Graphon estimation review

Chanwoo Lee
February 18, 2021

1 Graphon estimation

1.1 Problem setting

We consider a random graph with adjacency matrix $\{A_{ij}\} \in \{0,1\}^{n\times n}$ whose sampling procedure is determined by

$$(\xi_1,\dots,\xi_n) \sim \mathbb{P}_\xi, \quad A_{ij}|(\xi_i,\xi_j) \sim \text{Bernoulli}(\theta_{ij}), \quad \text{where } \theta_{ij} = f(\xi_i,\xi_j). \tag{1}$$

We assume that A_{ij}'s are mutually independent conditioned on $(\xi_1,\dots,\xi_n)$ across all $i<j$. The sequence $\{\xi_i\}$ are random variables sampled from a distribution $\mathbb{P}_\xi$. A common choice for the probability $\mathbb{P}_\xi$ is i.i.d. uniform distribution on $[0,1]$, but can be allowed to be any distributions. The function f on $[0,1]^2$, which is assumed to be symmetric, is called graphon. We can consider more generalized adding sparsity parameter $\rho \in [0,1]$ as $\rho\theta_{ij}$ in Bernoulli probability [Xu, 2018, Ghoshdastidar and Dukkipati, 2017].

We view the setting (1) as modeling the mean of A_{ij} by a regression function $f(\xi_i,\xi_j)$ without observing $\{(\xi_i,\xi_j)\}$. This causes an identifiability problem, because without observing the design, there is no way to associate the value of $f(x,y)$ with (x,y). There are some lines of work to overcome this identifiability issue. One way is to restrict the function class of the graphons. In Chan and Airoldi [2014], they assume that all graphons of interests satisfy the strict monotonicity condition such that for the graphon f,

$$g(u) \overset{\text{def}}{=} \int_0^1 f(u,v)dv \tag{2}$$

is strictly increasing (or decreasing). However, this function restriction is too strict and cannot handle row and column permutations. Another way to overcome the identifiability issue is to consider the following loss function [Gao et al., 2015, Xu, 2018]:

$$\frac{1}{n^2}\sum_{i,j\in[n]}(\hat{\theta}_{ij}-\theta_{ij})^2.$$

This is identical to the loss function used in the classical nonparametric regression problem of the form:

$$\frac{1}{n^2}\sum_{ij\in[n]}(\hat{f}(\xi_i,\xi_j)-f(\xi_i,\xi_j))^2.$$

Estimation of the matrix $\{\theta_{ij}\}$ is still possible without observing the design. For the estimation consistency, I will follow the second approach considering (2).

1.2 Estimation method

I introduce two different ways of estimating $\{\theta_{ij}\}$.

1. Least square minimization [Gao et al., 2015]:

   Let $z \in [k]^n$ define a clustering structure on the n nodes. For any $Q = \{Q_{ab}\} \in \mathbb{R}^{k\times k}$ and $z \in [k]^n$, define the objective function

$$L(Q,z) = \sum_{a,b\in[k]}\sum_{(i,j)\in z^{-1}(a)\times z^{-1}(b), i\neq j}(A_{ij}-Q_{ab})^2.$$

For any optimizer of the objective function $(\hat{Q}, \hat{z}) \in \arg\min_{Q \in \mathbb{R}^{k\times k}, z\in[k]^n} L(Q, z)$, the estimator of θ_{ij} is defined as

$$\hat{\theta}_{ij} = \hat{Q}_{\hat{z}(i)\hat{z}(j)}, \quad i > j.$$

and $\hat{\theta}_{ij} = \hat{\theta}_{ji}$ for $i < j$.

