A Theoretical Case Study of Structured Variational Inference for Community Detection

Mingzhang Yin[⋆], Y. X. Rachel Wang[†], Purnamrita Sarkar[⋆]

The University of Texas at Austin[⋆]
University of Sydney[†]

AISTATS 2020

Stochastic block model (SBM)

SBM(n, K, π, B) is a generative latent variable model

- Each node $i \in [n] := \{1, \cdots, n\}$ has a cluster label $h_i \in [K]$
- Probability that a node is in cluster $K = \pi_k$
- Probability that nodes i, j are connected $= B_{h_i, h_j}$
- Observation: adjacency matrix $A_{n \times n}$, latent variables: $\boldsymbol{h}_{n \times 1}$

In this paper, we theoretically study the convergence properties of structured variational inference (VI) on SBM

Variational inference

Basic idea: approximate the intractable $P(\boldsymbol{h}|X)$ with variational distribution $Q(\boldsymbol{h})$ by optimization

- The maginal likelihood (evidence) can be decomposed as

$$
\begin{aligned}
\log P(X) &= \int Q(\boldsymbol{h}) \log P(X) d\boldsymbol{h} \\
&= \int Q(\boldsymbol{h}) \log \frac{P(X, \boldsymbol{h})}{Q(\boldsymbol{h})} d\boldsymbol{h} + \int Q(\boldsymbol{h}) \log \frac{Q(\boldsymbol{h})}{P(\boldsymbol{h} \mid X)} d\boldsymbol{h} \\
&= \text{ELBO} + \mathcal{D}_{\text{KL}}(Q(\boldsymbol{h}) || P(\boldsymbol{h} \mid X))
\end{aligned}
$$

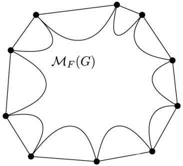
- For inference problem, $P(X)$ is considered as a constant, so

$$
\min_{Q} \mathcal{D}_{\text{KL}}(Q(\boldsymbol{h}) || P(\boldsymbol{h} \mid X)) \Leftrightarrow \max_{Q} \text{ELBO}
$$

Mean-field variational inference

- Mean-field variational inference (MFVI) assumes factorized Q distribution of $\boldsymbol{h} = (h_1, \ldots, h_n)^T$

$$Q(\boldsymbol{h}) = \prod_{i=1}^{n} q_{\theta_i}(h_i)$$



Wainwright & Jordan, 2008

- The factorized assumption allows closed-form coordinate ascent

- However, mean-field approximations makes the nonconvexity as an intrinsic property
  - $\rightarrow$ multiple local optima
  - $\rightarrow$ sensitivity to initialization

Motivations

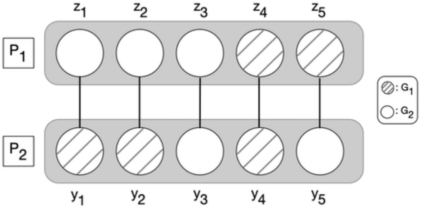
A gap between what are used in practice and what is known in theory for VI:

- On the one end, theoretical explanation of the success of “modern" VI with a variety of dependence structure is an open problem
- For SBM, the full theory is lacking for methods that model the node dependency, such as the belief propagation (BP)
- On the other hand, MFVI’s behavior has been well understood in SBM with two equal size clusters

Question: Theoretically, can additional dependence structure improve the VI objective landscape?
Approach: A case study by constructing VI with pairwise structure (VIPS)

Pairwise dependence structure [1]

- The n nodes are ramdomly partitioned to two sets: $P_1 = \{z_1, \cdots, z_{n/2}\}$, $P_2 = \{y_1, \cdots, y_{n/2}\}$
- Nodes z_i in P_1 are paired with nodes y_i in P_2
- VIPS: $q_\phi(\boldsymbol{h}) = \prod_{i=1}^{n/2} \text{Categorical}((h_{z_i}, h_{y_i}); \boldsymbol{\psi}_i)$ $\boldsymbol{\psi}_i = \sigma(\boldsymbol{\theta}_i)$ with softmax link function, logits $\boldsymbol{\theta}_i = (\theta_i^{00}, \theta_i^{01}, \theta_i^{10}, \theta_i^{11})$; $\boldsymbol{u} \in \mathbb{R}^n$ is MLE as the estimated membership vector
- MFVI: Variational distribution is a product of independent Bernoulli distributions



An illustration of a random pairwise partition, $n = 10$.

[1] We do not aim to design state-of-art method; rather we keep the dependence structure simple so the theoretical analysis is clear.

