Joint Embedding of Graphs

Shangsi Wang

March 20, 2016

1 Theory

Consider a 1-dimensional Multiple Random Eigen Graph model, that is

$$P_i = \lambda_i h_1 h_1^T$$

where λ_i follows an unknown distribution $\mathbb{F}$. In this paper, we consider the task of estimating h_1, and let

$$\rho(A_i, h) = ||A_i - \langle A_i, hh^T \rangle hh^T||_F^2$$

$$D_m(h, h_1) = \frac{1}{m} \sum_{i=1}^{m} \rho(A_i, h)$$

$$D(h, h_1) = E(\rho(A_i, h))$$

Assume,

$$\hat{h}_1^m = \underset{||h||=1}{\operatorname{argmin}} D_m(h, h_1) \tag{1}$$

This is an alternative formulation of 1-dimensional joint embedding of graphs. If we define,

$$(\hat{\Lambda}, \hat{h}_1^m) = \underset{\Lambda, ||h_1||=1}{\operatorname{argmin}} \sum_{i=1}^{m} \|A_i - \Lambda_i h_1 h_1^T\|^2$$

Then, $\hat{h}_1^m$ in both problems will be the same and $\hat{\Lambda}_i = \langle A_i, \hat{h}_1^m \hat{h}_1^{mT} \rangle$. We consider the formulation of equation 1.

Theorem 1.1 *If $D(h, h_1)$ has a unique global minimum at h', then $\hat{h}_1^m$ converges almost surely to h' as m goes to infinity, that is*

$$\hat{h}_1^m \overset{a.s.}{\to} h'$$

Proof The first part of proof is showing that $|D_n(h, h_1) - D(h, h_1)|$ converges uniformly to 0, that is

$$\sup_h |D_m(h, h_1) - D(h, h_1)| \overset{a.s.}{\to} 0$$

Clearly, by strong law of large numbers $D_m(h, h_1) \overset{a.s.}{\to} D(h, h_1)$ for all h. To establish uniform convergence, we will use a stochastic version of Arzelà–Ascoli theorem. In particular, we need to show that for almost all sample sequences

$\{D_m(h,h_1)\}_{m=1}^{\infty}$ is equicontinuous, that is there exists a set $M \subset \Omega$ where $P(M)=1$, fix an $\epsilon > 0$, $\exists\delta, \forall\omega \in M, \exists m(\omega)$

$$\sup_{||s-t||<\delta} |D_m(s,h_1,\omega) - D_m(t,h_1,\omega)| \leq \epsilon$$

for all $m > m(\omega)$. The condition is hard to verify in general; however, we could take advantage of the fact that there is only a finite number possibilities for each adjacency matrix. For an adjacency matrix $A \in \{0,1\}^{n\times n}$, $\rho(A,h)$ is a continuous function in h which lies in a compact set; therefore, there exists $\delta(A)$ such that

$$\sup_{\|s-t\|<\delta(A)} |\rho(A,s) - \rho(A,t)| \leq \epsilon$$

Let $\delta = \inf_A \delta(A)$, then

$$\begin{aligned}
|D_m(s,h_1,\omega) - D_m(t,h_1,\omega)| &= |\frac{1}{m}\sum_{i=1}^{m}\rho(A_i,s) - \frac{1}{m}\sum_{i=1}^{m}\rho(A_i,t)| \\
&\leq \frac{1}{m}\sum_{i=1}^{m}|\rho(A_i,s) - \rho(A_i,t)| \\
&\leq \epsilon
\end{aligned}$$

This shows $\{D_m(h,h_1)\}_{m=1}^{\infty}$ is equicontinuous on Ω. Together with the fact that parameter space is compact and $D_m(h,h_1)$ converges almost surely to $D(h,h_1)$, we incur Arzela-Ascoli Theorem and claim that

$$\sup_h |D_m(h,h_1) - D(h,h_1)| \overset{a.s.}{\to} 0$$

To prove the claim of theorem, we use a slightly modified version of proof for theorem 5.2.1 in Bickel and Doksum. By definition, we have $D_m(\hat{h}_1^m,h) \leq D_m(h',h)$ and $D(h',h) \leq D(\hat{h}_1^m,h)$. Using these two inequalities we have

$$D_m(h',h) - D(h',h) \geq D_m(\hat{h}_1^m,h) - D(h',h) \geq D_m(\hat{h}_1^m,h) - D(\hat{h}_1^m,h)$$

