Mean Field for the Stochastic Blockmodel: Optimization Landscape and Convergence Issues

Soumendu Sundar Mukherjee, Purnamrita Sarkar, Y. X. Rachel Wang, and Bowei Yan

Indian Statistical Institute, Kolkata; University of Texas, Austin; University of Sydney; and University of Texas, Austin

Stochastic Blockmodel

- K-block Stochastic Blockmodel (SBM) on n nodes (Holland et al., 1983)
  - Community labels: $n \times K$ membership matrix Z , $Z_{i\cdot}$ is the community membership vector of node i and has a $\texttt{Multinomial}(1;\pi)$ distribution, independently of the other rows.
  - Adjacency matrix $A \in \{0,1\}^{n\times n}$,
    $A_{ij}|(Z_{ia}=1, Z_{jb}=1) \sim \text{Bernoulli}(B_{ab}), \qquad i \neq j, \quad A_{ij} = A_{ji}$
- Estimate both Z and the parameters π_a, B_{ab}, $1 \le a, b \le K$.

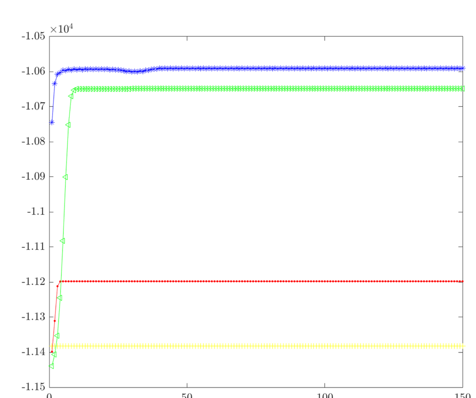
Mean field approximation

$$\log P(A; B, \pi) \overset{\text{(Jensen)}}{\geq} \sum_Z \log\left(\frac{P(A, Z; B, \pi)}{\psi(Z)}\right)\psi(Z) \quad \forall\psi \text{ prob. on } \mathcal{Z}.$$

- Equality holds for $\psi^*(Z) = P(Z|A; B, \pi)$.
- Mean field approximation with $\Psi_{MF} \equiv \{\psi : \psi(z_1, \ldots, z_n) = \prod_{j=1}^n \psi_j(z_j)\}$.

$$\ell_{MF}(\psi, B, \pi) = \sum_{i<j,a,b} \psi_{ia}\psi_{jb}(A_{ij}\log B_{ab} + (1 - A_{ij})\log(1 - B_{ab})) - \texttt{KL}(\psi||\pi^{\otimes n})$$

- Coordinate ascent, alternate between maximizing $\ell_{MF}(\psi, B, \pi)$ for MF parameters and model parameters
- **Pros:** Computationally fast, can easily be modified to allow more complex models.
- **Cons:** Suffers from many local optima.



- $K = 3,\ B = 0.5 \cdot \begin{bmatrix} 1 & 0.4 & 0.1 \\ 0.4 & 1 & 0.1 \\ 0.1 & 0.1 & 1 \end{bmatrix}$,
  $\pi = (1/3, 1/3, 1/3),\ n = 600.$
- truth; $\begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$; $\begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$; (1/3,1/3,1/3)

Related work

- SBM (Celisse et al. 2012, Bickel et al. 2013)
  - Reparametrize $B_{ab} = \rho_n S_{ab}$, $\rho_n \to 0$. $n\rho_n$ is roughly the average degree.
  - In the semi-dense regime $\rho_n n/\log n \to \infty$, closeness of maximum likelihood and maximum variational likelihood
  - Asymptotic equivalence of variational estimates and MLE
- Positive result from Zhang and Zhou 2017
  - Batch coordinate ascent updates (BCAVI), alternate between updating all ψ and the model parameters
  - When the initialization is sufficiently close to the truth, $\ell(\psi^{s+1}, Z) \le$ minimax error $+\, c_n\ell(\psi^s, Z)$, $c_n = o(1)$
- This paper: a more complete characterization for simple setting.
  - $K = 2$, $\pi = 1/2$, $B_{11} = B_{22} = p$, $B_{12} = B_{21} = q$, $p > q$. $p \asymp q \asymp \rho_n$, $\rho_n \to 0$ at some rate.

