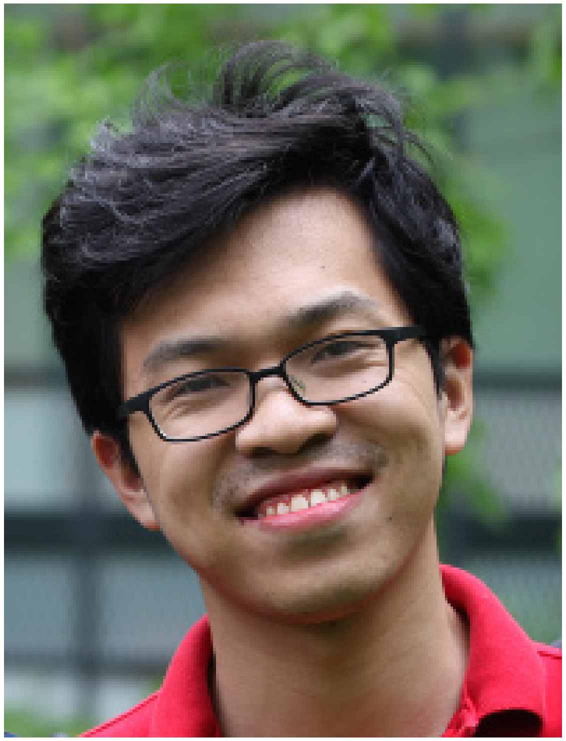
Three Siblings:
EM, Variational Inference, and Gibbs Sampling

Harrison H. Zhang
Department of Statistics and Data Science
Yale University



Anderson Y. Zhang

Outline

- **Stochastic Block Model**
- **EM, Variational Inference, and Gibbs Sampling**
- **Computational and Statistical Guarantees**
- **Questions and Extensions**

Stochastic Block Model

Model: Two Communities

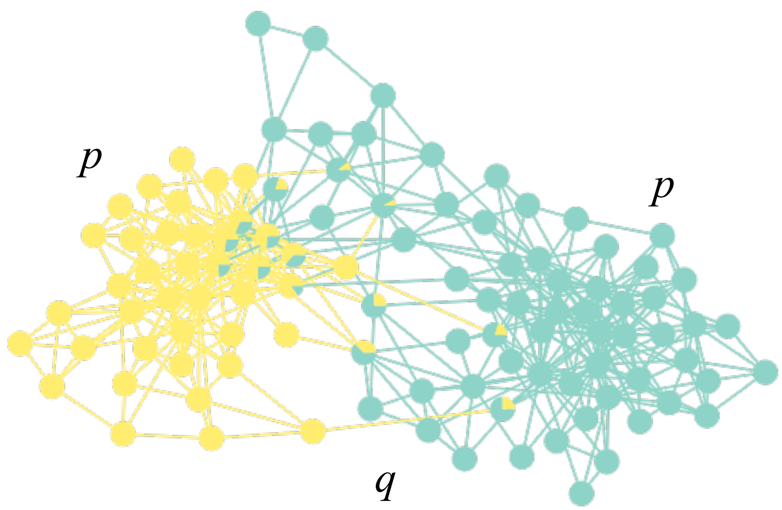
Partition

$$z : \{1, 2, ..., n\} \rightarrow \{0, 1\}^n.$$

Observation: Adjacency matrix A

$$A_{ij} \sim \text{Bernoulli}(P_{ij}), \text{ for } i > j,$$

where $P_{ij} = p$ when $z_i = z_j$, and $P_{ij} = q$ when $z_i \neq z_j$.



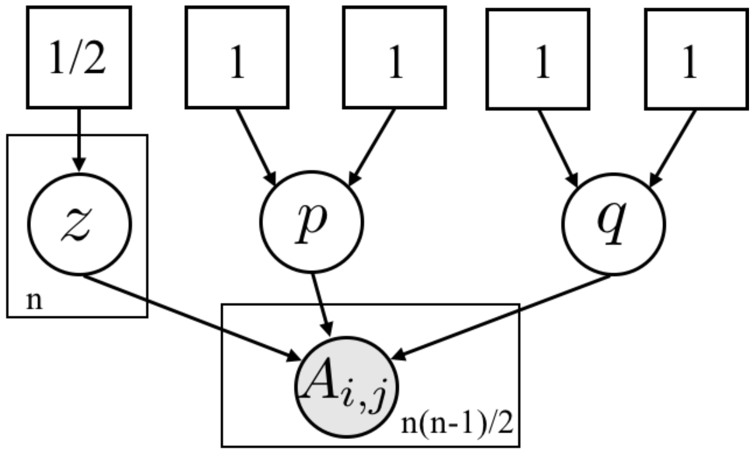
Goal: Recover z from A.

