Reading Group Notes

Joshua Agterberg

June 9, 2020

1 Notation

For ease of comparision, I'll use the same notation.

Throughout, let:

- $A \in \{0,1\}^{n \times n}$ be the symmetric adjacency matrix for a graph.
- B is the $K \times K$ symmetric block connectivity matrix.
- $Z \in \{0,1\}^{n \times K}$ is the matrix satisfying $Z_{ik} = 1$ if vertex i belongs to community k and 0 otherwise.
- For a given matrix M define the laplacian $\mathcal{L}(M) := D^{-1/2} M D^{-1/2}$, where D is the diagonal matrix $\text{Diag}(M1)$ (1 is the vector of all ones) (i.e. D is the degree matrix).
- Define $\Delta := \max_i \sum_{j=1}^n P_{ij}$; i.e. Δ is the maximum expected degree
- $\delta = \min_i \sum_{j=1}^n P_{ij}$.
- Define $\tau_n = \delta/\Delta$, the min expected degree divided by the max expected degree.
- I will use ρ_n as the sparsity parameter (see e.g. Cape et al. (2019); Tang et al. (2017); Tang and Priebe (2018))
- We let $P := \mathbb{E}A$. For a stochastic blockmodel, we note that $P = ZBZ^\top$ with self-loops and $ZBZ^\top - \text{Diag}(ZBZ^\top)$ without self-loops. I won't worry too much about self-loops here, but sometimes they can be pretty important (e.g. Han et al. (2019)).

Finally, when I refer to the Davis-Kahan Theorem, I mean the following.

Theorem 1 (Davis-Kahan, some variant)**.** *Suppose $\hat{M} = M + E$. Suppose M is rank d. Let $\hat{U}$ be the $n \times d$ matrix whose columns are the top d eigenvectors of $\hat{M}$ and similarly for U. Then*

$$\inf_{W \in \mathbb{O}(d)} ||U - \hat{U}W||_{2,F} \leq \frac{||E||_{2,F}}{\gamma},$$

where γ is such that the eigenvalues of M lie in an interval (a,b) and $n-d$ bottom eigenvalues of $\hat{M}$ lie outside the interval $(a-\gamma, b+\gamma)$.

If you do not know this already, lots of applications of Davis-Kahan really only need information about the order of the terms appearing on the right hand side.

For example, when everything is nice and dense (i.e. $\rho_n \equiv 1$), we have

- $||E||_2 \lesssim \sqrt{n}$ with high probability
- $\lambda_1, ..., \lambda_d(P) = \Theta(n)$ with high probability,

- $\lambda_{d+1}(P), ..., \lambda_n(P) \equiv 0$

Weyl's inequality shows that

$$|\lambda_i(P) - \lambda_i(A)| \leq ||E||_2 \lesssim \sqrt{n},$$

so that the bottom $n-d$ eigenvalues of A are of smaller order than $\sqrt{n}$ and the top d eigenvalues of A are of order n. Then the $n-d$ bottom eigenvalues of A lie outside the interval $(\lambda_d(P) - Cn, \lambda_1(P) + Cn)$ for some $C < \lambda_d(P)$ (since the order of this interval is n), so we apply the theorem above with $\gamma = Cn$. The theorem then reads that

$$\inf_{W \in \mathbb{O}(d)} ||U - \hat{U}W||_2 \lesssim \frac{1}{\sqrt{n}}.$$

Subsequent analyses focus on explicitly examining the dependence on the hidden constants (eigenvalues of P, number of blocks, rank, etc.) in the right hand side as well as allowing for sparsity.

2 Notes on Rohe et al. (2011)

The high-level overview of the paper is a study of the performance of spectral clustering using the leading eigenvectors of $\mathcal{L}(A)$. Their main idea is to combine a concentration result on the Laplacian with the Davis-Kahan Theorem, showing that (in Frobenius norm), the empirical eigenvectors are close to the true eigenvectors with high probability. They also use one of my favorite tricks when they square the matrices and note that the eigenvectors of the squared matrices are the same as the original eigenvectors – I used this trick this most recent December as part of a proof, even though I never ended up needing that proof.

Their first main result is a concentration bound.

Theorem 2 (Theorem 2.1)**.** *If $\tau_n^2 \log(n) > 2$ for all n, then*

$$||\mathcal{L}(A)^2 - \mathcal{L}(P)^2||_F = o\left(\frac{\log(n)}{\tau^2 n^{1/2}}\right)$$

almost surely.

