelihood estimator under biclustering setting has been considered by Gao et al., 2016. They proved the minimax estimation rate of MLE under Gaussian biclustering models. The biclustering model is related to stochastic block model. The latter one assumes the symmetric relationship between two modes. Celisse et al., 2012 established classification and estimation consistency for variational inference of SBM, while they do not provide precise error bound of estimators. Biclustering model is also related to graphon model where both observations and latent variables are assumed to be continuous between $[0, 1]$. Theory on graphon estimation includes Airoldi et al., 2013; Olhede and Wolfe, 2014; Choi, 2017; Gao et al., 2015; Klopp et al., 2017 and the references therein. Graphon models are estimated by nonparametric methods instead of using VI. The biclustering model also has connection to the matrix completion problem. The latter one imposes a low rank constraints on parameter Θ as opposed to the latent class structure. The theory for matrix completion problem can be found in Candès and Recht, 2009; Keshavan et al., 2010; Cai et al., 2010; Koltchinskii et al., 2011; Liu and Li, 2016; Chi et al., 2019; Cai and Li, 2020, etc.

When the number of classes are unknown. A nature question is what is the performance of variational estimator when J_1 and J_2 are misspecified. If J_1 or J_2 is over-specified, it is expected that the estimation is still consistent in the sense that some classes may be split into two (or more) sub-classes. If J_1 or J_2 is under-specified, different classes may merge together. The VI estimator should converge to a local stationary point. The precise characterization of such local optimum is of great interest in the future work.

4. Convergence of Variational Algorithm

Although the variational estimator entails nice theoretical properties, a natural question is how to compute the such estimator in practice. In this section, we consider a coordinate ascent variational inference (CAVI) algorithm via alternatively optimizing the model parameters and variational parameters and discuss the convergence issues.

The procedure of estimating the biclustering model is summarized as follows. We use $t \in \{0, \ldots, T\}$ to represent the iteration index and T to represent the total iteration numbers. We first set the initial model parameter as $\theta^{(0)}, \pi_1^{(0)}, \pi_2^{(0)}$ and initialize the variational parameters ϕ_{1i}'s and ϕ_{2j}'s such that $\sum_k \phi_{1i}^{(0)}[k] = 1$ and $\sum_l \phi_{2j}^{(0)}[l] = 1$ for all $i \in [m_1]$, $j \in [m_2]$. Next, we update model parameters θ, π_1, π_2 and variational parameters ϕ_{1i}'s, ϕ_{2j}'s alternatively. That is, we update $\theta^{(t)}$ by maximizing equation (1) with ϕ_{1i}, ϕ_{2j} fixed at $\phi_{1i}^{(t-1)}$ and $\phi_{2j}^{(t-1)}$. When density function $f_\theta(y|z_{1i}, z_{2j})$ belongs to some special distributional families, $\theta^{(t)}$ admits an explicit form. For example,

$$\theta_{kl}^{(t)} = \frac{\sum_{(i,j)\in A} \phi_{1i}^{(t-1)}[k]\phi_{2j}^{(t-1)}[l]y_{ij}}{\sum_{(i,j)\in A} \phi_{1i}^{(t-1)}[k]\phi_{2j}^{(t-1)}[l]}$$

when $f_\theta(y|z_{1i}, z_{2j})$ is the density of Bernoulli, Gaussian or Poisson distribution. We update $\phi_{1i}^{(t)}$ by maximizing equation (1) with model parameters fixed at value of t-th iteration and ϕ_{2j} fixed at $\phi_{2j}^{(t-1)}$. We update $\phi_{2j}^{(t)}$ in the same fashion. We then update $\pi_1^{(t)}$ and $\pi_2^{(t)}$ via using relations (5) and (5). The detailed mathematical formulas are presented in Algorithm 1.

Remark. We want to point out that we are not trying to propose a new algorithm in the current paper. The CAVI-type algorithm is a standard computational scheme in VI optimization problems (Blei et al., 2017). Our goal is to analyze the behavior of CAVI in general biclustering models. In below, we establish the local and global convergence results which may benefit the understanding of landscape of biclustering models.