2. Universal singular value thresholding [Chatterjee et al., 2015, Xu, 2018]:

   From given matrix $A \in \mathbb{R}^{n\times n}$, we use singular value decomposition of $A = \sum_{i=1}^{n} s_i u_i v_i^T$ to estimate θ. First we threshold the singular values: $S = \{i \colon s_i \geq \tau\}$ where $\tau = c\sqrt{n\rho}$ (in dense network case, set $\rho = 1$). Then we estimate θ_{ij} as

$$\theta_{ij} = \max\left(0, \min\left(\sum_{i\in S} s_i u_i v_i^T/\rho, 1\right)\right).$$

The first method directly targets to minimize least square loss [Gao et al., 2015]. While this approach enjoys rate-optimality, its computation is intractable. The second method can estimate the optimizers by polynomial time algorithm with the sacrifice of convergence rates. In this method, we consider the set of all symmetric matrices that have low ranks [Chatterjee et al., 2015, Xu, 2018]. To control the tails of eigenvalues of target matrix θ, Xu [2018] further assumes that eigenvalues of θ satisfies polynomial or super-polynomial decay. For the stochastic block model, those conditions are not needed anymore because $\lambda_i(\theta) = 0$ for $i > k$ where k is the number of block.

Definition 1 (Polynomial decay)**.** We say the eigenvalues of θ asymptotically satisfy a polynomial decay with rate $\beta > 0$ if all integers $0 \leq r \leq n-1$,

$$\frac{1}{n^2}\sum_{i\geq r+1} \mathbb{E}[\lambda_i^2(\theta)] \leq c_0 \frac{1}{r^\beta} + c_1 \frac{1}{n}$$

where c_0, c_1 are two constants independent of n and r.

Definition 2 (Super-polynomial decay)**.** We say the eigenvalues of θ asymptotically satisfy a polynomial decay with rate $\beta > 0$ if all integers $0 \leq r \leq n-1$,

$$\frac{1}{n^2}\sum_{i\geq r+1} \mathbb{E}[\lambda_i^2(\theta)] \leq c_0 e^{-c_2 r^\beta} + c_1 \frac{1}{n}$$

where c_0, c_1, c_2 are two constants independent of n and r.

1.3 Convergence rate

Convergence rates for graphon estimation are suggested with two perspectives: stochastic block model and nonparametric graphon estimation under Holder smoothness. We define mean square error as

$$\mathrm{MSE}(\hat{\theta}, \theta) = \mathbb{E}\left\{\frac{1}{n^2}\sum_{i,j\in[n]} (\hat{\theta}_{ij} - \theta_{ij})^2\right\}.$$

First, Gao et al. [2015] showed the minimax convergence rates depending on two different settings.

1. **Under stochastic block model with k blocks.** Define the class of SBM with k clusters as

$$\Theta_k = \left\{\{\theta_{ij}\} \in [0,1]^{n\times n} \colon \theta_{ii} = 0, \theta_{ij} = Q_{ab} = Q_{ba} \text{ for } (i,j) \in z^{-1}(a) \times z^{-1}(b), Q_{ab} \in [0,1], \text{ and } z \in [k]^n\right\}.$$

Then, the minimax convergence rate of the graphon estimation under stochastic block model is

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta_k} \mathrm{MSE}(\hat{\theta}, \theta) \gtrsim \frac{k^2}{n^2} + \frac{\log k}{n}.$$

2. **Under the Holder class.**

   The bounded Holder class is defined by $\mathcal{F}_\alpha(L) = \{f : 0 \le f \le 1, \|f\|_{\mathcal{H}_\alpha} \le L, f(x,y) = f(y,x) \text{ for } x \ge y\}$ where

$$\|f\|_{\mathcal{H}_\alpha} = \max_{j+k \le \lfloor \alpha \rfloor} \sup_{x,y \in \mathcal{D}} |\Delta_{jk} f(x,y)| + \max_{j+k = \lfloor \alpha \rfloor} \sup_{(x,y) \ne (x',y') \in \mathcal{D}} \frac{|\Delta_{jk} f(x,y) - \Delta_{jk} f(x',y')|}{(|x-x'| + |y-y'|)^{\alpha - \lfloor \alpha \rfloor}}.$$

   Here the derivatie operator is defined as $\Delta_{jk} f(x,y) = \partial^{j+k} f(x,y) / (\partial x)^j (\partial y)^k$.

   Then, the minimax convergence rate of the graphon estimation under the graphon $f \in \mathcal{F}_\alpha(L)$ is

$$\inf_{\hat{\theta}} \sup_{f \in \mathcal{F}_\alpha(L)} \sup_{\mathbb{P}_\xi} \mathrm{MSE}(\hat{\theta}, \theta) \gtrsim n^{-2\alpha/(\alpha+1)} + \frac{\log n}{n}.$$

   In fact, there is one more assumption on $\mathbb{P}_\xi$ to make sure that sampling points $\{\xi_i\}$ to well cover the interval $[0,1]$ for $\{f(\xi_i, \xi_j)\}$ representing the whole function f (See Gao et al. [2015] for more details).