Case 1: Known model parameters

We update parameters iteratively

$$\text{i-th meta iteration:}\quad \boldsymbol{\theta}^{10} \to \boldsymbol{u}_1^{(i)} \to \boldsymbol{\theta}^{01} \to \boldsymbol{u}_2^{(i)} \to \boldsymbol{\theta}^{11} \to \boldsymbol{u}_3^{(i)} \to \boldsymbol{\theta}^{10} \cdots$$

Theorem (Sample behavior for known parameters)

Assume $\boldsymbol{\theta}$ are initialized as $\mathbf{0}$ and the elements of $\boldsymbol{u}$ are initialized i.i.d from Bernoulli(0.5). When $p \asymp q \asymp \rho_n$, $p - q = \Omega(\rho_n)$, and $\sqrt{n}\rho_n = \Omega(\log(n))$, VIPS converges to the true labels asymptotically, in the sense that

$$\|\boldsymbol{u}_3^{(2)} - \boldsymbol{h}^*\|_1 = n\exp(-\Omega_P(n\rho_n))$$

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

Corollary: When $\boldsymbol{u}$ is initialized from a distribution with mean $\mu \neq 0.5$, $\|\boldsymbol{u}_3^{(3)} - \boldsymbol{h}^*\|_1 = n\exp(-\Omega_P(n\rho_n))$

Proof Sketch

- The proof hinges on SVD of $P = \mathbb{E}[A|U] = \frac{p+q}{2}\mathbf{1}_n\mathbf{1}_n^T + \frac{p-q}{2}v_2v_2^T - pI$, where $\|\boldsymbol{u} - \boldsymbol{h}^*\|_1 = n/2 - |\langle \boldsymbol{u}, v_2\rangle|$; we show signal $|\langle \boldsymbol{u}, v_2\rangle|$ increases at each iteration
- We use Littlewood-Offord type anti-concentration to ensure the signal is not too small

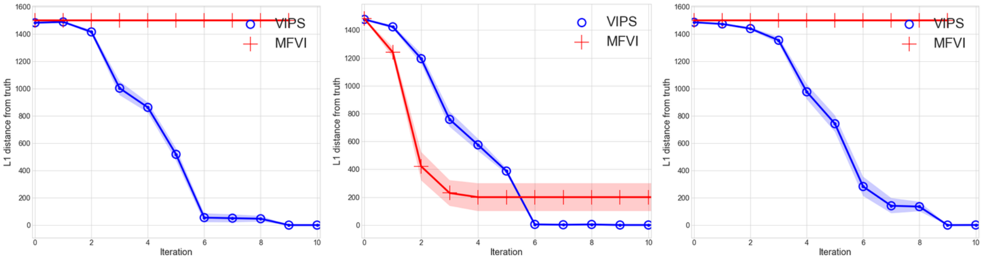
  We use a Berry-Esseen bound and a uniform bound based on Hoeffding inequality to handle the noise
- We show in the first three iterations (first meta-iteration)

$$\langle u_1^{(1)}, v_2\rangle = \Omega_P(n\sqrt{\rho_n})$$
$$\langle u_2^{(1)}, v_2\rangle \geq \frac{n}{8} - o_P(n)$$
$$\langle u_3^{(1)}, v_2\rangle \geq \frac{n}{4} - o_P(n);$$

  after the second meta-iteration

$$\langle u_3^{(2)}, v_2\rangle \geq \frac{n}{2} - n\exp(-\Omega_P(n\rho_n))$$

Simulation 1



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.

Case 2: Estimated, fixed model parameters

Proposition (Parameter robustness)

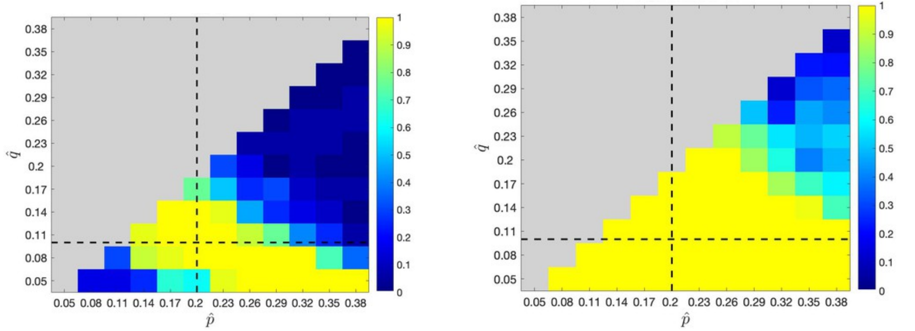
If we replace true p, q with some estimation $\hat{p}, \hat{q}$, we have

