Semidefinite Relaxation for Clustering and Community Detection

Bowei Yan

Department of Statistics and Data Sciences
University of Texas at Austin

Joint work with Purnamrita Sarkar

March 6, 2017@ HKUST

Mixture of Gaussians

Consider the Gaussian mixture model:

$$p(\boldsymbol{\theta}) = \sum_{k=1}^{r} \phi_i \mathcal{N}(\mu_i, \Sigma)$$

Introduce the latent variable $Z_{ik} = 1$(point i belongs to cluster k),

$$Y_i = \sum_{k=1}^{r} \mu_k Z_{ik} + U_i, U_i \sim \mathcal{N}(0, \Sigma). \tag{1}$$

GOAL Learn the latent labels Z.

k-means for clustering

k-means [Mac+67] minimizes the following loss function.

$$\sum_{k=1}^{r} \sum_{i:Z_{ik}=1} \|Y_i - \widehat{\boldsymbol{\mu}}_k\|^2$$

As it turns out, this can be reformalized as the following form [OW93].

$$\sum_{k=1}^{r} \sum_{i:Z_{ik}=1} \|Y_i - \widehat{\boldsymbol{\mu}}_k\|^2 = -\frac{r}{n}\text{trace}(YY^TZZ^T) + const$$

Semi-definite Relaxation for equal size clustering

- The problem is NP-hard.
- *Lifting*, or semi-definite relaxation: a technique dating back to max-cut [GW94].
- Let $X = ZZ^T \in \mathbb{R}^{n \times n}$, $X_{ij} = 1$ if and only if i, j belong to the same cluster.
- Consider the following SDP:

$$
\begin{aligned}
\max_{X} \quad & \langle YY^T, X \rangle \\
s.t. \quad & X \succeq 0, 0 \leq X \leq 1, X\mathbf{1} = \frac{n}{r}\mathbf{1}, \mathrm{diag}(X) = 1
\end{aligned}
$$

The "Kernel Trick"

Define the similarity among points by a kernel

$$K(i,j) = f(\|Y_i - Y_j\|^2)$$

The clustering framework

1. Transformation of K:

Kernel SVD	$\hat{X} = K$
K-PCA [SSM98]	$\hat{X} = K - K11^T/n - 11^TK/n + 11^TK11^T/n^2$;
Spectral clustering [NJW+02]	$\hat{X} = D^{-1/2}KD^{-1/2}$ where $D = \mathrm{diag}(K1_n)$;
SDP [**Y**S16]	$\hat{X} = \arg\max_{X \in \mathcal{F}} \quad \langle K, X \rangle$.

2. Do k-means on the r leading singular vectors V of $\hat{X}$.

Main Result - Kernel clustering via SDP

Theorem

Let $d_{k\ell} = \|\mu_k - \mu_\ell\|$*. Define the separation in kernel matrix as*

$$\gamma_{k\ell} := f(2\sigma_k^2) - f(d_{k\ell}^2 + \sigma_k^2 + \sigma_\ell^2); \quad \gamma_{\min} := \min_{\ell \neq k} \gamma_{k\ell}. \tag{2}$$

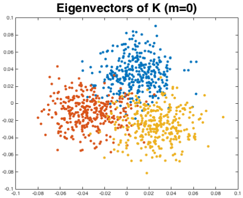
When $d_{k\ell}^2 > |\sigma_k^2 - \sigma_\ell^2|, \forall k \neq \ell$*, and* $\gamma_{\min} = \Omega\left(\sqrt{\frac{\log d}{d}}\right)$*. Denote* $X_0 = ZZ^T$*, then with probability going to 1,*

$$\|X_0 - \hat{X}\|_1 = o(1).$$

Bounding the misclassification rate

Combined with Davis-Kahan Theorem [YWS15], we can bound the number of mis-classified nodes in both cases [YS16].