I summarize convergence rates of different estimation methods depending on two different settings in the following table. As we can see, least square minimization achieves optimal convergence rates while universal singular value thresholding method is sub-optimal. This sub-optimal rate is the price one has to pay for the polynomial time algorithm.

Estimation method	Stochastic block model (with k block)	Holder class
Least square minimization [Gao et al., 2015]	$k^2/n^2 + \log k/n$	$n^{-2\alpha/(\alpha+1)} + \log n/n$
Universal singular value thresholding [Xu, 2018]	k/n	$n^{-2\alpha/(2\alpha+1)}$

Table 1: Convergence rates of different estimation methods

2 Extension to higher order networks: hypergraph community detection

There are some extensions of community detection to higher order [Ghoshdastidar and Dukkipati, 2015, 2017, Ahn et al., 2018, Chien et al., 2019]. Instead of considering adjacency matrix whose entries are 0 or 1, they consider m-th order symmetric tensor $\mathcal{A}_{i_1,\dots,i_m} \in [0,1]^{n \times \cdots \times n}$.

2.1 Problem setting

We consider the following random model. Let $\mathcal{V} = \{1, 2, \dots, n\}$ be a set of n nodes, and $z : [n] \to [k]$ be a partition of the nodes into k classes. A random weighted $m-$uniform hypergraph $\mathcal{A}_{i_1,\dots,i_m}$ is generated from

$$\mathbb{E}[\mathcal{A}_{i_1,\dots,i_m}] = \rho_n \Theta_{i_1,\dots,i_m}, \tag{3}$$

where ρ_n accounts for sparsity of the hypergraph. In particular, if $\rho_n = 1$ and $\mathcal{A}$ are independent Bernoulli random variable, (3) reduces to direct generalization to m-uniform hypergraph from (1) with stochastic block model.

$$(\xi_1, \dots, \xi_n) \sim \mathbb{P}_\xi, \quad \mathcal{A}_{i_1,\dots,i_m} | (\xi_{i_1}, \dots, \xi_{i_m}) \sim \text{Bernoulli}(\Theta_{i_1,\dots,i_m}), \text{ where } \Theta_{i_1,\dots,i_m} = f(\xi_1, \dots, \xi_m).$$

Ahn et al. [2018] assumes that $\Theta_{i_1,\dots,i_m} = p$ if $z(i_1) = \cdots = z_{i_m}$ and $\Theta_{i_1,\dots,i_m} = q$ otherwise where $1 > p > q > 0$. Chien et al. [2019] extends Ahn et al. [2018] to the case considering community relations and constructed minimax bound for misclassification error. To be specific, they define $\mathcal{K}_m$ as the set of all possible community relations under m-order. For example, when all $i_1, \dots, i_m$ are from the same cluster, they denote $r_1 = (m, 0, \dots, 0)$ and denote the only-1-different relation as $r_2 = (m-1, 1, 0 \dots, 0)$. In this way, they have $|\mathcal{K}_m|$ relations up to $r_{|\mathcal{K}_m|} = (1, \dots, 1)$, all different relation case. Based on this community relations, they assign probability as

$$\Theta_{i_1,\dots,i_m} = p_{r(i_1,\dots,i_m)},$$

where $\boldsymbol{p} = (p_1, \dots, p_{|\mathcal{K}_m|}) \in [0,1]^{|\mathcal{K}_m|}$ and $r\colon [n]^m \to [|\mathcal{K}_m|]$ such that $r(i_1, \dots, i_m) = l$ if the community relation of $i_1, \dots, i_m$ belongs to r_l. Ghoshdastidar and Dukkipati [2017] assumes the most general case such that

$$\Theta_{i_1,\dots,i_m} = \mathcal{Q}_{z(i_1),\dots,z(i_m)}, \tag{4}$$

where $\mathcal{Q} \in [0,1]^{k \times \cdots \times k}$ and $z\colon [n] \to [k]$ is a partition function.

Here, we stick to (4) for the purpose of the general cases.