$$\|\boldsymbol{u}_3^{(2)} - \boldsymbol{h}^*\|_1 = n \exp(-\Omega_P(n\rho_n))$$

if i) $\frac{p+q}{2} > \hat{\lambda}$, *ii)* $\hat{\lambda} - q = \Omega(\rho_n)$, *iii)* $\hat{t} = \Omega(1)$, *where* $\hat{t} = \frac{1}{2}\log\frac{\hat{p}/(1-\hat{p})}{\hat{q}/(1-\hat{q})}$, $\hat{\lambda} = \frac{1}{2\hat{t}}\log\frac{1-\hat{q}}{1-\hat{p}}$.

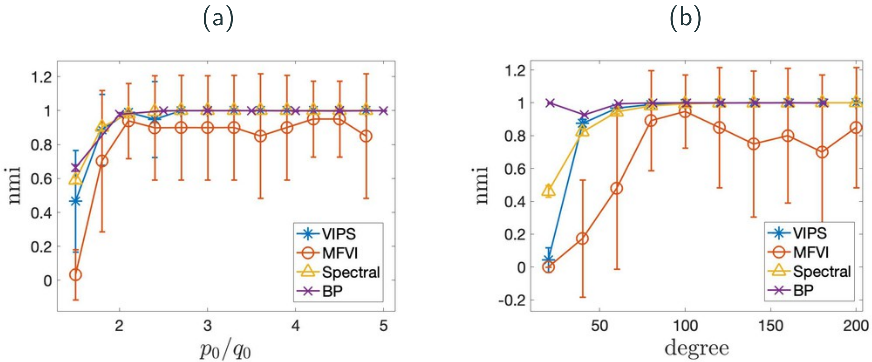
Simulation 2-1

(a) MFVI (b) VIPS



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.

Simulation 2-2



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 $n = 2000$.

Case 3: Updating model parameters

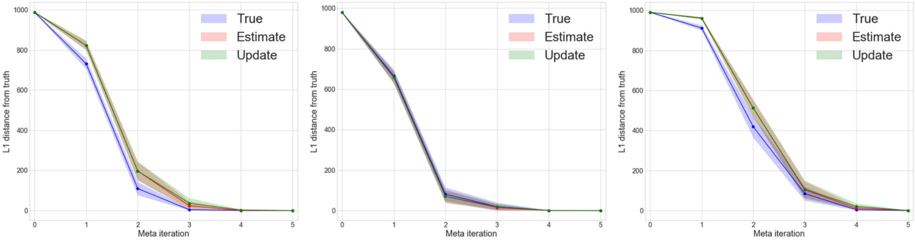
Theorem (Updating parameters and u simultaneously)

Suppose we initialize with some estimates of true (p, q) as $\hat{p} = p^{(0)}$, $\hat{q} = q^{(0)}$ satisfying the conditions in Proposition (Parameter robustness) and apply two meta iterations 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.

Simulation 3



Values of $\|u - z^*\|_1$ as the number of meta iterations increases. Each line is the mean curve of 50 random trials and the shaded area is the standard deviation. Here $n = 2000$ and $p_0 = 0.1, q_0 = 0.02$. u is initialized by Bernoulli distribution with mean $\mu = 0.1, 0.5, 0.9$ from the left to right.

Comparison to MFVI

<table>
<tr><th>MFVI [2]</th><th>VIPS</th></tr>
<tr><td>For unknown model parameters, MFVI with random initializations converges to the uninformative stationary points with non-negligible probability</td><td rowspan="3">Converges to the true membership vector with probability approaching 1</td></tr>
<tr><td>When the initialization is not centered at 0.5, MFVI converges to $\mathbf{0}_n$ or $\mathbf{1}_n$</td></tr>
<tr><td>When updating model parameters, MFVI with a random initialization converges to $\frac{1}{2}\mathbf{1}_n$</td></tr>
<tr><td>Less robust to mis-specified model parameters</td><td>More robust to mis-specified model parameters</td></tr>
</table>

[2] MFVI results are shown in (Mukherjee et al.,2018, Sarkar et al. 2019)

Future directions

- Study VIPS on SBM with multiple, unbalanced clusters
- Use similar methods to study the algorithms such as belief propagation on SBM
- Theoretically study structured VI with more general dependence structures and probabilistic models

  ......

Thank you!