EE290 Mathematics of Data Science Lecture 12 - 10/8/2019

Lecture 12: Correlated Recovery I

Lecturer: Jiantao Jiao *Scribe: Hansheng Jiang*

Today's topic: correlated recovery of SBM. This lecture is based on [1, Chapter 8].

1 SBM

We first briefly review the notation of SBM model. We have $SBM(n, 2, p, q)$ with scaling $p = a/n$, $q = b/n$.

$\sigma = (\sigma_1, \ldots, \sigma_n) \in \{-1, +1\}^n$ are iid sampled from $\mathrm{Rad}(\frac{1}{2})$.

$$\mathbb{P}(i \sim j) := \mathbb{P}(i \text{ connect with } j) = \begin{cases} p & \text{if } \sigma_i = \sigma_j \\ q & \text{if } \sigma_i \neq \sigma_j \end{cases}$$

$$A_{ij} = \mathbb{I}\{i \sim j\} \quad \text{iid for all } 1 \leqslant i < j \leqslant n.$$

2 Correlated Recovery

Definition 1 (Correlated recovery)**.** *For SBM, correlated recovery is to achieve a positive empirical correlation, i.e.,*

$$\frac{\mathbb{E}[\langle \sigma, \hat{\sigma} \rangle]}{n}$$

is strictly larger than 0 *as* $n \to \infty$.

Alternatively, correlated recovery is to achieve that

$$\min_{s \in \{+1, -1\}} \frac{\mathbb{E}[\|\sigma + s\hat{\sigma}\|_1]}{n}$$

strictly smaller than 1 *as* $n \to \infty$*, i.e., better than random guessing.*

Remark. We recall the following relation

$$\min_{s \in \{+1, -1\}} \|\sigma + s\hat{\sigma}\|_1 = n - |\langle \sigma, \hat{\sigma} \rangle|. \tag{1}$$

(1) can be proved by simply establishing that $|\sigma_i + \hat{\sigma}_i| = 1 + \sigma_i \hat{\sigma}_i$ and $|\sigma_i - \hat{\sigma}_i| = 1 - \sigma_i \hat{\sigma}_i$ for all $i = 1, \ldots, n$. The correlation $\langle \sigma, \hat{\sigma} \rangle$ is limited to symmetric SBM while the notion of $\min_{s \in \{+1, -1\}} \|\sigma + s\hat{\sigma}\|_1$ can be generalized to $k-$ community SBM.

3 Main Results

Theorem 2 (Mutual information characterization)**.** *Correlated recovery is possible if and only if* $I(\sigma_1, \sigma_2; G) > 0$ *as* $n \to \infty$.

Remark. Conditioned on $\sigma_1 = \sigma_2$ or $\sigma_1 \neq \sigma_2$, G is independent of (σ_1, σ_2). Since $\sigma_1 \sigma_2$ exactly tells $\sigma_1 = \sigma_2$ or $\sigma_1 \neq \sigma_2$, thus

$$I(\sigma_1, \sigma_2; G) = I(\sigma_1 \sigma_2; G).$$

In previous class, we have shown that in the Bayes hypothesis testing problem for $X \sim \mathrm{Rad}(\frac{1}{2})$ and observation Y

$$\min_{\hat{X}} \mathbb{P}(X \neq \hat{X}(Y)) = \frac{1}{2}(1 - \mathrm{TV}(P_+, P_-)),$$

where

$$P_+ := \mathcal{P}(Y \mid X = +1)$$
$$P_- := \mathcal{P}(Y \mid X = -1).$$

Before the proof of Theorem 2, we prove a lemma that relates mutual information to total variation.

Lemma 3.

$$\mathrm{TV}(P_+, P_-) = o(1) \Leftrightarrow I(X;Y) = o(1).$$

Proof [Proof of Lemma 3]
On the one hand, by Pinsker inequality

$$\begin{aligned} I(X;Y) &= \mathbb{E}_X D(P_{Y|X} \| P_Y) \\ &= \frac{1}{2} D(P_+ \| \bar{P}) + \frac{1}{2} D(P_- \| \bar{P}) \qquad \text{define } \bar{P} = (P_+ + P_-)/2 \\ &\geqslant \mathrm{TV}^2(P_+, \bar{P}) + \mathrm{TV}^2(P_-, \bar{P}) \\ &= \frac{1}{2} \mathrm{TV}^2(P_+, P_-). \end{aligned}$$