This is (essentially) just a concentration bound for the (squared) empirical Laplacian versus the true Laplacian in Frobenius norm. The theorem in L. Lu and X. Peng (2013) reads as

$$||\mathcal{L}(A) - \mathcal{L}(P)|| \lesssim \delta^{-1/2}$$

with high probability provided that $\delta \gg \log(n)$. In Rohe et al. (2011), the authors do not assume any sparsity, so they define $\tau_n = \delta/n$. Therefore, their τ_n is of order δ/n, so that their assumption reads that

$$\delta \geq C\frac{n}{\sqrt{\log(n)}},$$

which is only slightly slower than order n. So their requirement is much more stringent than Lu and Peng's. With this notation, their result reads as $o\left(\frac{n^{3/2}\log(n)}{\delta^2}\right)$, which, when $\delta > \frac{n}{\sqrt{\log(n)}}$ gives a bound of order $o\left(\frac{\log^2(n)}{\sqrt{n}}\right)$, which is a factor of $\log^2(n)$ worse under the same assumptions as L. Lu and X. Peng (2013) (though it is an $o(\cdot)$ and not an $O(\cdot)$). However, we should note that this was written before L. Lu and X. Peng (2013) came out, so they had to do it from scratch.

Their main result is as follows.

Theorem 3 (Theorem 3.1). *If $\mathcal{M}$ is the set of misclustered nodes, then*

$$|\mathcal{M}| = o\left(\frac{n_{max}\log^2(n)}{\lambda^2_{k_n}(\mathcal{L}(P))\tau_n^4 n}\right),$$

provided that $\frac{\log^2(n)}{n^{1/2}} = O(\lambda^2_{k_n}(\mathcal{L}(P)$ *and* $\tau^2\log(n) > 2$, *where* $n_{\max}$ *is the maximum number of nodes in each community.*

In the special case $B = \rho_n B_0$ for some (constant) rank K matrix B_0, we have that $\lambda_{k_n}(\mathcal{L}(P)) = \Theta(1)$. Recall $\tau_n = \frac{\delta}{n}$, where δ is the minimum expected degree. If $n_{\max} \asymp n$, then this bound reads as

$$\begin{aligned} o\left(\frac{n^4\log^2(n)}{\delta^4}\right) &= o\left(\frac{n^4\log^2(n)}{(n\rho_n)^4}\right) \\ &= o\left(\frac{\log^2(n)}{\rho_n^4}\right), \end{aligned}$$

so that the average number of miclustered nodes satisfies

$$\frac{1}{n}|\mathcal{M}| = o\left(\frac{\log^2(n)}{n\rho_n^4}\right).$$

General notes:

- The proof of their theorem combines the Davis-Kahan Theorem and their concentration results using the squared values.
- The requirement on the rate of convergence of $\lambda^2_{k_n}(\mathcal{L}(P))$ to zero is a requirement on the eigengap implicitly required by applying the Davis-Kahan Theorem. In other words, if this rate of convergence happens, all the top eigenvalues of $\mathcal{L}(A)$ will be of much larger order so that the Davis-Kahan Theorem can be applied. (Recall that eigenvalues of the Laplacian are between 0 and 2)
- Finally, the appearance of the $n_{\max}$ in their theorem comes from bounding the size of $\mathcal{M}$ directly.
- Slightly tighter results can be found now using stronger concentration results and $2 \to \infty$ type bounds, but I think the key point is that their results were amongst the first to combine these tools in a meaningful way.

Finally, I really like their Lemma 3.1, which shows that the eigenvectors of $\mathcal{L}(P)$ contain the community information.

3 Notes on Lei and Rinaldo (2015)

Again, the high level overview is what happens when applying spectral clustering to the eigenvectors of the adjacency matrix. Their results focus highly on the sparse regime wherein they assume an explicit sparsity parameter (they say α_n, but I will say ρ_n to use the slightly more common notation). I will not worry so much about their results on the degree-corrected stochastic blockmodel so as to compare to Rohe et al. (2011).

One of their main results concerns a concentration bound.

Theorem 4 (Theorem 5.2). *Suppose* $n\max_{i,j} P_{ij} \leq \Delta$ *for* $\Delta \geq c_0\log(n)$ *for some* $c_0 > 0$. *Then for any* r *there exists* $C = C(r, c_0)$ *such that*

$$||A - P|| \leq C\sqrt{\Delta}$$

with probability at least $1 - n^{-r}$.

As far as I know, this has the weakest assumptions and the strongest concentration (though L. Lu and X. Peng (2013) only requires $\Delta \gg \log^4(n)$, which is not that much less sparse). Their bound is still used often; it was used quite recently in a Biometrika paper with Liza Levina.

Their result on the spectral clustering performance is as follows.

Theorem 5. *Suppose $P = ZBZ^\top$ is of rank K, and suppose $\max_{k,l} B_{kl} \leq C\rho_n$ for some $\rho_n \geq c_0 \log(n)/n$. Then there exists an absolute constant c such that if*

$$(2+\varepsilon)\frac{Kn\rho_n}{\lambda_d(P)} < c,$$

then

$$\sum_{k=1}^{K} \frac{|S_k|}{n_k} \leq c^{-1}(2+\varepsilon)\frac{Kn\rho_n}{\lambda_d(P)},$$

where S_k are the sets of incorrectly clustered vertices.

This theorem is in the same spirit to the previous theorem; it gives a bound on the number of misclustered vertices in terms of stochastic blockmodel parameters. The ε is just for theoretical reasons but really makes no material difference.

They also present a corollary for interpretation.