	K-SVD	SDP
# mis-classified nodes	$O_P\left(\frac{nr\log n/d}{\gamma^2}\right)$	$o_P(1)$



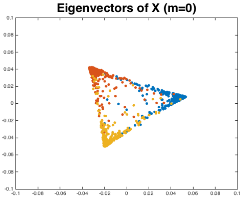


Figure: Leading eigenvectors for K and X, three true clusters are indicated in different colors.

Community Detection - an example

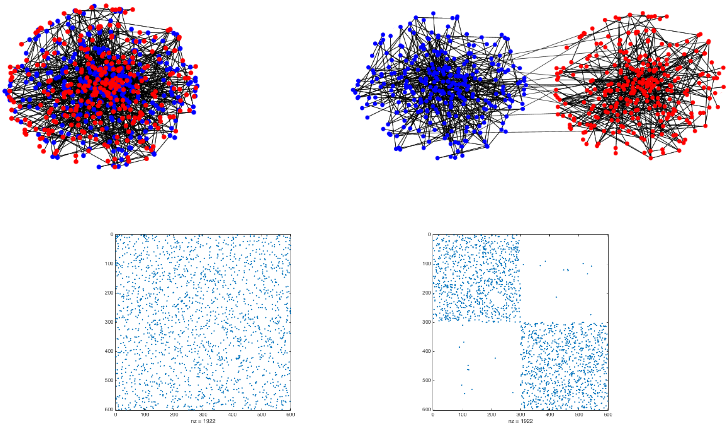


Figure: Stochastic Block Model with $B = \begin{pmatrix} 0.01 & 0.0002 \\ 0.0002 & 0.1 \end{pmatrix}$

Generative Community Model

- Stochastic Block Models [HLL83];
- Latent community matrix $Z \in \{0, 1\}^{n \times r}$;
- Each node belongs to exactly one cluster, $\sum_a Z_{ia} = 1$;
- Observe: adjacency matrix A

$$P(A_{ij} = 1 | Z_{ia} = 1, Z_{jb} = 1) = B_{ab}.$$

- Matrix representation $\mathbb{E}[A|B, Z] = ZBZ^T$.

Definition of Consistency

Definition:

Let $Z \in \{0,1\}^{n\times r}$ be the (unknown) assignment of nodes to blocks. Then any estimated assignment $\hat{Z} \in \{0,1\}^{n\times r}$ is *strongly consistent* (up to label permutations) iff

$$P[\hat{Z} = Z] \to 1 \qquad \text{As } n \to \infty.$$

$\hat{Z}$ is *weakly consistent* if

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(\hat{Z}^{(i)} \neq Z^{(i)}) = o_P(1).$$

Dense and *Sparse* Graphs

Dense: average degree $= \Omega(\log n)$;

- Spectral clustering [McS01; RCY11];
- Likelihood and modularity based methods [BC09];
- Convex relaxations [AL14; CL+15; CSX12]

Weak consistency for spectral method, strong consistency for convexified methods.

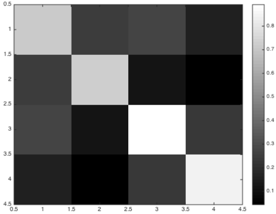
Sparse: average degree $=\Theta(1)$.

- Regularized Spectral Clustering [ACB+13; LLV15];
- Semidefinite relaxations of likelihood based methods [GV14]

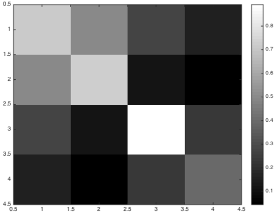
No consistency in sparse regime: a constant fraction of nodes are misclassified.

Assortativity

- Strongly Assortative: $\min_k B_{kk} > \max_k \max_{\ell \neq k} B_{k,\ell}$.
- Weakly Assortative: $\forall k, B_{kk} > \max_{\ell \neq k} B_{k,\ell}$.