Therefore,

$$|D_m(\hat{h}_1^m,h) - D(h',h)| \leq \max(|D_m(h',h) - D(h',h)|, |D_m(\hat{h}_1^m,h) - D(\hat{h}_1^m,h)|)$$

This implies

$$|D_m(\hat{h}_1^m,h) - D(h',h)| \leq \sup_h |D_m(h,h_1) - D(h,h_1)|$$

Hence, $|D_m(\hat{h}_1^m,h) - D(h',h)|$ must converge almost surely to 0. If $\hat{h}_1^m$ does not converge almost surely h', $||\hat{h}_1^m - h'|| \geq \epsilon$ for some ϵ and infinitely many ms. Since h' is the unique global minimum, $|D(\hat{h}_1^m,h) - D(h',h)| > \epsilon'$ for infinitely many ms. This contradicts the fact that

$$|D(\hat{h}_1^m,h) - D(h',h)| \leq |D_m(\hat{h}_1^m,h) - D(h',h)| + |D_m(\hat{h}_1^m,h) - D(\hat{h}_1^m,h)| \to 0$$

This concludes that $\hat{h}_1^m$ converges almost surely h'.

Theorem 1.2 *Consider a 1-dimensional MREG, that is $P_i = \lambda_i h_1 h_1^T$ with h_1 being norm 1 and $\lambda_i \sim F$. If h' is the global minimizer of $D(h, h_1)$, then*

$$||h' - h_1|| \leq \frac{2E(\lambda_i)}{E(\lambda_i^2)(h_1^T h')^2}$$

Proof The first part of proof is showing that h' is the eigenvector corresponds to the largest eigenvalue of $E(\langle A_i, h'h'^T\rangle A_i)$. Then, we show $E(\langle A_i, h'h'^T\rangle A_i)$ is close to a properly scaled $E(P_i)$. To conclude, we apply Davis-Kahan theorem to the top eigenvector of both matrices and get the desired result. First, we notice that

$$\begin{aligned}
\min_{||h||=1} D(h, h_1) &= \min_{||h||=1} E(\|A_i - \langle A_i, hh^T\rangle hh^T\|^2) \\
&= \min_{||h||=1} E(\langle A_i, A_i\rangle - \langle A_i, hh^T\rangle^2) \\
&= E(\langle A_i, A_i\rangle) - \max_{||h||=1} E(\langle A_i, hh^T\rangle^2)
\end{aligned}$$

Therefore,

$$h' = \underset{||h||=1}{\text{argmin}} D(h, h_1) = \underset{||h||=1}{\text{argmax}} E(\langle A_i, hh^T\rangle^2)$$

Taking the derivative of $E(\langle A_i, hh^T\rangle^2) + c(h^T h - 1)$ with respect to h, we have

$$\begin{aligned}
\frac{\partial E(\langle A_i, hh^T\rangle^2) + c(h^T h - 1)}{\partial h} &= E(\frac{\partial \langle A_i, hh^T\rangle^2}{\partial h}) + 2ch \\
&= E(\langle A_i, hh^T\rangle A_i)h + 2ch
\end{aligned}$$

Set it to 0,

$$\begin{aligned}
E(\langle A_i, h'h'^T\rangle A_i)h' &= -2ch' \\
&= (h'^T E(\langle A_i, h'h'^T\rangle A_i)h')h' \\
&= E(\langle A_i, h'h'^T\rangle^2)h'
\end{aligned}$$