BCAVI updates for $K = 2$

- Given $\psi^{(s-1)} \in [0,1]^n$, update $p^{(s)}$ and $q^{(s)}$ by averaging the entries of A using the soft membership vector $\psi^{(s-1)}$.
- Given $p^{(s)}$, $q^{(s)}$, update $\psi^{(s)}$,

$$\xi_i^{(s+1)} := \log\frac{\psi_i^{(s+1)}}{1 - \psi_i^{(s+1)}} = 4t^{(s)}\sum_{j\neq i}(\psi_j^{(s)} - \frac{1}{2})(A_{ij} - \lambda^{(s)}),$$
$$\psi_i^{(s+1)} = g(\xi_i^{(s+1)}), \qquad g \text{ is the sigmoid function},$$

where $t^{(s)} = \frac{1}{2}\log\left(\frac{p^{(s)}(1-q^{(s)})}{q^{(s)}(1-p^{(s)})}\right)$, $\lambda^{(s)} = \frac{1}{2t^{(s)}}\log\left(\frac{1-q^{(s)}}{1-p^{(s)}}\right)$.

Let p^*, q^* (corresponding λ^*, t^*) be the true parameters. Our analysis has two parts:

- Knowing p^*, q^*, updating ψ alone.
- Full updates with unknown p^*, q^*.

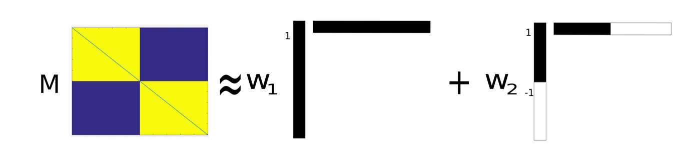
Known p^*, q^*

- Let $\mathbb{E}(A|Z) = ZBZ^\top - p^*I =: P - p^*I$, $M = P - p^*I - \lambda^*(J - I)$. Key decomposition:

$$\xi^{(s)} = 4t(A - \lambda(J - I))(\psi^{(s-1)} - \frac{1}{2}\mathbf{1})$$
$$= \underbrace{4tM(\psi^{(s-1)} - \frac{1}{2}\mathbf{1})}_{\text{population version}} + \underbrace{4t(A - \mathbb{E}(A|Z))(\psi^{(s-1)} - \frac{1}{2}\mathbf{1})}_{\text{sample noise}},$$

- M has a simple eigendecomposition:

$$w_1 = n\alpha_+ - (p^* - \lambda^*) \text{ with } \alpha_+ = \frac{p^* + q^*}{2} - \lambda^*, \quad \textit{eigenvector } u_1 = \mathbf{1}$$
$$w_2 = n\alpha_- - (p^* - \lambda^*) \text{ with } \alpha_- = \frac{p^* - q^*}{2}, \quad \textit{eigenvector } u_2 = \mathbf{1}_{\mathcal{C}_1} - \mathbf{1}_{\mathcal{C}_2}$$
$$w_j = -(p^* - \lambda^*),\ j = 3, \ldots, n$$



- Project $\psi^{(s)}$ on u_1, u_2. $\zeta_i^{(s)} = \langle\psi^{(s)}, u_i\rangle/\|u_i\|^2 = \langle\psi^{(s)}, u_i\rangle/n$, for $i = 1, 2$.

$$\psi^{(s)} = \zeta_1^{(s)}u_1 + \zeta_2^{(s)}u_2 + v^{(s)}.$$

$$\xi_i^{(s+1)} = 4tn\left((\zeta_1^{(s)} - \frac{1}{2})\alpha_+ + \sigma_i\zeta_2^{(s)}\alpha_-\right) + 4tw_3\left((\zeta_1^{(s)} - \frac{1}{2}) + \sigma_i\zeta_2^{(s)} + v_i^{(s)}\right)$$
$$=: na_{\sigma_i}^{(s)} + b_i^{(s)}, \qquad \sigma_i = \pm 1$$

- Key ψ to consider $\frac{1}{2}\mathbf{1}, \mathbf{1}, \mathbf{0}, \mathbf{1}_{\mathcal{C}_1}, \mathbf{1}_{\mathcal{C}_2}$.