Strongly Assortative



Weakly assortative

Related work

Table: Convex Relaxations for stochastic block models

Ref.	Dense	Sparse	Unequal Size	Weak assortativity	Tuning free
[HWX16]	✓				✓
[AL14]	✓			✓	✓
[CL+15]	✓		✓		
[GV14]		✓	✓		
[CSX14]	✓		✓		
This work	✓	✓	✓	✓	✓

Dealing with different cluster sizes

- Most existing work use binary clustering matrix.
- Hard to handle different cluster sizes.
- We use the following projection matrix instead.

Let m_k be the size of kth cluster,

$$X_0 = \begin{bmatrix} \frac{1}{m_1}E_{m_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{m_2}E_{m_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \frac{1}{m_r}E_{m_r} \end{bmatrix}$$

Consider the following semi-definite programming.

$$\begin{aligned} \max_X \quad & \langle A, X \rangle \\ s.t. \quad & X \succeq 0, 0 \leq X \leq 1, X\mathbf{1} = \mathbf{1}, \mathrm{trace}(X) = r \end{aligned}$$

Theoretical Guarantees

Theorem (Sparse graph)

Let $a_k = np_k, b_k = nq_k$ *are positive constants,* $\alpha := m_{\max}/m_{\min}$*.*
With probability tending to 1,

$$\frac{\|\hat{X} - X_0\|_F}{\|X_0\|_F} \leq \epsilon, \quad \text{if } \min_k (a_k - b_k) \geq \frac{C\alpha^2 r}{\epsilon^2}.$$

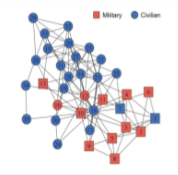
Theorem (Dense graph)

If $\min_k (p_k - q_k) > 0$*, then with probability tending to 1,*

$$\|\hat{X} - X_0\|_F = o(1) \quad \text{if} \quad \min_k (p_k - q_k)/r\alpha = \Omega(\sqrt{\max_k B_{kk}/n})$$

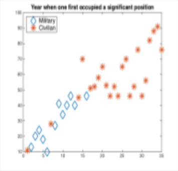
Combining Graph and Covariates

Network (Graph)



A+λYY'

Covariates (Features)

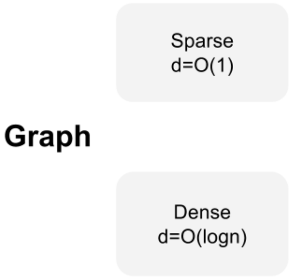


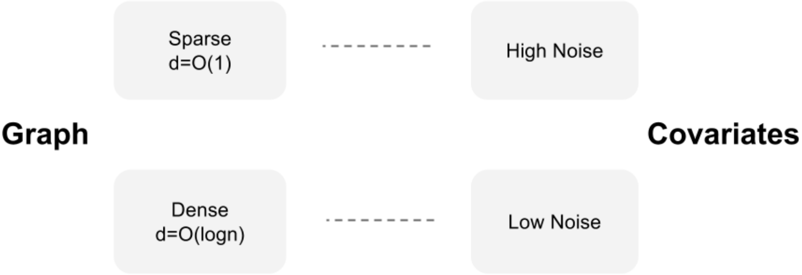
A

Y

Performance is dominated by the more informative one

Combining Graph and Covariates





$$Y_i = \sum_k \mu_k Z_{ik} + \frac{W_i}{\sqrt{d}} \quad , Cov(W_i) = \sum_k \sigma_k^2 Z_{ik} I_d$$

- Low noise: high dimension, $\sigma_k = O(1)$ [El +10];
- High noise: $\sigma_k^2 = \Theta(d)$.

