USC University of Southern California

On the Power of SVD in the Stochastic Block Model

Xinyu Mao and Jiapeng Zhang, University of Southern California

NeurIPS 2023

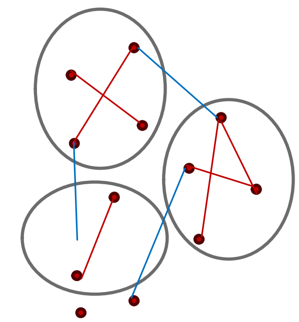
Heuristic for clustering: dimensionality reduction

- "Curse of dimensionality": higher dimensional data → worse performance.
  - In particular, this phenomenon is observed for clustering algorithms.
- Heuristic: Apply dimensionality reduction before clustering.
  - A widely-used dimensionality reduction tool: Spectral methods like Principal Component Analysis/Singular Value Decomposition.

Why do spectral methods (like PCA/SVD) help to cluster high-dimensional datasets?

Stochastic Block Model (SBM) and Vanilla-SVD algorithm

Well-known theoretical Benchmark for graph clustering



(Symmetric) SBM

1. Let V denote the set of vertices.
2. V is partitioned into k disjoint sets $V = \bigcup_{i=1}^{k} V_k$ uniformly at random.
3. A random (undirected) graph $\hat{G}$ is sampled in the following way: $\forall u, v \in V$,
   - an edge $\{u, v\}$ is added independently with probability p, if u, v are in the same set;
   - otherwise, an edge $\{u, v\}$ is added independently with probability q.

Task: Given $\hat{G}$, recover the partition $V_1, \ldots, V_k$.

$\hat{G}_u$: the column indexed by u in the adjacent matrix $\hat{G}$.

Vanilla-SVD algorithm

1. Let P_k be the projection to the subspace spanned by the first k eigenvectors of $\hat{G}$.
2. Compute $\rho(u) \coloneqq P_k \hat{G}_u$ for each vertex u.
3. Clustering according to the distances given by the vector representation ρ: put u, v in the same cluster if $\|\rho(u) - \rho(v)\| \leq 0.2(p-q)\sqrt{n / k}$.

Many existing spectral algorithms: [McSherry 01, Vu18]…

"Vanilla spectral algorithm": No extra steps, widely used in practice.

Our result:
Vanilla-SVD exhibits clustering power

Vanilla-SVD algorithm

1. Let P_k be the projection to the subspace spanned by the first k eigenvectors of $\widehat{G}$.
2. Compute $\rho(u) := P_k \widehat{G}_u$ for each vertex u.
3. Clustering according to the distances given by the vector representation ρ.

Main Theorem

In the symmetric SBM, Vanilla-SVD algorithm recovers all clusters with probability $1 - O(n^{-1})$ if

$$\max\{p(1-p), q(1-q)\} \geq \frac{C \log n}{n} \quad \text{and} \quad n \geq C \cdot k \left(\frac{\sqrt{kp} \log^6 n + \sqrt{\log n}}{p - q} \right)^2,$$

where n := # of vertices and C is a universal constant.

- Previous analysis only applies to either non-vanilla algorithms or more restricted parameter regimes.
  - E.g. for vanilla-SVD, only the case $k = O(1)$ is analyzed prior to our work.
- Provides **theoretical understanding of successful heuristics**.
- Technical Contribution: a new method to analyze the eigenspace under random perturbation.

Summary

Why do spectral methods (like PCA/SVD) help to cluster high-dimensional datasets?

Main Theorem

In the symmetric SBM, Vanilla-SVD algorithm recovers all clusters with probability $1 - O(n^{-1})$ if

$$\max\{p(1-p), q(1-q)\} \geq \frac{C \log n}{n} \quad \text{and} \quad n \geq C \cdot k \left(\frac{\sqrt{kp} \log^6 n + \sqrt{\log n}}{p-q} \right)^2,$$

where n := # of vertices and C is a universal constant.

Our result suggests that
vanilla spectral algorithms exhibit clustering power itself.

Future works: better parameters, apply our method to other models…

Thank you for listening ☺