Known p^*, q^*

Proposition 1 (Saddle point) *$\psi = \frac{1}{2}\mathbf{1}$ is a saddle point of the population mean field log-likelihood when p^* and q^* are known, for all n large enough.*

Theorem 1 (Population behavior) *The limit behavior of the population BCAVI updates is characterized by the signs of α_+ and $a_{\pm 1}^{(0)}$. Assume that $|na_{\pm 1}^{(0)}| \to \infty$, $\rho_n \to 0$. Define $\ell(\psi^{(0)}) = \mathbb{1}(a_{+1}^{(0)} > 0)\mathbf{1}_{\mathcal{C}_1} + \mathbb{1}(a_{-1}^{(0)} > 0)\mathbf{1}_{\mathcal{C}_2}$. Then, we have*

$$\frac{\|\psi^{(1)} - \ell(\psi^{(0)})\|^2}{n} = O(\exp(-\Theta(n\min\{|a_{+1}^{(0)}|, |a_{-1}^{(0)}|\}))) = o(1).$$

We also have for any $s \ge 2$,

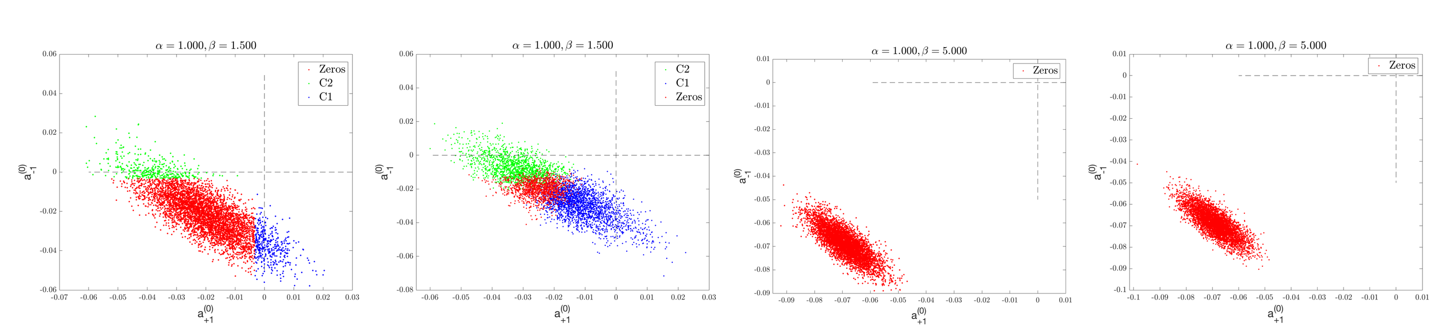
$$\frac{\|\psi^{(s)} - \ell(\psi^{(0)})\|^2}{n} = \begin{cases} O(\exp(-\Theta(nt\alpha_-))), & \textit{If } a_{+1}^{(0)}a_{-1}^{(0)} < 0 \\ O(\exp(-\Theta(nt\alpha_+))), & \textit{If } a_{+1}^{(0)}a_{-1}^{(0)} > 0. \end{cases}$$

For example, $a_{+1}^{(0)}a_{-1}^{(0)} < 0$, $\ell(\psi^{(0)}) = \mathbf{1}_{\mathcal{C}_1}$ or $\mathbf{1}_{\mathcal{C}_2}$; $a_{+1}^{(0)}a_{-1}^{(0)} > 0$, $\ell(\psi^{(0)}) = \mathbf{1}$ or $\mathbf{0}$.