4.1. Local Convergence

There exist a literature of convergence analysis for EM algorithm under mixture models. Local convergence property of EM are studied by Jin et al., 2016; Yan et al., 2017; Zhao et al., 2020. Global convergence property of EM is considered in Xu et al., 2016. Landscape of stochastic block model is described in Mukherjee et al., 2018. Local convergence theory of CAVI for SBM model has also been studied in Zhang and Zhou, 2020. However, there is no literature on VI algorithm for biclustering models. We fill this gap in the literature. We first study the local convergence of Algorithm 1. In this section, we show that algorithm returns a consistent estimator of model parameter when the initialization is good enough.

Assumption A5: We assume that there exist constants D_1 and D_2 such that $\phi_{1i}^{(0)}$'s satisfy

$$\frac{\sum_{i:z_{1i}^*=k} \phi_{1i}^{(0)}[k]}{\sum_{k'\neq k}\sum_{i:z_{1i}^*=k} \phi_{1i}^{(0)}[k']} > D_1 \tag{7}$$

for all $k \in [J_1]$ and $\phi_{2j}^{(0)}$'s satisfy

$$\frac{\sum_{j:z_{2j}^*=l} \phi_{2j}^{(0)}[l]}{\sum_{l'\neq l}\sum_{j:z_{2j}^*=l} \phi_{2j}^{(0)}[l']} > D_2 \tag{8}$$

for all $l \in [J_2]$.

Here, (7) and (8) guarantee that, for each latent class, initial distribution should put relatively large mass on that true

label. In other words, the initialization of variational distribution should be good enough to concentrate locally around true labels.

Algorithm 1 CAVI for Biclustering Model.

1: **Input.** Observations: $\{y_{ij}\}$
2: **Output.** Estimated parameter: $\hat{\theta}$
3: **Initialization.**
Randomly sample $\theta^{(0)}$ from parameter space $\mathcal{B}$ and choose $\pi_1^{(0)} = (\frac{1}{J_1}, \ldots, \frac{1}{J_1})$ and $\pi_2^{(0)} = (\frac{1}{J_2}, \ldots, \frac{1}{J_2})$.
Sample variational parameter $\phi_{1i}^{(0)}$ independently so that $\sum_{k\in[J_1]} \phi_{1i}^{(0)}[k] = 1$.
Sample $\phi_{2j}^{(0)}$ independently so that $\sum_{l\in[J_2]} \phi_{2j}^{(0)}[l] = 1$.
4: **while** not converged **do**
5: Increase the time index: $t = t + 1$.
6: For each $k \in [J_1]$ and $l \in [J_2]$, update θ_{kl} by

$$\theta_{kl}^{(t)} = \arg\max_{\theta} \sum_{i,j} \phi_{1i}^{(t-1)}[k]\phi_{2j}^{(t-1)}[l] \log f_\theta(y_{ij}|k,l).$$

7: For each $i \in [m_1]$, update ϕ_{1i} by

$$\begin{aligned}\phi_{1i}^{(t)} \;=\; & \arg\max_{\phi} \sum_{k\in[J_1],l\in[J_2]} \sum_{j\in[m_2]} \big\{\phi[k]\phi_{2j}^{(t-1)}[l] \\ & \cdot \log f_{\theta^{(t)}}(y_{ij}|k,l)\big\} - \sum_{k\in[J_1]} \phi[k](\log \phi[k] \\ & - \log \pi_1^{(t-1)}[k]).\end{aligned}$$

8: For each $j \in [m_2]$, update ϕ_{2j} by

$$\begin{aligned}\phi_{2j}^{(t)} \;=\; & \arg\max_{\phi} \sum_{k\in[J_1],l\in[J_2]} \sum_{i\in[m_1]} \big\{\phi[l]\phi_{1i}^{(t)}[k] \\ & \cdot \log F_{\theta^{(t)}}(y_{ij}|k,l)\big\} - \sum_{l\in[J_2]} \phi[l](\log \phi[l] \\ & - \log \pi_2^{(t-1)}[l]).\end{aligned}$$

9: For $k \in [J_1]$, update

$$\pi_1^{(t)}[k] = \frac{1}{m_1} \sum_{i\in[m_1]} \phi_{1i}^{(t)}[k].$$

10: For $l \in [J_2]$, update

$$\pi_2^{(t)}[l] = \frac{1}{m_2} \sum_{j\in[m_2]} \phi_{2j}^{(t)}[l].$$

11: **end while**
12: Set $\hat{\theta} = \theta^{(T_c)}$, where T_c is the time index when the algorithm converges.