2.2 Estimation method and consistency

Estimation procedure

Ghoshdastidar and Dukkipati [2017] focused on finding the partition of the vertices $z\colon [n] \to [k]$. For this purpose, we define a few terms. The degree of any mode $i \in [n]$ is the total weight of edges on which i is incident. For a collection of nodes $z^{-1}(a)$ for $a \in [k]$, we define its volume and its associativity are defined as

$$\mathrm{Vol}(z^{-1}(a)) = \sum_{i \in z^{-1}(a)} \deg(i), \qquad \mathrm{Assoc}(z^{-1}(a)) = \sum_{i_1,\dots,i_m \in z^{-1}(a)} \mathcal{A}_{i_1,\dots,i_m}.$$

The normalised associativity of a partition $z^{-1}(1), \dots z^{-1}(k)$ is given as

$$\text{N-Assoc}(z^{-1}(1), \dots z^{-1}(k)) = \sum_{i=1}^{k} \frac{\mathrm{Assoc}(z^{-1}(i))}{\mathrm{Vol}(z^{-1}(i))}. \tag{5}$$

They find cluster function z that maximises the normalized associativity (5). Ghoshdastidar and Dukkipati [2017] showed that maximization problem (5) is equivalent to tensor trace maximization (TTM) problem and relaxed optimization as

$$\hat{\boldsymbol{Z}} = \underset{\boldsymbol{Z} \in \mathbb{R}^{n \times k}\colon \boldsymbol{Z}^T \boldsymbol{Z} = \boldsymbol{I}}{\arg\max} \mathrm{Trace}(\mathcal{A} \times_1 \boldsymbol{Z}^T \cdots \times_m \boldsymbol{Z}^T). \tag{6}$$

They estimate the cluster function z running $k-$means on the rows of $\hat{\boldsymbol{Z}}$ in (6).

Theoretical guarantee

Define the multi-class 0-1 loss as

$$\mathrm{Err}(z, \hat{z}) = \min_{\sigma} \sum_{i=1}^{n} \mathbb{1}\{z(i) \neq \sigma(\hat{z}(i))\},$$

where z and $\hat{z}$ denote the true and the estimated partitions, respectively, and σ is any permutation on $[k]$. Let $\boldsymbol{A}_{ij} = \sum_{i_3,\dots,i_m=1}^{n} \mathcal{A}_{i,j,i_3,\dots,i_m}$ and $\boldsymbol{G} \in \mathbb{R}^{k \times k}$ such that $\boldsymbol{A}_{ij} = \boldsymbol{G}_{z^{-1}(i) z^{-1}(j)}$. Without loss of generality,

assume that the cluster sizes are $n_1 \geq n_2 \geq \cdots \geq n_k$. Define

$$\delta = \lambda_k(\boldsymbol{G}) \min_{1\leq i\leq d} \frac{n_{z(i)}}{\sum_{j=1}^n \boldsymbol{A}_{ij}} - \max_{1\leq i,j\leq n} \left| \frac{\boldsymbol{A}_{ii}}{\sum_{j=1}^n \boldsymbol{A}_{ij}} - \frac{\boldsymbol{A}_{jj}}{\sum_{j=1}^n \boldsymbol{A}_{ij}} \right|. \tag{7}$$

Then, we have the following consistency theorem.

Theorem 2.1. Define $d = \min_{1\leq i\leq n} \mathbb{E}[\deg(i)]$. Let δ be as defined in (7). Assume that there exists on absolute constant $C > 0$, such that, if $\delta > 0$ and

$$\delta^2 d > \frac{Ckn_1(\log n)^2}{n_k},$$

for all large n, then with probability $1 - o(1)$, the partitioning error for TTM is

$$\mathrm{Err}(z, \hat{z}) = \mathcal{O}\left(\frac{kn_1 \log n}{\delta^2 d}\right).$$

3 Extension to higher order networks: stochastic block model for multi-layer networks.

Several authors have studied the problem of detecting an underlying community structure in multi-layer networks[Paul et al., 2016, 2020, De Bacco et al., 2017, Wilson et al., 2017, Lei et al., 2020].

