Nearest neighbor smoother

Chanwoo Lee

1 Proof of Lemmas from Wang's 0624 note based on [2]

Lemma 1 (Permutation error)**.**

$$\text{Loss}(\sigma, \hat{\sigma}) := \frac{1}{d} \max_i |\sigma(i) - \hat{\sigma}(i)| \leq d^{-(m-1)\beta/2}.$$

Lemma 2 (Estimation error due to permutation)**.**

$$\mathbb{E}\|\text{Block}_k(\mathcal{Y} \circ \hat{\sigma}^{-1}) - \text{Block}_k(\mathcal{Y} \circ \sigma^{-1})\|_F^2 \leq d^m \text{Loss}^2(\sigma, \hat{\sigma}).$$

Proof. Define $\mathcal{A} := \mathcal{Y} \circ \sigma^{-1}$ and $\hat{\mathcal{A}} := \mathcal{Y} \circ \hat{\sigma}^{-1}$. Notice that

$$\mathbb{E}\|\text{Block}_k(\mathcal{Y} \circ \hat{\sigma}^{-1}) - \text{Block}_k(\mathcal{Y} \circ \sigma^{-1})\|_F^2 =$$
$$h^m \mathbb{E}\left(\sum_{\substack{k_i \in \{0,\ldots,k-1\} \\ i=1,\ldots,m}} \left(\frac{1}{h^m} \sum_{\substack{h_j \in \{0,\ldots,h-1\} \\ j=1,\ldots,m}} \hat{\mathcal{A}}_{\omega(k_1,\ldots,k_m,h_1,\ldots h_m)} - \mathcal{A}_{\omega(k_1,\ldots,k_m,h_1,\ldots h_m)} \right)^2 \right), \tag{1}$$

where $\omega(k_1, \ldots, k_m, h_1, \ldots h_m) = (k_1 h + h_1, \ldots, k_m h + h_m)$.

Then, we bound

$$\mathbb{E}\left(\hat{\mathcal{A}}_\omega - \mathcal{A}_\omega\right)^2 = \mathbb{E}\left(\underbrace{(\hat{\mathcal{A}}_\omega - [\Theta \circ \sigma \circ \hat{\sigma}^{-1}]_\omega)^2}_{(a)} + \underbrace{(\mathcal{A}_\omega - \Theta_\omega)^2}_{(b)} + \underbrace{([\Theta \circ \sigma \circ \hat{\sigma}^{-1}]_\omega - \Theta_\omega)^2}_{(c)} \right)$$

Notice that the (a) and (b) equal to $\text{Var}(\hat{\mathcal{A}}_\omega)$ and $\text{Var}(\mathcal{A}_\omega)$ respectively, bounded by 1. For (c),

$$\begin{aligned} ([\Theta \circ \sigma \circ \hat{\sigma}^{-1}]_\omega - \Theta_\omega)^2 &= ([\Theta \circ \sigma]_{\omega'} - [\Theta \circ \hat{\sigma}]_{\omega'})^2, \quad \text{for some } \omega' \in [d]^m \\ &\leq \frac{L^2 |\sigma(\omega') - \hat{\sigma}(\omega'|_1^2}{d^2} \\ &\lesssim \text{Loss}^2(\sigma, \hat{\sigma}). \end{aligned}$$

Going back to (1), we show that

$$\begin{aligned} &h^m \mathbb{E}\left(\sum_{\substack{k_i \in \{0,\ldots,k-1\} \\ i=1,\ldots,m}} \left(\frac{1}{h^m} \sum_{\substack{h_j \in \{0,\ldots,h-1\} \\ j=1,\ldots,m}} \hat{\mathcal{A}}_{\omega(k_1,\ldots,k_m,h_1,\ldots h_m)} - \mathcal{A}_{\omega(k_1,\ldots,k_m,h_1,\ldots h_m)} \right)^2 \right) \\ &\leq \frac{k^m}{h^m} \left(h^m \left(2 + \text{Loss}^2(\sigma, \hat{\sigma})\right) + \frac{h^m(h^m - 1)}{2} \text{Loss}^2(\sigma, \hat{\sigma}) \right) \\ &\lesssim d^m \text{Loss}^2(\sigma, \hat{\sigma}). \end{aligned}$$

□

Lemma 3 (Denoising error)**.**

$$\mathbb{E}\|\mathrm{Block}_k(\mathcal{Y}\circ\sigma^{-1}) - \mathrm{Block}_k(\Theta)\|_F^2 \le k^m.$$

