Overlapping Clustering Models, and One (class) SVM to Bind Them All

Xueyu Mao

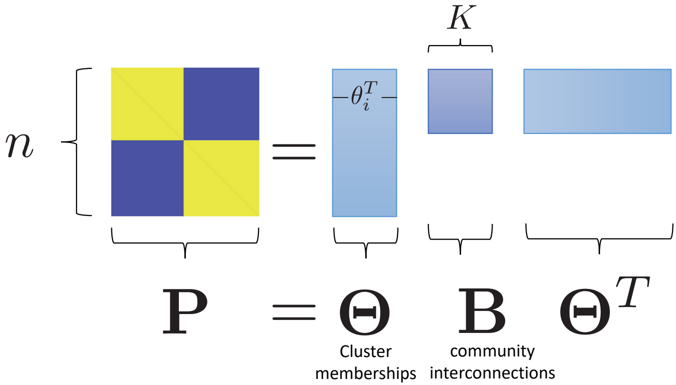
Department of Computer Science
The University of Texas at Austin

Neural Information Processing Systems
December 6, 2018

Joint work with **Purnamrita Sarkar** and **Deepayan Chakrabarti**

(Poster: Today 10:45 AM – 12:45 PM @ Room 517 AB #114)

Stochastic Blockmodel



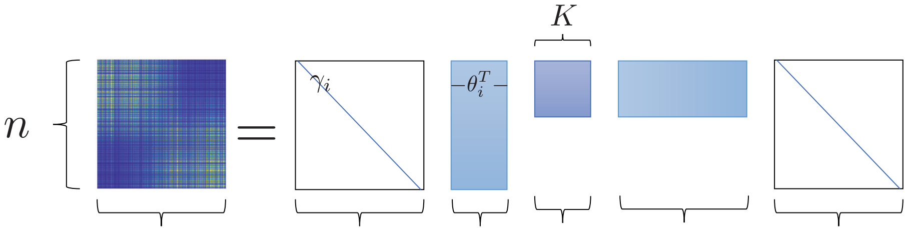
Limitations:

- Each node belongs to exactly one community
- All nodes in the same community have the same expected degree

Extensions of Stochastic Blockmodel

- Mixed membership blockmodels (Airoldi et al. 2008) extend this to **allow overlap**
  - $\boldsymbol{\theta}_i$ is a distribution over K communities
- Degree-corrected blockmodels (Karrer and Newman 2011) extend this to **allow heterogeneous degree distributions**
  - Each node has a degree parameter γ_i
- There are many other extensions to model the above two properties
  - DCMMSB (Jin et al., 2017)
  - OCCAM (Zhang et al. 2014)
  - SBMO (Kaufmann et al. 2016)

Overlapping clustering model



$$\mathbf{P} = \mathbf{\Gamma} \; \mathbf{\Theta} \; \mathbf{B} \; \mathbf{\Theta}^T \; \mathbf{\Gamma}$$

Degree parameters ($\mathbf{\Gamma}$), Cluster memberships ($\mathbf{\Theta}$), community interconnections ($\mathbf{B}$)

- This covers many well-known overlapping clustering models:

$\|\boldsymbol{\theta}_i\|_1 = 1$	DCMMSB
$\|\boldsymbol{\theta}_i\|_2 = 1$	OCCAM
$\boldsymbol{\theta}_i \in \{0,1\}^K$	SBMO

- The LDA topic model (Blei et al. 2003) is also a special case

Main idea

	Model	Main idea
(Zhang et al. 2014)	OCCAM	k-median on regularized eigenvectors
(Kaufmann et al. 2016)	SBMO	Alternating minimization
(Mao et al., 2017)	MMSB	Finding K corners of a simplex in $\mathbb{R}^K$
(Jin et al., 2017)	DCMMSB	Finding K corners of a simplex in $\mathbb{R}^{K-1}$
(Arora et al., 2013)	Topic Models	Finding K corners of a simplex in $\mathbb{R}^V$
This work	All	Finding extreme rays of a **convex cone**

- Let $\mathbf{V} \in \mathbb{R}^{n \times K}$ be the top-K eigenvectors of $\mathbf{P}$
- Rows of $\mathbf{V}$ form a **cone**

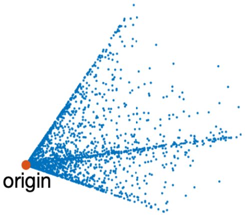
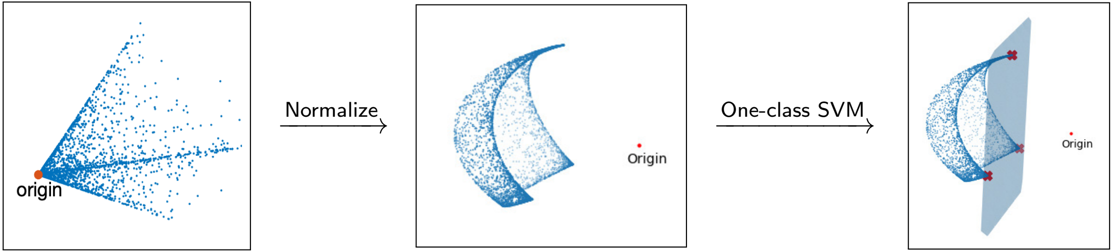


Figure: Each point is a row of $\mathbf{V}$

Main idea



- SVM-cone:
  - Normalize rows $\mathbf{v}_i$ of $\mathbf{V}$ to unit ℓ_2 norm
    - Each node lies on the intersection of the cone and the unit sphere
  - Run a one-class SVM $\Longrightarrow$ **support vectors are the corners**
  - Estimate community memberships by regression $\mathbf{v}_i$ on these corners
- This is for the ideal "population" version
  - Similar ideas provably work for the "empirical" version

Per-node Consistency Guarantees

- This one algorithm yields consistency guarantees for
  - community memberships of **each node**
    - most algorithms show guarantees for the whole matrix
  - for **all overlapping clustering models** mentioned earlier
- Example

Per-node consistency guarantee for DCMMSB (informal)

If $\boldsymbol{\theta}_i \sim \mathrm{Dirichlet}(\boldsymbol{\alpha})$, under a broad parameter regime, with high probability,

$$\max_i \|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\| = \tilde{O}\left(\frac{g}{\sqrt{\rho n}}\right),$$

where g depends on model parameters.

Conclusions

- A simple and scalable algorithm

$$\boxed{\text{Eigendecomposition} \Rightarrow \text{Row-normalize} \Rightarrow \text{One-class SVM} \Rightarrow \text{Regression}}$$

  - infers community memberships for a **broad class** of overlapping clustering models
  - with **per-node** consistency guarantees

- Good performance on several large scale real-world datasets.

$$\boxed{\text{Poster: Today 10:45 AM – 12:45 PM @ Room 517 AB \#114}}$$