3.1 Problem setting

We consider an undirected multi0lyaer graph $G = \{V, E\}$, where the vertex set V consists of n vertices and the edge set E consists of edges of m different types of representing different relationships. We can view the multi-graph as a adjacency tensor $\mathcal{A} \in \{0,1\}^{n\times n\times m}$ such that $\mathcal{A}(:,:,i_3) = A^{(i_3)}$ for $i_3 \in [m]$ where $A^{(i_3)}$ denotes i_3-th adjacency matrix. The generating model is

$$\mathcal{A}_{i_1,i_2,i_3}|(z(i_1), z(i_2), i_3) \sim \text{Bernoulli}(\mathcal{Q}_{z(i_1),z(i_2),i_3}), \tag{8}$$

where $\mathcal{Q} \in [0,1]^{k\times k\times m}$ and $z\colon [n] \to [k]$ is a partition function.

Paul et al. [2016] assumes additive logistic model on (8) such that

$$\text{logit}(\mathcal{Q}_{z(i_1),z(i_2),i_3}) = \pi_{z(i_1),z(i_2)} + \beta_{i_3}$$

for some $\pi \in \mathbb{R}^{k\times k}$ and $\beta \in \mathbb{R}^m$. Paul et al. [2016] estimates parameters π and β based on M-estimation. They relax this restricted model later to (8) [Paul et al., 2020] using matrix factorization on each adjacency matrix $\mathcal{A}(:,:,i_3)$.

3.2 Estimation method

I will briefly compare the estimation method for the model (8) in the following table.

Paper	Estimation method
Paul et al. [2020]	Spectral matrix factorization on each adjacency matrix
De Bacco et al. [2017]	EM algorithm assuming poisson distribution in (8)
Wilson et al. [2017]	Minimization of the extended modularity score of the partition
Lei et al. [2020]	Minimization of the least square error $\sum_{i_3\in[m]} \sum_{1\leq i_1\neq i_3\leq n} (\mathcal{A}_{i_1,i_2,i_3} - \mathcal{Q}_{z(i_1),z(i_2),i_3})^2$.

Table 2: Caption

References

Kwangjun Ahn, Kangwook Lee, and Changho Suh. Hypergraph spectral clustering in the weighted stochastic block model. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):959–974, 2018.

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

Sourav Chatterjee et al. Matrix estimation by universal singular value thresholding. *Annals of Statistics*, 43 (1):177–214, 2015.

I Eli Chien, Chung-Yi Lin, and I-Hsiang Wang. On the minimax misclassification ratio of hypergraph community detection. *IEEE Transactions on Information Theory*, 65(12):8095–8118, 2019.

Caterina De Bacco, Eleanor A Power, Daniel B Larremore, and Cristopher Moore. Community detection, link prediction, and layer interdependence in multilayer networks. *Physical Review E*, 95(4):042317, 2017.

Chao Gao, Yu Lu, Harrison H Zhou, et al. Rate-optimal graphon estimation. *Annals of Statistics*, 43(6): 2624–2652, 2015.

Debarghya Ghoshdastidar and Ambedkar Dukkipati. A provable generalized tensor spectral method for uniform hypergraph partitioning. In *International Conference on Machine Learning*, pages 400–409. PMLR, 2015.

Debarghya Ghoshdastidar and Ambedkar Dukkipati. Uniform hypergraph partitioning: Provable tensor methods and sampling techniques. *The Journal of Machine Learning Research*, 18(1):1638–1678, 2017.

Jing Lei, Kehui Chen, and Brian Lynch. Consistent community detection in multi-layer network data. *Biometrika*, 107(1):61–73, 2020.

Subhadeep Paul, Yuguo Chen, et al. Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *Electronic Journal of Statistics*, 10(2):3807–3870, 2016.

Subhadeep Paul, Yuguo Chen, et al. Spectral and matrix factorization methods for consistent community detection in multi-layer networks. *The Annals of Statistics*, 48(1):230–250, 2020.

James D Wilson, John Palowitch, Shankar Bhamidi, and Andrew B Nobel. Community extraction in multilayer networks with heterogeneous community structure. *The Journal of Machine Learning Research*, 18(1):5458–5506, 2017.

Jiaming Xu. Rates of convergence of spectral methods for graphon estimation. In *International Conference on Machine Learning*, pages 5433–5442. PMLR, 2018.