Theorem 6 *Under Assumptions A1 - A4, we assume (7) and (8) hold with sufficiently large D_1 and D_2, then Algorithm 1 will return a consistent estimator with probability tending to 1 as $m_1, m_2 \to \infty$.*

Note that VI estimator is biased under the finite sample cases. Here, $m_1, m_2 \to \infty$ means that m_1 and m_2 go to infinity at the same rate. In Algorithm 1, the order of updating model parameter θ and variational parameters ϕ's can be exchanged (i.e., we first implement lines 6-7 and then implement line 5 in each iteration). We can still show the local convergence if we additionally assume that $\theta^{(0)}$ is in the neighborhood of θ^*.

Theorem 7 *Under Assumptions A1 - A4, we assume that (8) holds with sufficiently large D_2 and $\theta^{(0)} \in B(\theta^*, \delta)$ for small radius δ, then Algorithm 1 returns a consistent estimator with probability tending to 1 as $m_1, m_2 \to \infty$.*

4.2. Global Convergence

However, we do not have the knowledge of true model parameters or true latent memberships so that Condition I1 cannot be guaranteed in practice. Can we obtain the global convergence of Algorithm 1. In the following, we give the partial answer to this question.

We first observe that there exist saddle points in optimizing (1). For example, if we let $\phi_{1i}^{(0)} = (\frac{1}{J_1}, \ldots, \frac{1}{J_1})$ and $\phi_{2j}^{(0)} = (\frac{1}{J_2}, \ldots, \frac{1}{J_2})$. Then whatever true parameter is, the algorithm will always return that

$$\theta^{(t)} \equiv \breve{\theta},\ \pi_1^{(t)} \equiv (\frac{1}{J_1}, \ldots, \frac{1}{J_1}),\ \pi_2^{(t)} \equiv (\frac{1}{J_2}, \ldots, \frac{1}{J_2}),$$

where $\breve{\theta}$ is a matrix with all entries equal to $\arg\max_\theta \sum_{i,j} \log f_\theta(y_{ij})$. Hence, we need to add random noise into the initialization of variational parameters to escape from this saddle point.

We consider the following random initialization. Specifically, we sample

$$\phi_{1i}^{(0)} \sim \text{Dir}(\boldsymbol{\alpha}_1) \text{ and } \phi_{2j}^{(0)} \sim \text{Dir}(\boldsymbol{\alpha}_2) \tag{9}$$

for $i \in [m_1]$ and $j \in [m_2]$ independently. Here $\text{Dir}(\boldsymbol{\alpha})$ represents the Dirichlet distribution with parameter $\boldsymbol{\alpha}$; $\boldsymbol{\alpha}_1$ is a vector of length J_1 with all entries being 1 and $\boldsymbol{\alpha}_2$ is a vector of length J_2 with all entries being 1. In other words, $\text{Dir}(\boldsymbol{\alpha}_1)$ and $\text{Dir}(\boldsymbol{\alpha}_2)$ are non-informative priors.

Degenerate Case. We first consider a degenerate situation when $J_2 = 1$. Then the biclustering model reduces to a latent class model. Under this simplified setting, we aim to find out the global convergence properties.

We first additionally assume that $J_1 = 2$. Then we are able to show that the algorithm can always gives a consistent

estimator as long as the model is identifiable (i.e., model satisfies Assumption $A1$, $\theta^*_{11} \neq \theta^*_{21}$).

Theorem 8 *Under Assumptions $A1 - A4$ and the setting that $J_1 = 2$ and $J_2 = 1$, then Algorithm 1 will return a consistent estimator with probability tending to 1 as both $m_1, m_2 \to \infty$ when the initialization satisfies (9).*

Actually, we have the stronger result that the algorithm only fails when ϕ_{1i}'s lie on certain measure zero set. A precise statement is stated as follows.