On the other hand,

$$\begin{aligned} I(X;Y) &= \frac{1}{2} D(P_+ \| \bar{P}) + \frac{1}{2} D(P_- \| \bar{P}) \\ &\leqslant \frac{1}{2} \int \frac{(P_+ - \bar{P})^2}{\bar{P}} + \frac{1}{2} \int \frac{(P_- - \bar{P})^2}{\bar{P}} \qquad \text{because } D(P\|Q) \leqslant \log(1 + \chi^2(P\|Q)) \leqslant \chi^2(P\|Q) \\ &= \int \frac{(P_+ - P_-)^2}{2(P_+ + P_-)} \\ &\leqslant \int \frac{1}{2} |P_+ - P_-| \\ &= \mathrm{TV}(P_+, P_-). \end{aligned}$$

□

Proof [Proof of Theorem 2]
Based on Lemma 3, we only need to show correlated recovery is possible if and only if $\mathrm{TV}(P_+, P_-) = \Omega(1)$.
($\Leftarrow$)
We show if $\mathrm{TV}(P_+, P_-) \geqslant \epsilon$, then we can do correlated recovery.
For all $i \neq j$, we denote $T_{ij} = \sigma_i \sigma_j$, then there exists test $\hat{T}_{ij}(G)$ such that

$$\mathbb{P}(\hat{T}_{ij} = T_{ij}) \geqslant \frac{1}{2} + \epsilon, \epsilon > 0.$$

We define an estimator $\hat{\sigma}$ by

$$\hat{\sigma}_1 = +1, \hat{\sigma}_i = \hat{T}_{1i}, i = 2, \ldots, n.$$

Then

$$\max_{s\in\{-1,+1\}} \sum_{i=1}^{n} \mathbb{P}(\sigma_i = s\hat{\sigma}_i) \geqslant \sum_{i=1}^{n} \mathbb{P}(\sigma_i = \hat{\sigma}_i) = \frac{1}{2} + \sum_{i=2}^{n} \mathbb{P}(T_{1i} = \hat{T}_{1i}) \geqslant \frac{1}{2} + (1/2+\epsilon)(n-1).$$

($\Rightarrow$)

If $\mathrm{TV}(P_+, P_-) = o(1)$, then for any $\hat{T}_{ij}$, $\mathbb{P}(\hat{T}_{ij} = \sigma_i\sigma_j) = \frac{1}{2} + o(1)$. For any estimator $(\hat{\sigma}_1, \ldots, \hat{\sigma}_n)$,

$$\begin{aligned}\mathbb{E}\|\sigma\sigma^{\mathrm{T}} - \hat{\sigma}\hat{\sigma}^{\mathrm{T}}\|_F^2 &= \sum_{i\neq j} \mathbb{P}(\sigma_i\sigma_j \neq \hat{\sigma}_i\hat{\sigma}_j) \\ &= 2n^2 - o(n^2).\end{aligned}$$

On the other hand, $\mathbb{E}\|\sigma\sigma^{\mathrm{T}} - \hat{\sigma}\hat{\sigma}^{\mathrm{T}}\|_F^2 = 2n^2 - \mathbb{E}|\langle\sigma, \hat{\sigma}\rangle|^2$.

Combining two equations, we have

$$\frac{\mathbb{E}|\langle\sigma, \hat{\sigma}\rangle|^2}{n^2} = o(1),$$

which implies that

$$\frac{\mathbb{E}|\langle\sigma, \hat{\sigma}\rangle|}{n} = o(1)$$

due to Jensen's inequality.

□

Corollary 4. *If $\tau := \frac{(a-b)^2}{2(a+b)} < 1$, then correlated recovery is possible.*

Proof [Proof of Corollary 4]

It suffices to show that $\tau < 1 \Rightarrow \mathrm{TV}(P_+, P_-) \Rightarrow I(\sigma_1, \sigma_2; G) = o(1)$.

We first show that total variation has the following variational characterization

$$\mathrm{TV}(P, Q) = \frac{1}{2} \inf_R \sqrt{\frac{(P-Q)^2}{Q}},$$

where the equality is achieved by taking $R^* = \frac{|P-Q|}{\int |P-Q|}$, and the infimum is due to Cauchy-Schwarz inequality

$$\int \frac{(P-Q)^2}{R} = \int \left(\frac{P-Q}{\sqrt{R}}\right)^2 \int (\sqrt{R})^2 \geqslant (\int |P-Q|)^2 = 4\mathrm{TV}^2(P, Q)$$

Since it suffices to show

$$\begin{aligned}\int \frac{(P_+ - P_-)^2}{R} &= \int \frac{P_+^2 + P_-^2 - 2P_+P_-}{R} \\ &= \int \frac{P_+^2}{R} + \int \frac{P_-^2}{R} - 2\int \frac{P_+P_-}{R} \\ &= o(1),\end{aligned}$$

we know it suffices to show that $\int \frac{P_z P_{\tilde{z}}}{R} = C + o(1)$ for all $z, \tilde{z} \in \{+1, -1\}$ for an R of choice, where C is some constant that is independent of $z, \tilde{z}$.