Proof. Notice that $\mathbb{E}(\mathrm{Block}_k(\mathcal{Y}\circ\sigma^{-1})) = \mathrm{Block}_k(\Theta)$. Consequently, we have

$$\mathbb{E}\left([\mathrm{Block}_k(\mathcal{Y}\circ\sigma^{-1})]_\omega - [\mathrm{Block}_k(\Theta)]_\omega\right)^2 = \mathbb{E}\left([\mathrm{Block}_k(\mathcal{Y}\circ\sigma^{-1})]_\omega^2\right) - [\mathrm{Block}_k(\Theta)]_\omega^2. \tag{2}$$

Hence we bound

$$\begin{aligned}
\mathbb{E}\left([\mathrm{Block}_k(\mathcal{Y}\circ\sigma^{-1})]_\omega^2\right) &= \mathbb{E}\left(\frac{1}{h^m}\sum_{\omega'\in\left[\lfloor\frac{\omega-1}{h}\rfloor h+1,\lfloor\frac{\omega-1}{h}\rfloor h+h\right]}[\mathcal{Y}\circ\sigma^{-1})]_{\omega'}\right)^2\\
&= \frac{1}{h^{2m}}\left(\sum_{\substack{\omega',\omega''\in\left[\lfloor\frac{\omega-1}{h}\rfloor h+1,\lfloor\frac{\omega-1}{h}\rfloor h+h\right]\\ \omega'\neq\omega''}}\Theta_{\omega'}\Theta_{\omega''} + \sum_{\omega'\in\left[\lfloor\frac{\omega-1}{h}\rfloor h+1,\lfloor\frac{\omega-1}{h}\rfloor h+h\right]}\Theta_{\omega'}^2 + \mathrm{Var}([\mathcal{Y}\circ\sigma^{-1}]_{\omega'})\right)\\
&\le \left(\frac{1}{h^m}\sum_{\omega'\in\left[\lfloor\frac{\omega-1}{h}\rfloor h+1,\lfloor\frac{\omega-1}{h}\rfloor h+h\right]}\Theta'_\omega\right)^2 + \frac{c}{h^m}\\
&= [\mathrm{Block}_k(\Theta)]_\omega^2 + \frac{c}{h^m},
\end{aligned}$$

for some constant $c>0$. Notice that we set $c = 1/4$ when $\mathcal{Y}\circ\sigma^{-1}$ is Bernoulli distribution, while $c = \mathrm{Var}(\mathcal{E}_\omega)$ if $\mathcal{E}$ is from standard normal. Therefore, combining the above inequality and (2) gives us

$$\mathbb{E}\|\mathrm{Block}_k(\mathcal{Y}\circ\sigma^{-1}) - \mathrm{Block}_k(\Theta)\|_F^2 \le \frac{d^m}{h^m} \le k^m.$$

$\square$

Lemma 4 (Approximation error)**.** For every fixed integer $k \le d$, we have

$$\|\mathrm{Block}_k(\Theta) - \Theta\|_F^2 \lesssim \frac{d^m}{k^2}.$$

Proof. Notice that for any $\omega\in[d]^m$,

$$([\mathrm{Block}_k(\Theta)]_\omega - \Theta_\omega)^2 = \left(\frac{1}{h^m}\sum_{\omega'\in\left[\lfloor\frac{\omega-1}{h}\rfloor h+1,\lfloor\frac{\omega-1}{h}\rfloor h+h\right]}(\Theta_{\omega'} - \Theta_\omega)\right)^2.$$

Notice that

$$|\Theta_{\omega'} - \Theta_\omega|^2 \le \frac{L^2\,|\omega'-\omega|_1^2}{d^2}$$

$$\lesssim \frac{1}{k^2},$$

where the last inequality uses the fact that $\omega'_i \in \left[\lfloor \frac{\omega_i - 1}{h} \rfloor h + 1, \lfloor \frac{\omega_i - 1}{h} \rfloor h + h\right]$ for all $i \in [m]$. $\square$

2 Intuition of [1, 4] and distinction from histogram method [2]

We consider the following model,

$$A_{ij} = \Theta_{ij} + \epsilon_{ij} = \Theta(\xi_i, \xi_j) + \epsilon_{ij},$$

where ϵ_{ij} denote the Bernoulli error depending on Θ_{ij} and ξ_i's are the latent variables that has not been observed. Generally speaking, neighbor-based methods [1, 4] set $\mathcal{N}_i$ for each node $i \in [d]$ and obtain probability matrix averaging corresponding observed matrix. To be specific, for given node $i \in [d]$, we define the neighbor set $\mathcal{N}_i$ of a node i, according to how close nodes are from i-th node based on certain criteria. Then, the probability matrix is estimated by

- Probability matrix estimation [1]:

$$\hat{\Theta}_{ij} = \frac{\sum_{i' \in \mathcal{N}_i, j' \in \mathcal{N}_j} A_{i'j'}}{|\mathcal{N}_i||\mathcal{N}_j|}.$$

- Probability matrix estimation [4]:

$$\hat{\Theta}_{ij} = \frac{\sum_{i' \in \mathcal{N}_i} A_{i'j}}{|\mathcal{N}_i|}.$$

As one can see, the major difference between [1] and [4] is how to estimate probability matrix estimation given the neighbor set $\mathcal{N}_i$.