Theorem 9 *Under Assumptions $A1 - A4$ and the setting that $J_1 = 2$ and $J_2 = 1$, then Algorithm 1 will fail to return a consistent estimator if and only if*

$$\begin{aligned} & \arg\max_{\theta} \sum_{i,j} \phi^{(0)}_{1i}[1] \log f_\theta(y_{ij}) \\ = \quad & \arg\max_{\theta} \sum_{i,j} \phi^{(0)}_{1i}[2] \log f_\theta(y_{ij}). \end{aligned} \tag{10}$$

We can easily see that naive random initialization will make (10) held with probability zero. This explains the usefulness of random initialization. This result is related to EM method in estimating mixture Gaussian models. In Xu et al., 2016, they fully characterize the global convergence of EM algorithm for two equal-proportion Gaussian distributions.

When $J_1 \geq 3$, the story is different. The algorithm might be trapped at local optimizers. When $J_1 = 3$, we can relabel the latent classes such that

$$\theta^*_{11} > \theta^*_{21} > \theta^*_{31}.$$

We define $\bar{\theta}$ as follows,

$$\bar{\theta} := \arg\max_{\theta} \{\sum_{k=1}^{3} \pi_k \mathbb{E}_{y \sim f_{\theta^*_{k1}}} \log f_\theta(y)\},$$

which can be viewed as population mean. We similarly define

$$\begin{aligned} \bar{\theta}_{k_1 k_2} \quad := \quad & \arg\max_{\theta} \{\pi_{k_1} \mathbb{E}_{y \sim f_{\theta^*_{k_1 1}}} \log f_\theta(y) \\ & + \pi_{k_2} \mathbb{E}_{y \sim f_{\theta^*_{k_2 1}}} \log f_\theta(y)\}, \end{aligned}$$

which can be viewed as the population parameter of groups k_1 and k_2. We further define

$$g_k = \nabla \mathbb{E}_{y \sim f_{\theta^*_{k1}}} [\log f_{\bar{\theta}}(y)], \quad k \in \{1, 2, 3\}. \tag{11}$$

The slope g_k quantifies the gap between θ^*_{k1} and $\bar{\theta}$. When $\theta^*_{k1} = \bar{\theta}$, then $g_k = 0$. In addition, we define several variance quantities,

$$\begin{aligned} v_1 = v_2 &= \mathbb{E}[(\phi^{(0)}_{1i}[1] - 1/3)^2], \\ v_{12} &= \mathbb{E}[(\phi^{(0)}_{1i}[1] - 1/3)(\phi^{(0)}_{1i}[2] - 1/3)], \end{aligned}$$

and let $\mathbf{V} = \begin{pmatrix} v_1 & v_{12} \\ v_{12} & v_2 \end{pmatrix}$.

Theorem 10 *Under Assumptions A1 - A4 and the setting with $J_1 = 3$, $J_2 = 1$, $\theta^*_{31} < \theta^*_{21} < \theta^*_{11}$ and*

$$\mathbb{E}_{y \sim f_{\theta^*_{21}}} \log f_{\bar{\theta}_{12}}(y) > \mathbb{E}_{y \sim f_{\theta^*_{21}}} \log f_{\theta^*_{31}}(y), \tag{12}$$

the probability that Algorithm 1 fails to return a consistent estimator is at least $P(g_2 Z_1 + g_3 Z_2 > 0, Z_1 > 0, Z_2 > 0)$ with $(Z_1, Z_2) \sim N(\mathbf{0}, \mathbf{V})$, when both $m_1, m_2 \to \infty$.

Here condition (12) implies that Group 2 is closer to Group 1 rather than Group 3. When $\bar{\theta}$ is close to θ^*_{21}, it is easier for algorithm to find global optimum. By contrast, when $\bar{\theta}$ is close to θ^*_{31}, then Class 3 becomes the dominating group. The algorithm may be stuck at the local optimum, since Class 1 and Class 2 are classified into one group and the dominating Class 3 could be split into two groups. The algorithm can never jump out of this local optimum due to the constraint (12). By Theorem 10, we know that the algorithm may not always converge to the global maximizer and hence will give inconsistent estimator when $J_1 \geq 3$.