Bayesian Inference

Prior: $z_i \sim \text{Ber}(1/2), \forall i \in [n]$ and $p, q \sim \text{Beta}(1,1)$ independently.

Challenge: The posterior is computationally intractable:

$$\mathbf{p}(z,p,q|A) = \frac{\mathbf{p}(z,p,q,A)}{\sum_{z\in\{0,1\}^n} \int_{p,q} \mathbf{p}(z,p,q,A)}.$$



Three Siblings:
EM, Variational Inference, and Gibbs Sampling

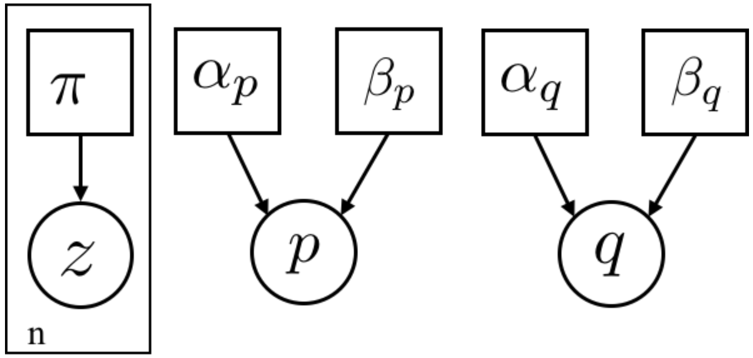
Mean Field Variational Inference

Main Idea: Approximate by a product measure $\mathbf{q}(z,p,q) = \prod_i \mathbf{q}_i(z_i)\mathbf{q}(p)\mathbf{q}(q)$ where $z_i \sim \mathrm{Ber}(\pi_i), \forall i \in [n], p \sim \mathrm{Beta}(\alpha_p, \beta_p), q \sim \mathrm{Beta}(\alpha_q, \beta_q)$ and the Kullback-Leibler divergence is minimized:

$$\hat{\mathbf{q}} = \arg\min_{\mathbf{q}\in\mathbf{Q}} \mathrm{KL}\Big[\mathbf{q}(z,p,q)\Big\|\mathbf{p}(z,p,q|A)\Big].$$

Iterative Algorithm: Approach $\hat{\mathbf{q}}$ by an iterative algorithm:

$$\{\pi^{(s)}, \alpha_p^{(s)}, \beta_p^{(s)}, \alpha_q^{(s)}, \beta_q^{(s)}\}_{s\geq 0}.$$



Mean Field Variational Inference

General Setting

- approximate $\mathbf{p}(\theta|\mathbf{x})$ by some product measure $\mathbf{q}(\theta) = \prod \mathbf{q}_i(\theta_i)$
- minimize the Kullback-Leibler divergence

$$\hat{\mathbf{q}}^{\mathrm{MF}} = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}\left[\mathbf{q}(\theta) \middle\| \mathbf{p}(\theta|\mathbf{x})\right]$$

- Iterative Algorithm:

$$\hat{\mathbf{q}}_i = \arg\min_{\mathbf{q}_i} \mathrm{KL}\left[\mathbf{q}_i(\theta_i) \times \prod_{j \neq i} \mathbf{q}_j(\theta_j) \middle\| p(\theta|\mathbf{x})\right], \forall i \in [n]$$

$$\iff \quad \hat{\mathbf{q}}_i(\theta_i) \propto \exp\{\mathbb{E}_{\mathbf{q}_{-i}}[\log \mathbf{p}(\theta_i|\theta_{-i}, \mathbf{x})]\}, \forall i \in [n].$$

Iterative Algorithm

- Updates on $\alpha_p^{(s+1)}, \beta_p^{(s+1)}, \alpha_q^{(s+1)}, \beta_q^{(s+1)}$:

$$\alpha_p^{(s+1)} = 1 + \sum_{a=1}^{k} \sum_{i<j} A_{i,j} \pi_{i,a}^{(s)} \pi_{j,a}^{(s)}, \quad \beta_p^{(s+1)} = 1 + \sum_{a=1}^{k} \sum_{i<j} (1 - A_{i,j}) \pi_{i,a}^{(s)} \pi_{j,a}^{(s)},$$

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

- Updates on $\{\pi_i^{(s+1)}\}$:

$$\pi_i^{(s+1)} = \frac{\exp\left(2t^{(s)} \sum_{j \neq i} \pi_j^{(s)} (A_{i,j} - \lambda^{(s)})\right)}{\exp\left(2t^{(s)} \sum_{j \neq i} \pi_j^{(s)} (A_{i,j} - \lambda^{(s)})\right) + \exp\left(2t^{(s)} \sum_{j \neq i} (1 - \pi_j^{(s)}) (A_{i,j} - \lambda^{(s)})\right)},$$

where $t^{(s)}, \lambda^{(s)}$ are functions of $\alpha_p^{(s)}, \beta_p^{(s)}, \alpha_q^{(s)}, \beta_q^{(s)}$.