Based on the Correspondence

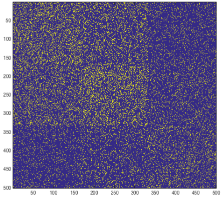
The combined SDP:

$$\begin{aligned} \max_X \quad & \langle A + \lambda K, X \rangle, \\ s.t. \quad & X \succeq 0, \\ & 0 \le X \le 1/m_{\min}, \\ & X\mathbf{1}_n = \mathbf{1}_n, \\ & \mathrm{trace}(X) = r \end{aligned} \qquad \text{(SDP-comb)}$$

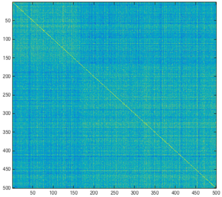
Outperforms clustering from using one source alone, especially when the information from graph and covariates are "orthogonal".

An example for "orthogonal information"

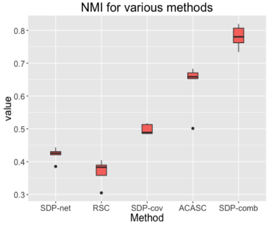
- Generate the graph with $n = 500, r = 3,\ B = \begin{bmatrix} 0.2 & 0.16 & 0.08 \\ 0.16 & 0.2 & 0.1 \\ 0.08 & 0.1 & 0.12 \end{bmatrix}$.
- Generate the covariates where $d = 100, \sigma = 1$, the centers such that $d_{12}^2 = d_{13}^2 = 0.17, d_{23}^2 = 0.02$.



(a) Network



(b) Kernel



(c) Performance

Main Results

Theorem

- *Dense graph plus low noise covariates*
  Define $\nu_k := f(2\sigma_k^2) - \max_{\ell \neq k} f(d_{k\ell}^2 + \sigma_k^2 + \sigma_\ell^2)$. *For* $\gamma' = \min_k \left(\frac{p_k - q_k}{1+\lambda} + \frac{\lambda}{1+\lambda}\nu_k\right) \geq 0$, *we have:*

$$\frac{\|\hat{X} - X_0\|_F}{\|X_0\|_F} \leq \frac{\sqrt{2\alpha^2 r}}{\gamma'} \left(\frac{1}{1+\sqrt{\lambda}} C_G \sqrt{\frac{r p_{\max}}{n}} + \frac{\sqrt{\lambda}}{1+\sqrt{\lambda}} C_K \sqrt{\frac{\log n}{\min(d,n)}}\right)$$

- *Sparse graph plus high noise kernel*
  Let $p_k = a_k/n, q_k = b_k/n, g = \bar{p}/n$. *Using* $\lambda = \ell/n$, $\pi_{\min} = n/m_{\min}$, *we have:*

$$\frac{\|\hat{X} - X_0\|_F^2}{\|X_0\|_F^2} \leq \frac{C_G + \ell C_K(f, d_{k\ell}, \sigma_{k,\ell})}{r\pi_{\min}^2 \min_k(a_k - b_k + \ell\nu_k)}$$

Improved error bound

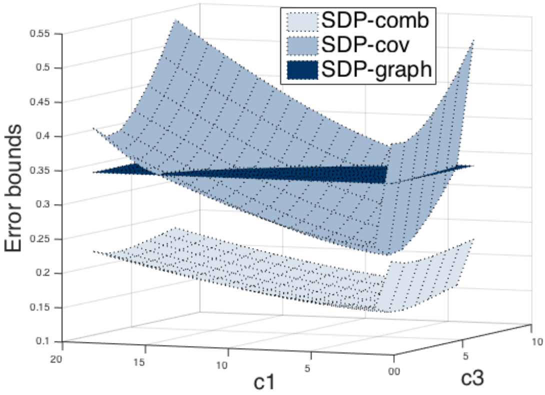


Figure: Error surfaces for sparse graph, high noise covariates and their combination.

Algorithms

- Alternating Direction Method of Multipliers [BPC+11];

$$\min_{X} \quad -\langle A, X\rangle + 1(\mathcal{L}(X) = b) + 1(Y \succeq 0) + 1(0 \leq Z \leq 1),$$
$$s.t. \quad X = Y, X = Z$$

Algorithms

Algorithm 1 ADMM

Input: Network A, node covariate matrix Y, tuning parameter ρ.