Case $J_1 = 2$ and $J_2 = 2$

Theorem 11 *Under the Gaussian/Bernoulli/Poisson model satisfying Assumptions A1 - A3 with $J_1 = J_2 = 2$ and $\pi_1 = \pi_2 = (\frac{1}{2}, \frac{1}{2})$, Algorithm 1 returns a consistent estimator with probability tending to 1 as $m_1, m_2 \to \infty$ when initialization satisfies (9), and*

$$(\theta^*_{11} - \theta^*_{21})(\theta^*_{12} - \theta^*_{22}) > 0, \ (\theta^*_{11} - \theta^*_{12})(\theta^*_{21} - \theta^*_{22}) > 0. \tag{13}$$

Theorem 11 guarantees the global convergence when true parameters are well separated in the sense of (13). For example, consider a Bernoulli model with $\pi_1 = \pi_2 = (0.5, 0.5)$ and

$$\theta = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix},$$

the algorithm will be trapped at local optimum. Condition (13) is only sufficient, it is of interest to obtain the necessary condition under setting of $J_1 = J_2 = 2$ in future work.

Technical challenges. 1) To establish the consistency results, different types of concentration inequalities are needed for variational parameter and model parameters separately. 2) To compute the estimation bound under the setting of general response distribution functions, the calculation is more involved and tedious. 3) To study the global convergence of CAVI, we need carefully calculating the differences between class-specific model parameters. The computation is overwhelming.

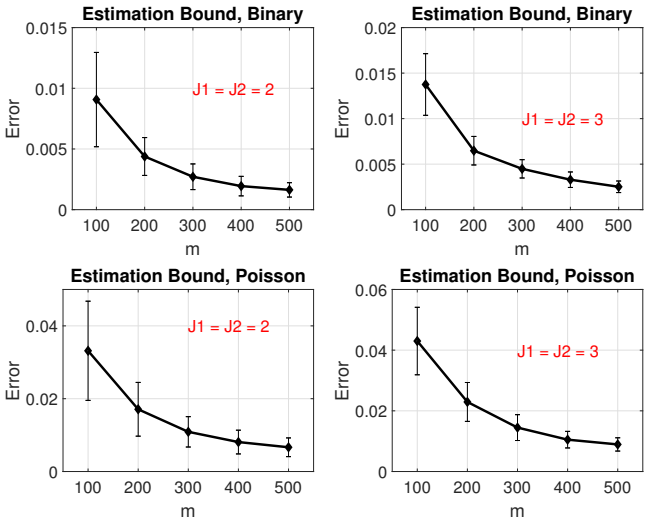

Figure 1. Plots of estimation errors under different settings. Upper two plots are for Bernoulli models. Bottom two plots are for Poisson models. The standard error bars are also plotted.

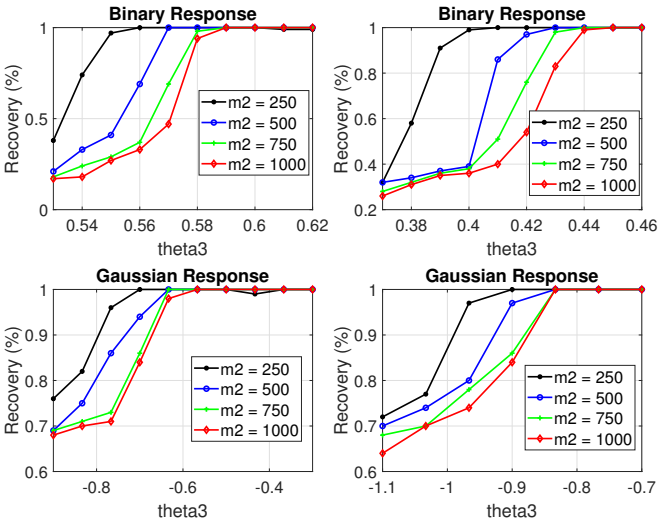

Figure 2. Probability of global convergence for Bernoulli and Poisson models under different settings. The parameter choice is specified as follows. Upper left: $\pi_1 = (0.1, 0.2, 0.7)$, $\theta_{11} = 0.9$, $\theta_{21} = 0.7$. Upper right: $\pi_1 = (0.1, 0.2, 0.7)$, $\theta_{11} = 0.9$, $\theta_{21} = 0.6$. Bottom left: $\pi_1 = (0.2, 0.2, 0.6)$, $\theta_{11} = 1, \theta_{21} = 0$. Bottom right: $\pi_1 = (0.3, 0.1, 0.6)$, $\theta_{11} = 1, \theta_{21} = 0$. For all four cases, $m_1 = 500$ and $m_2 \in \{250, 500, 750, 1000\}$.