Three Siblings

Similarity: Coordinate updates with $\mathbf{p}(\theta_i|\theta_{-i}, \mathbf{x})$.

Expectation Maximization	Mean Field Variational Inference	Gibbs Sampling
For *local* variables (e.g., $\{z_i\}_{i=1}^n$) $\exp\{\mathbb{E}_{\mathbf{q}_{-i}}[\log \mathbf{p}(\theta_i\|\theta_{-i}, \mathbf{x})]\}$ For *global* variables (e.g., p, q) $\arg\max_{\theta_i} \mathbb{E}_{\mathbf{q}_{-i}}[\log \mathbf{p}(\theta_i\|\theta_{-i}, \mathbf{x})]$	For all $i \in [n]$, $\exp\{\mathbb{E}_{\mathbf{q}_{-i}}[\log \mathbf{p}(\theta_i\|\theta_{-i}, \mathbf{x})]\}$	$\forall i \in [n]$, sample from $\mathbf{p}(\theta_i\|\theta_{-i}, \mathbf{x})$

Statistical and Computational Guarantees

Loss Function

- Let z^*, p^*, q^* be the underlying ground truth from which A is generated.
- We consider the distance between z^* and $\pi^{(s)}$.

Loss Function

$$
\begin{aligned}
\ell(\pi, z^*) &= \frac{1}{n} \min \left\{ \|\pi - z^*\|_1 , \|\pi - (1_n - z^*)\|_1 \right\} \\
&= \frac{1}{n} \min \left\{ \sum_{i=1}^{n} |\pi_i - z_i^*| , \sum_{i=1}^{n} |\pi_i - (1 - z_i^*)| \right\}
\end{aligned}
$$

Theoretical Guarantees for Mean Field

Theorem (Zhang & Z. 2017)**.** *Let* $I = (\sqrt{p^*} - \sqrt{q^*})^2$ *[signal-to-noise ratio]. Assume the initializer* $\pi^{(0)}$ *satisfies* $\ell(\pi^{(0)}, z^*) \leq c$ *for some small constant* c *with high probability. Further we assume the following equation holds*

$$nI \to \infty.$$

Then in each iteration of the iterative algorithm for mean field, with high probability

$$\ell(\pi^{(s+1)}, z^*) \leq \exp\left(-(1 - o(1))nI/2\right) + \frac{1}{\sqrt{nI}}\ell(\pi^{(s)}, z^*).$$

Corollary 1 (Zhang & Z. 2017)**.** *For $s \geq \log n$, we have with high probability*

$$\ell(\hat{\pi}^{(s)}, z^*) \leq \exp(-(1-o(1))nI/2)) + o(n^{-D}), k = 2$$

for every $D > 1$.

Remark: Rate-optimal. Matches with the minimax rate in Zhang and Z. (2016, AoS).

Theoretical Guarantees for Gibbs Sampling

Theorem (Zhang & Z. 2017)**.** *Let* $I = (\sqrt{p^*} - \sqrt{q^*})^2$ *[signal-to-noise ratio]. Assume the initializer* $z^{(0)}$ *satisfies* $\ell(z^{(0)}, z^*) \le c$ *for some small constant* c *with high probability. Further we assume the following equation holds*

$$nI \to \infty.$$

Then in each iteration of the batched Gibbs sampling, with high probability

$$\mathbb{E}\left[\ell(z^{(s)}, z^*)|A, z^{(0)}\right] \le \exp\left(-(1-o(1))nI/2\right) + \left[\frac{1}{\sqrt{nI}}\right]^s \ell(z^{(0)}, z^*) + sb_n,$$

where $b_n = o(\exp\left(-n^{3/2}I\right))$.

Questions and Extensions

Questions

- **Initializer:** What is the sharp condition for initializer?(c.f., analysis on Lloyd's Algorithm in Lu & Z. 2016 to have $c = 1/2 - o(1)$)

- **Parameter Estimation:** Can we prove that $\{p^{(s)}\}_{s\geq 1}$ and $\{q^{(s)}\}_{s\geq 1}$ converge to the optimal rates?

Extensions

- **Models:** Computational and statistical guarantees of variational inference for a general class of latent variable models.

- **Algorithms:** A unified framework to understand iterative algorithm.