1: Compute kernel matrix K where $K(i,j) = f(\|Y_j - Y_j\|_2^2)$;
2: **while** not converge **do**
3: $\quad X^{(k+1)} = \Pi_L(\frac{1}{2}(Z^k - U^k + Y^k - V^k) + \frac{1}{\rho}(A + \lambda K))$;
4: $\quad Z^{(k+1)} = \min(\max(0, X^{k+1} + U^k), 1)$;
5: $\quad Y^{(k+1)} = \Pi_{S^+}(X^{(k+1)+V^k})$;
6: $\quad U^{(k+1)} = U^k + X^{(k+1)} - Z^{(k+1)}$;
7: $\quad V^{(k+1)} = V^k + X^{(k+1)} - Y^{(k+1)}$;
8: **end while**
9: Return X^k.

Algorithms

- Alternating Direction Method of Multipliers [BPC+11];

$$\begin{aligned} \min_{X} \quad & -\langle A, X\rangle + 1(\mathcal{L}(X) = b) + 1(Y \succeq 0) + 1(0 \le Z \le 1), \\ s.t. \quad & X = Y, X = Z \end{aligned}$$

- SDPLR [BM03] - non-convex low rank decomposition;

$$X = VV^T, \quad V \in \mathbb{R}^{n \times r}$$

Augmented Lagrangian Method

$$\begin{aligned} L(V, \alpha, \sigma) := & -\operatorname{trace}(V^T A V) + \langle \alpha, \mathcal{L}(VV^T) - b\rangle \\ & + \frac{\sigma}{2}\left((\|\mathcal{L}(VV^T) - b\|_F^2\right) \end{aligned}$$

Algorithms

Algorithm 2 Burer-Monteiro

Input: Network A, initialization $V^{(0)}$, hyper-parameters η, ϕ;

```
while not converge do
    V^(k) = arg min_V L(V, α^(k-1), σ^(k-1));
    u^k = ||𝓛(V^(k) V^(k)T) - b||_F^2;
    if u^k < η u^(k-1) then
        α^(k) = α^(k-1) + σ^(k-1)(𝓛(V^(k) V^(k)T) - b);
        u^k = u^(k-1).
    else
        σ^(k) = φ σ^(k-1);
    end if
end while
Return V^(k).
```

- **while** not converge **do**
  - $V^{(k)} = \arg\min_V L(V, \alpha^{(k-1)}, \sigma^{(k-1)})$;
  - $u^k = \|\mathcal{L}(V^{(k)} V^{(k)T}) - b\|_F^2$;
  - **if** $u^k < \eta u^{k-1}$ **then**
    - $\alpha^{(k)} = \alpha^{(k-1)} + \sigma^{(k-1)}(\mathcal{L}(V^{(k)} V^{(k)T}) - b)$;
    - $u^k = u^{k-1}$.
  - **else**
    - $\sigma^{(k)} = \phi\sigma^{(k-1)}$;
  - **end if**
- **end while**
- Return $V^{(k)}$.

Summary

- Semi-definite programming achieves stronger guarantees than spectral methods;
- By using projection matrix instead of binary matrix we can achieve provable recovery for a broader family of problems;
- Combining graph with node covariates improves the accuracy.
- It has some computational challenges when the scale of the problem increases.

Questions?