Corollary 6 (Corollary 3.2)**.** *Suppose $B = \rho_n B_0$ for some ρ_n satisfying $n\rho \geq C\log(n)$, and suppose $\max_{k,l}(B_0)_{k,l} = 1$. Then there exists an absolute constant c such that if*

$$(2+\varepsilon)\frac{Kn}{n_{min}^2 \lambda_K(B_0)^2 \rho_n} < c$$

then with probability $1 - n^{-1}$,

$$\frac{1}{n}|\mathcal{M}| \leq c^{-1}(2+\varepsilon)\frac{Kn_{\max}}{n_{\min}^2 \lambda_K(B_0)^2 \rho_n}$$

The remark under this Corollary shows that to match their result to Rohe et al. (2011), they only require $\rho_n > \frac{log(n)}{n}$, whereas Rohe et al. (2011) requires $\rho_n > \frac{1}{\log(n)}$ (though this was a little buried earlier).

In particular, under the special conditions we analyzed in the previous section with approximately equal cluster sizes $n_{\max} \asymp n_{\min}$, we see that

$$\frac{1}{n}|\mathcal{M}| \lesssim \frac{1}{n\rho_n},$$

which is better then Rohe et al. (2011) by a factor of $\frac{\log^2(n)}{\rho_n^3}$. Recall we often examine the regime in which $\rho_n \to 0$, in which case this bound does much better.

4 Comparisons and Talking Points

- Two of my favorite lemmas are Lemma 3.1 in Rohe et al. (2011) and Lemma 2.1 in Lei and Rinaldo (2015), in which the authors compare the eigenstructure of the population matrices. This to me clarifies a few things in particular:
  - It shows why we care about the rows of the leading eigenvectors for community recovery in the SBM (and shows why it is a nice jumping-off point for the scaled eigenvectors in the analysis of RDPGs)

  - K-means will eventually do great clustering if n is sufficiently large
  - If we expect a row-wise CLT, the scaling will be by n in the dense case ($\rho_n \equiv 1$) (indeed, this is the scaling in both Cape et al. (2019) and Tang and Priebe (2018))
  - (Exercise) staring at the proof of Lemma 3.1 in Rohe et al. (2011) and Lemma 2.1 in Lei and Rinaldo (2015) shows that the eigenvector differences of the Laplacian are the same order and magnitude as in the adjacency matrix. In other words, one can show that the eigenvectors of the Laplacian have similar properties to those of $ZBZ^\top$ in community recovery.

- How do these bounds translate to $2 \to \infty$ or entrywise max bounds?
- Why might we care about $2 \to \infty$ or entrywise max bounds?
- How much do these results depend on the full rank assumption of B?
- Lei and Rinaldo (2015) also studies the DCSBM. For what types of models will we want to scale by the square roots of the eigenvalues?
- How do current entrywise eigenvector bounds extend and refine these bounds (both the results and proofs)?
- How do these results inform our understanding of more complicated models like the MMSBM, DCMMSBM, or (G)RDPG? (recall the MMSBM has the form $ZBZ^\top$ like the SBM, only Z is a matrix whose rows sum to 1, and the DCMMSBM is of the form $DZBZ^\top D$, where D is the diagonal degree-correction parameter. The (G)RDPG is just the property that P is a fixed rank).

References

J. Cape, M. Tang, and C. E. Priebe. Signal-plus-noise matrix models: eigenvector deviations and fluctuations. *Biometrika*, 106(1):243–250, Mar. 2019. ISSN 0006-3444. doi: 10.1093/biomet/asy070. URL https://academic.oup.com/biomet/article/106/1/243/5280315.

X. Han, Q. Yang, and Y. Fan. Universal Rank Inference via Residual Subsampling with Application to Large Networks. *arXiv:1912.11583 [math, stat]*, Dec. 2019. URL http://arxiv.org/abs/1912.11583. arXiv: 1912.11583.

L. Lu and X. Peng. Spectra of edge-independent random graphs. *Electronic Journal of Combinatorics*, 20, 2013.

J. Lei and A. Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, Feb. 2015. ISSN 0090-5364, 2168-8966. doi: 10.1214/14-AOS1274. URL https://projecteuclid.org/euclid.aos/1418135620.

K. Rohe, S. Chatterjee, and B. Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, Aug. 2011. ISSN 0090-5364. doi: 10.1214/11-AOS887. URL http://arxiv.org/abs/1007.1684. arXiv: 1007.1684.

M. Tang and C. E. Priebe. Limit theorems for eigenvectors of the normalized Laplacian for random graphs. *The Annals of Statistics*, 46(5):2360–2415, Oct. 2018. ISSN 0090-5364, 2168-8966. doi: 10.1214/17-AOS1623. URL https://projecteuclid.org/euclid.aos/1534492839.

M. Tang, J. Cape, and C. E. Priebe. Asymptotically efficient estimators for stochastic blockmodels: the naive MLE, the rank-constrained MLE, and the spectral. *arXiv:1710.10936 [stat]*, Oct. 2017. URL http://arxiv.org/abs/1710.10936. arXiv: 1710.10936.