For example, $a_{+1}^{(0)}a_{-1}^{(0)} < 0$, $\ell(\psi^{(0)}) = \mathbf{1}_{\mathcal{C}_1}$ or $\mathbf{1}_{\mathcal{C}_2}$; $a_{+1}^{(0)}a_{-1}^{(0)} > 0$, $\ell(\psi^{(0)}) = \mathbf{1}$ or $\mathbf{0}$. Consider the sample updates with iid initialization $\psi^{(0)}$.

Theorem 2 (Sample behavior) *For $s \ge 1$, the same conclusion holds for the sample BCAVI updates with high probability as long as $n|a_{\pm 1}^{(0)}| \gg \max\{\sqrt{n\rho_n}\|\psi^{(0)} - \frac{1}{2}\|_\infty, 1\}$, $\sqrt{n\rho_n} = \Omega(\log n)$ and $\psi^{(0)}$ is independent of A.*

Remark 1 *The above condition is not satisfied when $\mathbb{E}\psi_i^{(0)} = 1/2$. In this case, $\zeta_1^{(0)} - 1/2 = O_P(n^{-1/2})$, $\zeta_2^{(0)} = O_P(n^{-1/2})$, $n|a_{\pm 1}^{(0)}| = O_P(\sqrt{n}\rho_n)$.*



(A) (B) (C) (D)

Figure: $p^* = 0.4, q^* = 0.025, n = 200$, 5000 initializations with iid $Beta(\alpha, \beta)$: (A), (C) represent population behavior and (B), (D) represent the corresponding sample behavior.

Known p^*, q^*

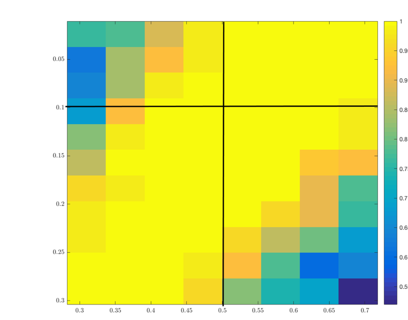


Figure: Robustness to estimation error in p, q.

- x axis has different p values and y axis has different q values.
- The lines represent p^* and q^*.
- The numbers represent average accuracy from 50 random initializations $Unif(0, 1)$.
- $p^* = 0.5$, $q^* = 0.1$, $n = 400$

Unknown p^*, q^*

Proposition 2 (Optimization landscape) *For n large enough, $(\psi, p, q) = (\frac{1}{2}\mathbf{1}, \frac{p^*+q^*}{2}, \frac{p^*+q^*}{2})$ is a strict local maximum of the population mean field log-likelihood.*

Proposition 3 *Consider the population updates of BCAVI with unknown p^*, q^* and $\rho_n \to 0$, $n\rho_n \to \infty$. Let $(\psi, \tilde{p}, \tilde{q})$ be a stationary point of the population mean field log-likelihood. If $\psi = \psi_u + \psi_{u^\perp}$, where $\psi_u \in$ span$\{u_1, u_2\}$ and $\psi_{u^\perp} \perp$ span$\{u_1, u_2\}$, then $\|\psi_{u^\perp}\| = o(\sqrt{n})$ as $n \to \infty$.*

Lemma 3 (Futility of random initializations) *Consider the initial distribution $\psi_i^{(0)} \overset{iid}{\sim} f_\mu$ where f is a distribution supported on $(0, 1)$ with mean μ. If μ is bounded away from 0 and 1 and $n\rho_n \to \infty$, then $\psi_i^{(s)} = \frac{1}{2} + O_P(\sqrt{\rho_n/n})$ for $s \ge 1$, where $\psi^{(s)}$ is computed using the full updates.*