Reference I

Preprint: https://arxiv.org/abs/1607.02675

- A. A. Amini, A. Chen, P. J. Bickel, E. Levina, *et al.*, "Pseudo-likelihood methods for community detection in large sparse networks", *The Annals of Statistics*, vol. 41, no. 4, pp. 2097–2122, 2013.
- A. A. Amini and E. Levina, "On semidefinite relaxations for the block model", *ArXiv preprint arXiv:1406.5647*, 2014.
- P. J. Bickel and A. Chen, "A nonparametric view of network models and newman–girvan and other modularities", *Proceedings of the National Academy of Sciences*, vol. 106, no. 50, pp. 21 068–21 073, 2009.
- S. Burer and R. D. Monteiro, "A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization", *Mathematical Programming*, vol. 95, no. 2, pp. 329–357, 2003.
- S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers", *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

Reference II

- T. T. Cai, X. Li, *et al.*, "Robust and computationally feasible community detection in the presence of arbitrary outlier nodes", *The Annals of Statistics*, vol. 43, no. 3, pp. 1027–1059, 2015.
- Y. Chen, S. Sanghavi, and H. Xu, "Clustering sparse graphs", in *Advances in neural information processing systems*, 2012, pp. 2204–2212.
- ——, "Improved graph clustering", *IEEE Transactions on Information Theory*, vol. 60, no. 10, pp. 6440–6455, 2014.
- N. El Karoui *et al.*, "On information plus noise kernel random matrices", *The Annals of Statistics*, vol. 38, no. 5, pp. 3191–3216, 2010.
- O. Guédon and R. Vershynin, "Community detection in sparse networks via grothendieck's inequality", *ArXiv preprint arXiv:1411.4686*, 2014.
- M. X. Goemans and D. P. Williamson, ". 879-approximation algorithms for max cut and max 2sat", in *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, ACM, 1994, pp. 422–431.
- P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps", *Social networks*, vol. 5, no. 2, pp. 109–137, 1983.
- B. Hajek, Y. Wu, and J. Xu, "Achieving exact cluster recovery threshold via semidefinite programming", *IEEE Transactions on Information Theory*, vol. 62, no. 5, pp. 2788–2797, 2016.

Reference III

- C. M. Le, E. Levina, and R. Vershynin, "Sparse random graphs: Regularization and concentration of the laplacian", *ArXiv preprint arXiv:1502.03049*, 2015.
- J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations", in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Oakland, CA, USA., vol. 1, 1967, pp. 281–297.
- F. McSherry, "Spectral partitioning of random graphs", in *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, IEEE, 2001, pp. 529–537.
- A. Y. Ng, M. I. Jordan, Y. Weiss, *et al.*, "On spectral clustering: Analysis and an algorithm", *Advances in neural information processing systems*, vol. 2, pp. 849–856, 2002.
- M. L. Overton and R. S. Womersley, "Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices", *Mathematical Programming*, vol. 62, no. 1-3, pp. 321–357, 1993.
- K. Rohe, S. Chatterjee, and B. Yu, "Spectral clustering and the high-dimensional stochastic blockmodel", *The Annals of Statistics*, pp. 1878–1915, 2011.

Reference IV

- B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem", *Neural computation*, vol. 10, no. 5, pp. 1299–1319, 1998.
- B. Yan and P. Sarkar, "On robustness of kernel clustering", in *Advances in Neural Information Processing Systems*, 2016, pp. 3090–3098.
- Y. Yu, T. Wang, and R. Samworth, "A useful variant of the davis–kahan theorem for statisticians", *Biometrika*, vol. 102, no. 2, pp. 315–323, 2015.

back

Davis-Kahan Theorem

Theorem ([YWS15])