To calculate $\int \frac{P_z P_{\tilde{z}}}{R}$, we need to use the second moment trick. If $P_z = \int P_\theta \pi(\mathrm{d}\theta)$, and $P_{\tilde{z}} = \int P_{\tilde{\theta}} \pi(\mathrm{d}\tilde{\theta})$ then by Fubini theorem

$$\int \frac{P_z P_{\tilde{z}}}{R} = \int \pi(\mathrm{d}\theta)\pi(\mathrm{d}\tilde{\theta}) \int \frac{P_\theta(x) P_{\tilde{\theta}}(x)}{R(x)} \mathrm{d}x.$$

Specifically, the distribution of the adjacency matrix A is

$$P(A \mid \sigma) = \prod_{i<j}(P\mathbb{I}\{\sigma_i = \sigma_j\} + Q\mathbb{I}\{\sigma_i \neq \sigma_j\}) = \prod_{i<j}(\frac{P+Q}{2} + \frac{P-Q}{2}\sigma_i\sigma_j),$$

where $P = \mathsf{Bern}(p), Q = \mathsf{Bern}(q)$. We pick R to be the distribution of $G(n, \frac{a+b}{2n})$, i.e. $R = \prod_{i<j} \frac{P+Q}{2}$, then after calculations, we can obtain that

$$\int \frac{P_\sigma P_{\tilde{\sigma}}}{R} = \prod_{i<j}(1 + \rho\sigma_i\sigma_j\tilde{\sigma}_i\tilde{\sigma}_j),$$

where $\rho = \tau/n + O(1/n^2)$.

Now our goal is to calculate the conditional expectation

$$\begin{aligned}
&\mathbb{E}\left[\prod_{i<j}(1 + \rho\sigma_i\sigma_j\tilde{\sigma}_i\tilde{\sigma}_j) \mid \sigma_1\sigma_2 = z, \tilde{\sigma}_1\tilde{\sigma}_2 = \tilde{z}\right] \\
=&\mathbb{E}\left[\prod_{i<j} e^{\rho\sigma_i\sigma_j\tilde{\sigma}_i\tilde{\sigma}_j - \rho^2/2 + O(\rho^3)} \mid \sigma_1\sigma_2 = z, \tilde{\sigma}_1\tilde{\sigma}_2 = \tilde{z}\right] \\
=&e^{(n(n-1)/2)(-\rho^2/2)}\mathbb{E}\left[e^{\sum_{i<j}\rho\sigma_i\sigma_j\tilde{\sigma}_i\tilde{\sigma}_j + O(1/n)} \mid \sigma_1\sigma_2 = z, \tilde{\sigma}_1\tilde{\sigma}_2 = \tilde{z}\right] \\
\overset{(i)}{=}&e^{-\tau^2/4 - \tau/2}\mathbb{E}\left[e^{\frac{\tau}{2}(\frac{1}{n}(\sum_{i=1}^n \sigma_i\tilde{\sigma}_i)^2) + o(1)} \mid \sigma_1\sigma_2 = z, \tilde{\sigma}_1\tilde{\sigma}_2 = \tilde{z}\right],
\end{aligned} \tag{2}$$

where in (i) we used the expansion $(\sum_i \sigma_i\tilde{\sigma}_i)^2 = 2\sum_{i<j}\sigma_i\tilde{\sigma}_i\sigma_j\tilde{\sigma}_j + \sum_i \sigma_i\sigma_i\tilde{\sigma}_i\tilde{\sigma}_i = 2\sum_{i<j}\sigma_i\tilde{\sigma}_i\sigma_j\tilde{\sigma}_j + n$, and before it we used the expansion $\log(1+x) = x - x^2/2 + O(x^3)$.

Finally, we observe that $\frac{1}{\sqrt{n}}(\sum_{i=1}^n \sigma_i\tilde{\sigma}_i)$ weakly converges to $N(0,1)$ by central limit theorem. To make the quantity in (2) converges to a constant as n goes to infinity, it is effectively very similar to controlling $\exp\{\frac{\tau}{2}N^2(0,1)\}$. One can imagine $\tau < 1$ is sufficient by considering the limiting case $\mathbb{E}_{Z\sim N(0,1)}e^{\frac{\tau}{2}Z^2} = (2\pi)^{-1/2}e^{\frac{\tau}{2}z^2}e^{-z^2/2}$. This last argument is explained intuitively but it can be justified using truncation and uniform integrability arguments. □

References

[1] Yihong Wu and Jiaming Xu. Statistical inference on graphs: Selected topics, October 2019. `http://www.stat.yale.edu/~yw562/teaching/stats-graphs.pdf`.