Another difference is how to define the empirical distance between two different nodes. The distance between two nodes are defined as follows

- Distance in [1]: $d^2(i,j) = \frac{1}{2}\left(\int_0^1 |\Theta(\xi_i, v) - \Theta(\xi_j, v)|^2 dv + \int_0^1 |\Theta(v, \xi_i) - \Theta(v, \xi_j)|^2 dv\right)$.
- Distance in [4]: $d^2(i,j) = \int_0^1 |\Theta(\xi_i, v) - \Theta(\xi_j, v)|^2 dv$.

For symmetric probability function Θ, two distances are the same. However, two paper take different empirical distance between two different nodes. In addition, [1] requires more than two observed adjacency matrix to obtain the empirical distance while [4] needs only one observed adjacency matrix.

We view the probability estimation procedure as kernel smoothing methods with nearest smoother. For given $(X_i, Y_i)_{i=1}^N$, kernel smoothing methods estimates $Y(X_0)$ by

$$\hat{Y}(X_0) = \frac{\sum_{i=1}^{N} K_{h_\lambda}(X_0, X_i) Y(X_i)}{\sum_{i=1}^{N} K_{h_\lambda}(X_0, X_i)}, \text{ where } K_{h_\lambda}(X_0, X) = D\left(\frac{\|X - X_0\|}{h_\lambda(X_0)}\right). \tag{3}$$

Here $\|\cdot\|$ is the Euclidean norm, $h_\lambda(X_0)$ is a parameter (kernel radius), and $D(t)$ is typically a positive real valued function, whose value is decreasing (or not increasing) for the increasing distance between the X and X_0.

Examples of kernel smoothers are as follow.

- The Gaussian kernel is one of the most widely used kernels, and is expressed with the equation below.

$$K(x^*, x_i) = \exp\left(-\frac{(x^* - x_i)^2}{2b^2}\right) K(x^*, x_i) = \exp\left(-\frac{(x^* - x_i)^2}{2b^2}\right)$$

- Nearest smoother is expressed with setting functions as follow, $h_m(X_0) = \left\|X_0 - X_{[m]}\right\|$, where $X_{[m]}$ is the m-th closest to X_0 neighbor, and

$$D(t) = \begin{cases} 1/m & \text{if } |t| \le 1 \\ 0 & \text{otherwise} \end{cases}$$

  Let me express nearest smoother in our context. Define $\mathcal{N}_0 = \{X_{[1]}, \ldots, X_{[m]}\}$ as m-closest X_0 neighbor. Then, (3) is

$$\hat{Y}(X_0) = \frac{\sum_{X \in \mathcal{N}_0} Y(X)}{|\mathcal{N}_0|}. \tag{4}$$

Notice that $\hat{Y}$ in (4) corresponds to $\hat{\Theta}$ and $Y(X)$ to A_{ij} where an index (i, j) becomes a predictor. This correspondence shows the connection between kernel smoother method and proposed methods in [1, 4]. One distinction between kernel smoother method and that of the network context is that the distance of predictors is a simple Euclidean distance for kernel smooth method while we need to define the distance of predictors (nodes) and estimate empirical one based on observed network.

Remark 1. From this point of view, relationship between histogram method in [2] and neighbor-based method in [1, 4] is similar to one between histogram density estimation and kernel density estimation.

Denote $\deg_{[i]}$ as i-th largest degrees of nodes. Define distance between two nodes as

$$d(i, j) = \max_{k \in [K]} \mathbb{1}\{\deg(i), \deg(j) \in \mathcal{N}_k\}, \text{ where } \mathcal{N}_k = \{\deg_{[(h(k-1)+1]}, \ldots, \deg_{[hk]}\}. \tag{5}$$

Then, we check that histogram estimation in [2] is a special case of neighborhood-based method with distance function (5). Therefore, what matters is how to define the distance between two nodes to define neighbor $\mathcal{N}_i$ for $i \in [d]$.

3 Comparison between [1] and [4]

Table 1 summarizes estimation methods, optimal block size, neighborhood size with respect to the block size, and convergence rate for each method.