5. Numerical Results

In this section, we collect several numerical experiments to support our theory, i.e., validating the tightness of estimation bound and global convergence of CAVI algorithm.

Estimation bound. We consider the Bernoulli biclustering model (i.e., $Y_{ij} \sim \text{Bernoulli}(\theta_{z_i z_j})$) and Poisson biclustering model (i.e., $Y_{ij} \sim \text{Poisson}(\theta_{z_i z_j})$). We set sample size $m_1 = m_2 = m$, where m take values from $\{100, 200, 300, 400, 500\}$. We set number of classes $J_1 = J_2 = J = 2$ or 3. True parameter θ is randomly generated and π_1, π_2 are set to be uniform prior. For each setting, we run 100 replications. Estimation errors ($\frac{1}{\sqrt{J_1 J_2}}\|\hat{\theta} - \theta^*\|_F$) with corresponding standard errors are shown in Figure 1. Based on curves, we can see that the estimation error decreases as sample sizes increase. When the number of classes increases, the error will become larger. In addition, when J is fixed, the estimation error decays at rate of $\frac{1}{\sqrt{m}}$ with variational approximation error vanishing (less than 1e-9). This matches the results stated in Theorem 3.

Global convergence. We take $J_1 = 3, J_2 = 1$ and consider mixture Bernoulli model and mixture Gaussian model. We fix parameters θ_{11}, θ_{21}, population prior π_1 and but let θ_{31} vary. We want to study the probability of global convergence as θ_{31} changes. The detailed settings and corresponding results under different sample sizes can be found in caption of Figure 2. Each setting is replicated for 100 times.

It is straightforward to see that group 3 is the dominating class (i.e., $\pi_1[3]$ has the largest value). From Figure 2, when θ_{31} is much smaller than θ_{11}, θ_{21}, then it becomes harder to find the global optimum. There exists a non-zero probability of failure in recovering the true parameter. As θ_{31} increases, the recovery probability increases up to 1. This phenomenon matches our theory.

6. Conclusion

In this paper, we develop a theory of variational inference estimation in biclustering models. Our result is general in the sense that we do not assume any specific response functions. The assumptions are mild and they are satisfied by most probabilistic models, including but not limited to SBM model, mixture Gaussian model and mxiture Poisson model. We establish the classification consistency, population consistency and parameter consistency. Both upper and lower bounds of estimation errors are also obtained. Our theory answers the question why variational inference works well for biclustering problem. In addition, we consider a coordinate ascent variational inference (CAVI) algorithm for parameter estimation. The algorithm is shown to have local convergence property under a reasonable initialization requirement. On the other hand, with the random initialization, global convergence results are also established under several important special settings. This work not only bridges the gap in the literature around VI theory and but also gives a deeper understanding of the landscape of biclustering models.

Acknowledgement

The authors sincerely thank the anonymous reviewers and area chair for their constructive comments.

References

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

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.

Edoardo M. Airoldi, Thiago B. Costa, and Stanley H. Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 692–700, Lake Tahoe, NV, 2013.

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.

Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

Yunfeng Cai and Ping Li. Solving the robust matrix completion problem via a system of nonlinear equations. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 4162–4172, Online [Palermo, Sicily, Italy], 2020.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

Alain Celisse, Jean-Jacques Daudin, and Laurent Pierre. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6:1847–1899, 2012.

Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67 (20):5239–5269, 2019.

David Choi. Co-clustering of nonsmooth graphons. *Annals of Statistics*, 45(4):1488–1515, 2017.

Sara Dolnicar, Sebastian Kaiser, Katie Lazarevski, and Friedrich Leisch. Biclustering: Overcoming data dimensionality problems in market segmentation. *Journal of Travel Research*, 51(1):41–49, 2012.

Chao Gao, Yu Lu, and Harrison H Zhou. Rate-optimal graphon estimation. *Annals of Statistics*, 43(6):2624–2652, 2015.

Chao Gao, Yu Lu, Zongming Ma, and Harrison H Zhou. Optimal estimation and completion of matrices with biclustering structures. *Journal of Machine Learning Research*, 17(1):5602–5630, 2016.

Chao Gao, Zongming Ma, Anderson Y Zhang, and Harrison H Zhou. Achieving optimal misclassification proportion in stochastic block models. *Journal of Machine Learning Research*, 18(1):1980–2024, 2017.