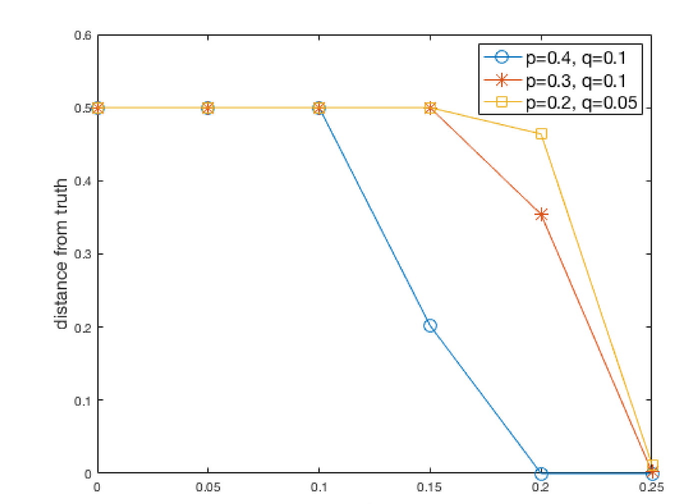
Lemma 4 (Initializations correlated with truth) *Consider an initial $\psi^{(0)}$ such that*

$$\zeta_1 = \frac{\mu_1 + \mu_2}{2} + O_P(1/\sqrt{n}) \qquad \zeta_2 = \frac{\mu_1 - \mu_2}{2} + O_P(1/\sqrt{n}).$$

If μ_1, μ_2 are bounded away from 0 and 1 and satisfy

$$|\mu_1 - \mu_2| > \max\left(2|\mu_1 + \mu_2 - 1| + O_P\left(\rho_n/\sqrt{n}\right), \left(\frac{\rho_n \log n}{n(p^* - q^*)^2}\right)^{1/3}\right),$$

and $n\rho_n \to \infty$, then $\psi^{(1)} = \mathbf{1}_{\mathcal{C}_1} + O_P(\exp(-\Omega(\log n)))$ or $\mathbf{1}_{\mathcal{C}_2} + O_P(\exp(-\Omega(\log n)))$.



Average distance between the estimated ψ and the true Z with respect to c_0, where $\mathbb{E}(\psi^{(0)}) = (1/2 + c_0)\mathbf{1}_{\mathcal{C}_1} + (1/2 - c_0)\mathbf{1}_{\mathcal{C}_2}$.

Generalizations - $K > 2$

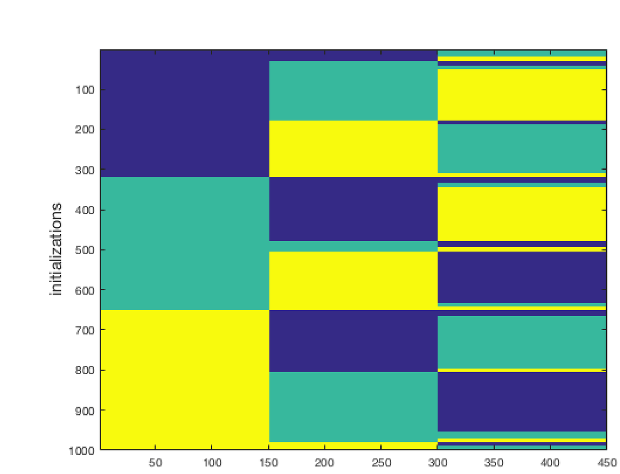


Figure: Convergence from random initialization for $K = 3$ with known p, q.

- $K = 3$, $p^* = 0.5$, $q^* = 0.01$, equal class, $n = 450$, initialized with $Dirichlet(0.1, 0.1, 0.1)$.
- For each iteration (each row) we represent the node membership with different colors.
- All stationary points lie in the span of $\{\mathbf{1}_{\mathcal{C}_1}, \mathbf{1}_{\mathcal{C}_2}, \mathbf{1}_{\mathcal{C}_3}\}$.
- We conjecture that the number of stationary points grow exponentially in K.
- Unknown p^*, q^* and random initializations lead to $(1/3, 1/3, 1/3)$.