Method	[1]	[2]	[4]
Estimation ($\hat{\Theta}$)	$\hat{\Theta}_{ij} = \frac{\sum_{i' \in \mathcal{N}_i, j' \in \mathcal{N}_j} A_{i'j'}}{\lvert\mathcal{N}_i\rvert\lvert\mathcal{N}_j\rvert}$	$\hat{\Theta}_{ij} = \frac{\sum_{i' \in \mathcal{N}_i, j' \in \mathcal{N}_j} A_{i'j'}}{\lvert\mathcal{N}_i\rvert\lvert\mathcal{N}_j\rvert}$	$\hat{\Theta}_{ij} = \frac{\sum_{i' \in \mathcal{N}_i} A_{i'j}}{\lvert\mathcal{N}_i\rvert}$
Block size (K)	NA*	$\sqrt{n}$	$\mathcal{O}(\sqrt{\frac{n}{\log n}})$
Neighborhood size	NA*	K^2	K
Convergence rate	$\mathcal{O}\left(\left(\frac{\log n}{n}\right)^{1/4}\right)$	$\mathcal{O}\left(\frac{\log n}{n}\right)$	$\mathcal{O}\left(\left(\frac{\log n}{n}\right)^{1/2}\right)$

* In this paper, the neighbor set is defined as $\mathcal{N}_i = \{j \in [d] \colon \hat{d}(i,j) \leq \Delta\}$ for a prespecified precision parameter Δ. But theoretically, they showed $K > \left(\frac{n}{\log n}\right)^{1/4}$ with high probability.

Table 1: Summary of three different estimation methods for probability matrix. The convergence rates are the upper bound of $\frac{1}{n^2}\|\hat{\Theta} - \Theta\|_F^2$.

Strict monotonicity condition: [2] does require strict monotonicity condition on the degree function. Main reason for this condition is to resolve identifiability issue. [3] explains the necessity of strict monotonicity condition well with the following example

$$
\begin{aligned}
f(u,v) &:= \mathbb{1}_{[0,1/2]^2}(u,v) + \mathbb{1}_{[1/2,1]^2}(u,v), \\
f_0(u,v) &:= \mathbb{1}_{[0,1/2]\times[1/2,1]}(u,v) + \mathbb{1}_{[1/2,1]\times[0,1/2]}(u,v),
\end{aligned}
$$

which give monotone non-decreasing degree function $\deg(u) = \deg_0(u) = 1/2$, generate a same graph, yet are different for a.e. $(u,v) \in [0,1]^2$.

However, [2] did not use this monotonicity when they prove the upper bound of estimation error (In addition, their proof is not correct and I think this condition should be used to guarantee the convergence).

For our setting, we use the strict monotonicity condition to control $\text{Loss}(\sigma, \hat{\sigma}) := \frac{1}{d}\max_i |i - \hat{\sigma}(i)|$. Intuitively, if there is no strict monotonicity condition, degree does not give any information about indices. For example, consider the equal degree case, i.e., $\deg(1) = \cdots \deg(d)$. In this case, we cannot control $\text{Loss}(\sigma, \hat{\sigma})$, due to the lack of information from the degree.

Extension to hypergraphon: We consider the extension to hypergraphon. Here we observe an adjacency tensor $\mathcal{A} \in \{0,1\}^{d_1 \times \cdots d_m}$ and want to estimate probability tensor $\Theta \in [0,1]^{d_1 \times \cdots \times d_m}$

1. Probability matrix estimation in [4] can be generalized to tensor case as

$$
\hat{\Theta}_\omega = \frac{\sum_{\omega_1' \in \mathcal{N}_{\omega_1}, \ldots, \omega_{m-1}' \in \mathcal{N}_{\omega_{m-1}}} \mathcal{A}_{\omega_1', \ldots, \omega_{m-1}', \omega_m}}{\prod_{\ell=1}^{m-1} |\mathcal{N}_{\omega_\ell}|}
$$

2. Probability matrix estimation in [1] can be generalized to tensor case as

$$
\hat{\Theta}_\omega = \frac{\sum_{\omega_1' \in \mathcal{N}_{\omega_1}, \ldots \omega_m' \in \mathcal{N}_{\omega_m'}} A_{\omega'}}{\prod_{\ell=1}^{m} |\mathcal{N}_{\omega_\ell}|}.
$$

References

[1] Edoardo M Airoldi, Thiago B Costa, and Stanley H Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. *arXiv preprint arXiv:1311.1731*, 2013.

[2] Stanley Chan and Edoardo Airoldi. A consistent histogram estimator for exchangeable graph models. In *International Conference on Machine Learning*, pages 208–216. PMLR, 2014.

[3] Justin Yang, Christina Han, and Edoardo Airoldi. Nonparametric estimation and testing of exchangeable graph models. In *Artificial Intelligence and Statistics*, pages 1060–1067. PMLR, 2014.

[4] Yuan Zhang, Elizaveta Levina, and Ji Zhu. Estimating network edge probabilities by neighborhood smoothing. *arXiv preprint arXiv:1509.08588*, 2015.