Gérard Govaert and Mohamed Nadif. Latent block model for contingency table. *Communications in Statistics—Theory and Methods*, 39(3):416–425, 2010.

Jiajun Gu and Jun S Liu. Bayesian biclustering of gene expression data. *BMC Genomics*, 9(1):1–10, 2008.

Yue Guan, Jennifer G Dy, Donglin Niu, and Zoubin Ghahramani. Variational inference for nonparametric multiple clustering. In *MultiClust Workshop, KDD-2010*. Citeseer, 2010.

John A Hartigan. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):123–129, 1972.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(5), 2013.

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

Chi Jin, Yuchen Zhang, Sivaraman Balakrishnan, Martin J. Wainwright, and Michael I. Jordan. Local maxima in the likelihood of gaussian mixture models: Structural results and algorithmic consequences. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4116–4124, Barcelona, Spain, 2016.

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

Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

Olga Klopp, Alexandre B Tsybakov, and Nicolas Verzelen. Oracle inequalities for network models and sparse graphon estimation. *Annals of Statistics*, 45(1):316–354, 2017.

Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates

for noisy low-rank matrix completion. *Annals of Statistics*, 39(5):2302–2329, 2011.

Guangcan Liu and Ping Li. Low-rank matrix completion in the presence of high coherence. *IEEE Trans. Signal Process.*, 64(21):5623–5633, 2016.

Eleni Matechou, Ivy Liu, Daniel Fernández, Miguel Farias, and Bergljot Gjelsvik. Biclustering models for two-mode ordinal data. *Psychometrika*, 81(3):611–624, 2016.

Vichi Maurizio. Double k-means clustering for simultaneous classification of objects and variables. In *Advances in classification and data analysis*, pages 43–52. Springer, 2001.

Ted Meeds and Sam Roweis. Nonparametric bayesian biclustering. Technical Report UTML TR 2007–001, January 2007.

Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for binary symmetric block models. *arXiv preprint arXiv:1407.1591*, 3(5), 2014.

Soumendu Sundar Mukherjee, Purnamrita Sarkar, Y. X. Rachel Wang, and Bowei Yan. Mean field for the stochastic blockmodel: Optimization landscape and convergence issues. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10717–10727, Montréal, Canada, 2018.

Donglin Niu, Jennifer G. Dy, and Zoubin Ghahramani. A nonparametric bayesian model for multiple clustering with overlapping feature views. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 814–822, La Palma, Canary Islands, Spain, 2012.

Sofia C Olhede and Patrick J Wolfe. Network histograms and universality of blockmodel approximation. *Proceedings of the National Academy of Sciences*, 111(41):14722–14727, 2014.

Shirley Pledger and Richard Arnold. Multivariate methods using mixtures: Correspondence analysis, scaling and pattern-detection. *Computational Statistics & Data Analysis*, 71:241–261, 2014.

Amela Prelić, Stefan Bleuler, Philip Zimmermann, Anja Wille, Peter Bühlmann, Wilhelm Gruissem, Lars Hennig, Lothar Thiele, and Eckart Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129, 2006.

Jonathan Templin and Robert A Henson. *Diagnostic measurement: Theory, methods, and applications*. Guilford Press, 2010.

Duy Vu and Murray Aitkin. Variational algorithms for biclustering models. *Computational Statistics & Data Analysis*, 89:12–24, 2015.

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 (NIPS)*, pages 2676–2684, Barcelona, Spain, 2016.

Bowei Yan, Mingzhang Yin, and Purnamrita Sarkar. Convergence of gradient EM on multi-component mixture of gaussians. In *Advances in Neural Information Processing Systems (NIPS)*, pages 6956–6966, Long Beach, CA, 2017.

Anderson Y Zhang and Harrison H Zhou. Theoretical and computational guarantees of mean field variational inference for community detection. *Annals of Statistics*, 48 (5):2575–2598, 2020.

Ruofei Zhao, Yuanzhi Li, and Yuekai Sun. Statistical convergence of the em algorithm on gaussian mixture models. *Electronic Journal of Statistics*, 14(1):632–660, 2020.

Zhixin Zhou and Ping Li. Rate optimal chernoff bound and application to community detection in the stochastic block models. *Electronic Journal of Statistics*, 14(1): 1302–1347, 2020.