Let $\Sigma, \hat{\Sigma} \in \mathbb{R}^{p\times p}$ be symmetric, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ respectively. Fix $1 \leq r \leq s \leq p$ and assume that $\min(\lambda_{r-1} - \lambda_r, \lambda_{s-1} - \lambda_s) > 0$, where $\lambda_0 := \infty$ and $\lambda_{p+1} := -\infty$. Let $d := s - r + 1$, and let $V = (v_r, v_{r+1}, \cdots, v_s) \in \mathbb{R}^{p\times d}$ and $\hat{V} = (\hat{v}_r, \hat{v}_{r+1}, \cdots, \hat{v}_s) \in \mathbb{R}^{p\times d}$ have orthonormal columns satisfying $\Sigma v_j = \lambda_j v_j$ and $\hat{\Sigma}\hat{v}_j = \hat{\lambda}_j \hat{v}_j$, for $j = r, r+1, \cdots, s$. Then there exists an orthogonal matrix $\hat{O} \in \mathbb{R}^{d\times d}$ such that

$$\|\hat{V}\hat{O} - V\|_F \leq \frac{2^{3/2}\|\hat{\Sigma} - \Sigma\|_F}{\min(\lambda_{r-1} - \lambda_r, \lambda_{s-1} - \lambda_s)}.$$

back

Example of high dimensional covariate matrix

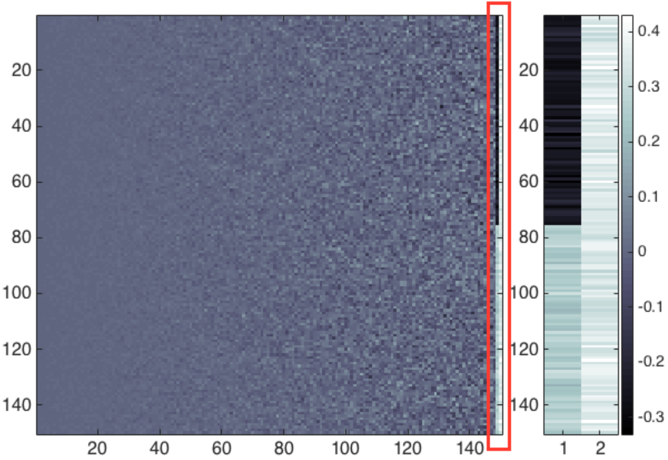


Figure: $YY^T = K$

Nonparametric Asymptotic Model

- Given $\xi_1, \ldots, \xi_n$ i.i.d. $\mathcal{U}(0,1)$ associated with vertices $1, \ldots, n$, let:

$$h : [0,1]^2 \to [0,1] \qquad , h \text{ symmetric.}$$

$$P[A_{ij} = 1 | \xi_1, \ldots, \xi_n] = P[A_{ij} = 1 | \xi_i, \xi_j] = h(\xi_i, \xi_j)$$

- Determines P_h on $n \times n$ symmetric matrices with 0/1 elements, for all n. (Aldous, Hoover (1983))
- Analogous to de Finetti's Theorem.

Tuning λ

Pick the λ that maximizes the eigengap of $\hat{X}$.

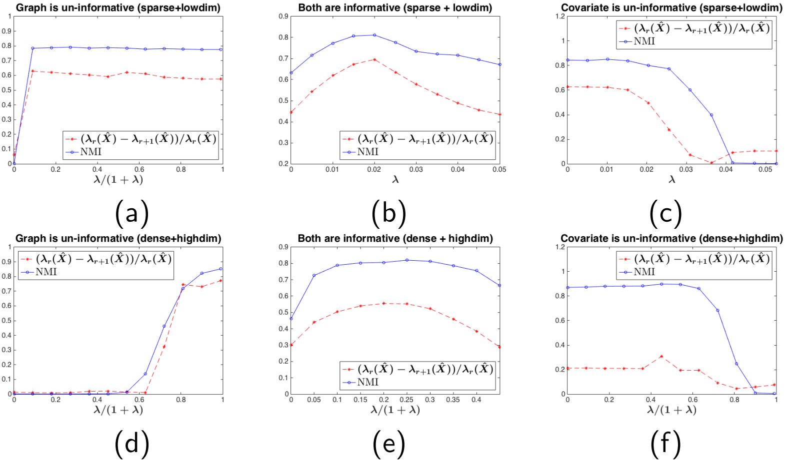


Figure: NMI and eigengap as λ changes