It implies h' is an eigenvector of $E(\langle A_i, h'h'^T\rangle A_i)$ and the corresponding eigenvalue is $E(\langle A_i, h'h'^T\rangle^2)$. Furthermore, $E(\langle A_i, h'h'^T\rangle^2)$ must be the eigenvalue with the largest magnitude. Assume not, then there exists a h'' with norm 1 such that

$$|h''^T E(\langle A_i, h'h'^T\rangle A_i)h''| = |E(\langle A_i, h'h'^T\rangle\langle A_i, h''h''^T\rangle)| > E(\langle A_i, h'h'^T\rangle^2)$$

However, by Cauchy-Schwarz inequality we must have

$$E(\langle A_i, h''h''^T\rangle^2)E(\langle A_i, h'h'^T\rangle^2) > |E(\langle A_i, h'h'^T\rangle\langle A_i, h''h''^T\rangle)|^2$$

This implies $E(\langle A_i, h''h''^T\rangle^2) \geq E(\langle A_i, h'h'^T\rangle^2)$ which contradicts the definition of h'. This concludes that h' is the eigenvector corresponds to the largest eigenvalue of $E(\langle A_i, h'h'^T\rangle A_i)$.

Next, we compute $E(\langle A_i, h'h'^T\rangle A_i)$.

$$\begin{aligned}
E(\langle A_i, h'h'^T\rangle A_i|P_i) &= E(\langle A_i - P_i, h'h'^T\rangle (A_i - P_i)|P_i) + E(\langle A_i - P_i, h'h'^T\rangle P_i|P_i) \\
&\quad + E(\langle P_i, h'h'^T\rangle (A_i - P_i|P_i) + E(\langle P_i, h'h'^T\rangle P_i|P_i) \\
&= E(\langle A_i - P_i, h'h'^T\rangle (A_i - P_i)|P_i) + \lambda_i (h_1^T h')^2 P_i \\
&= 2h'h'^T * P_i * (J - P_i) - DIAG(h_1 h_1^T * P_i * (J - P_i)) + \lambda_i (h_1^T h')^2 P_i \\
&= 2h'h'^T * P_i * (J - P_i) - DIAG(h'h'^T * P_i * (J - P_i)) + \lambda_i (h_1^T h')^2 P_i
\end{aligned}$$

Here, DIAG() means only keep the diagonal of matrix, $*$ means the hadmard product and J is a matrix of all ones. Using the fact that $P_i = \lambda_i h_1 h_1^T$, we have

$$\begin{aligned}
E(E(\langle A_i, h'h'^T\rangle A_i|P_i) - \lambda_i (h_1^T h')^2 P_i) &= E(\langle A_i, h'h'^T\rangle A_i) - E(\lambda_i^2)(h_1^T h')^2 h_1 h_1^T \\
&= E(2h'h'^T * P_i * (J - P_i) - DIAG(h'h'^T * P_i * (J - P_i)))
\end{aligned}$$

If we consider the norm difference between $E(\langle A_i, h'h'^T\rangle A_i)$ and $E(\lambda_i^2)(h_1^T h')^2 h_1 h_1^T$, we have

$$\begin{aligned}
||E(\langle A_i, h'h'^T\rangle A_i) - E(\lambda_i^2)(h_1^T h')^2 h_1 h_1^T|| &= ||E(2h'h'^T * P_i * (J - P_i) - DIAG(h'h'^T * P_i * (J - P_i)))|| \\
&\leq E(||2h'h'^T * P_i * (J - P_i) - DIAG(h'h'^T * P_i * (J - P_i))||) \\
&\leq E(||2h'h'^T * P_i * (J - P_i)||) \\
&\leq E(||2h'h'^T * P_i||) \\
&\leq 2E(\lambda_i)||h'h'^T * h_1 h_1^T|| \\
&= 2E(\lambda_i)
\end{aligned}$$

