Nearest neighbor smoother

Chanwoo Lee

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

Lemma 1 (Permutation error)**.**

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

Lemmas 2-3 hold in expectation. Could you show they also hold in high probability? Perhaps combine current expectation bounds + Hoeffding inequality —> high probability bounds

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

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

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

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

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

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 $\mathrm{Var}(\hat{\mathcal{A}}_\omega)$ and $\mathrm{Var}(\mathcal{A}_\omega)$ respectively, bounded by 1. For (c),

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

Loss^2(..)

Going back to (1), we show that

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

□

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

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

Hence we bound

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

holds only for Bernoulli Y

Add remark for Bernoulli. vs Gaussian.
Express the total error in terms of noise variance.

Therefore, combining the above inequality and (2) gives us

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

□

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

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

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

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

Notice that

$$\begin{aligned}
|\Theta_{\omega'} - \Theta_\omega|^2 &\leq \frac{L^2\,|\omega'-\omega|_1^2}{d^2} \\
&\lesssim \frac{1}{k^2},
\end{aligned}$$

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 $i \in [m]$. □

2 Intuition of [1, 3] 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, 3] 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 [3]:

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

As one can see, the major difference between [1] and [3] 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]: $d(i,j)^2 = \frac{1}{2}\left(\int_0^1 |\Theta(\xi_i, v) - \Theta(\xi_j, v)|^2 dv + \int_0^1 |\Theta(v, \xi_i) - \Theta(v, \xi_j)|^2 dv\right)$.
- Distance in [3]: $d(i,j)^2 = \int_0^1 |\Theta(\xi_i, v) - \Theta(\xi_j, v)|^2 dv$.

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 [3] 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

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

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

bor. 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, 3]. 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.

From this point of view, relationship between histogram method in [2] and neighbor-based method in [1, 3] 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

good!

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

Add remark:
The comparison also justifies why [3] chooses h-th quantile (rather than absolute value) when determining the neighborhood size.

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 $i \in [d]$.

Both methods have same neighborhood sizes under optimal tuning parameters.

histogram [2]: neighborhood size h ~ 1/k ~ 1/sqrt(n)
NN smoothing [3]: neighborhood size h-th quantile~1/sqrt(n)

what's the intuition for requiring monotonic degree in [2]?

3 Things to do

1. List of the corresponding Lemmas 1-4 for the estimator in [3]: The difficulty of this is from the fact that [3] choose to use row-wise convergence $\|\hat{\Theta}_{i\cdot} - \Theta_{i\cdot}\|_2$. I am thinking about how to generalize probability matrix estimation [3]:

   convergence rate: 1/\sqrt{n}

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

   green vs. blue:
   which one has better accuracy (in theory)?

   to tensor case well. The point is whether it is reasonable to use only $A_{\omega_1', \omega_2, \ldots, \omega_m}$ for $\omega_1' \in \mathcal{N}_{\omega_1}$ to estimate $A_{\omega_1, \ldots, \omega_m}$. i.e.

   I doubt. This is essentially still a matrix problem of size d-by-d^2. No gain from tensor.

   $$\hat{\Theta}_{\omega} = \frac{\sum_{\omega_1' \in \mathcal{N}_{\omega_1}} \mathcal{A}_{\omega_1', \omega_2, \ldots, \omega_m}}{|\mathcal{N}_{\omega_1}|}.$$

   Notice that the generalization of probability matrix estimation [1]:

   what is the convergence rate?

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

   is direct as follows given the neighbor set $\mathcal{N}$

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

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