Notice that the only non-zero eigenvector of $E(\lambda_i^2)(h_1^T h')^2 h_1 h_1^T$ is h_1 and the corresponding eigenvalue is $E(\lambda_i^2)(h_1^T h')^2$. Applying Davis-Kahan theorem to eigenvector corresponding to the largest eigenvalue of matrices $E(\langle A_i, h'h'^T\rangle A_i)$ and $E(\lambda_i^2)(h_1^T h')^2 h_1 h_1^T$, we conclude

$$||h' - h_1|| \leq \frac{2E(\lambda_i)}{E(\lambda_i^2)(h_1^T h')^2}$$

Remark In proof of theorem 1, we require h' to be the unique global minimizer of $D(h, h_1)$ or global maximizer $E(\langle A_i, hh^T\rangle^2)$ subject to being norm 1. However, the global optimizer is definitely not unique due to symmetry in the sign of h, that is $D(h, h_1) = D(-h, h_1)$ for any h. This problem can be fixed by forcing an orientation of $\hat{h}_1^m$ or stating that the convergence is up to a sign flip. It is also possible that there are multiple global minimizers of $D(h, h_1)$ which are not sign flip of each other. In this case, theorem 1 will not apply. We are only certain that when all graphs are from Erdos-Renyi random graph model, the global minimizer is unique up to a sign flip.

Remark To see an application of theorem 2, we consider Erdos-Renyi graphs with 100 vertices and edge probability 0.5. Under this setting, theorem 2 implies $||h' - h_1|| \in [0, 0.04] \cup [1.28, 1.52]$. The second interval is disturbing. It is due to the fact that when $h_1^T h'$ is small the bound is useless. We provide some insights why the second interval is there and how we can get rid of it with more

assumptions. If we take a closer look on $E(\langle A_i, hh^T\rangle^2)$,

$$\begin{aligned} E(\langle A_i, hh^T\rangle^2) &= E(\langle P_i, hh^T\rangle^2) + 2E(\langle P_i, hh^T\rangle\langle A_i - P_i, hh^T\rangle) + E(\langle A_i - P_i, hh^T\rangle^2) \\ &= E(\langle P_i, hh^T\rangle^2) + E(\langle A_i - P_i, hh^T\rangle^2) \\ &= E(\lambda_i^2)(h_1^T h)^4 + E((h^T(A_i - P_i)h)^2) \end{aligned}$$

Therefore,

$$h' = \underset{||h||=1}{\operatorname{argmax}} E(\lambda_i^2)(h_1^T h)^4 + E((h^T(A_i - P_i)h)^2)$$

We see $E(\lambda_i^2)(h_1^T h)^4$ is clearly maximized when $h = h_1$; however, the noise term $E((h^T(A_i - P_i)h)^2)$ is generally not maximized at $h = h_1$. If we assume n is large, we can apply concentration inequality to $(h^T(A_i - P_i)h)^2$ and have an upper bound on $E((h^T(A_i - P_i)h)^2)$. If we further assume A_i is not too sparse, that is $E(\lambda_i^2)$ grows with n fast enough, then the sum of these two terms are dominated by the first term. This provides a way to have a lower bound on $h_1^T h'$. We may then remove the second interval. In general, if n is small, the noise term may cause h' differs with h_1 by a significant amount. We focus on the case that n is fixed in this paper. The case that n goes to infinity is considered in papers ***.

Remark The two theorems above only concern the estimation of h_1, but not λ_i. To estimate λ_i, we have

$$\hat{\lambda}_i^m = \langle A_i, \hat{h}_1^m \hat{h}_1^{mT}\rangle$$

If n is fixed, it is impossible to prove any consistency result on estimation of λ_i due to the fact we only observe one finite graph A_i associated with λ_i. When m goes to infinity, we can apply theorem 1

$$\hat{\lambda}_i^m \to \langle A_i, h'h'^T\rangle = h'^T A_i h'$$

Then applying theorem 2 and utilizing the fact that $h^T A_i h$ is continuous in h, we may have a upper bound on $|\hat{\lambda}_i^m - h_1^T A_i h_1|$. In general, $h_1^T A_i h_1$ is concentrated about λ_i with high probability.