

Manifold Clustering in the Setting of Generalized Random Dot Product Graphs

SDSS Lightning Presentation

May 2023



John Koo,
Postdoctoral Fellow,
Indiana University

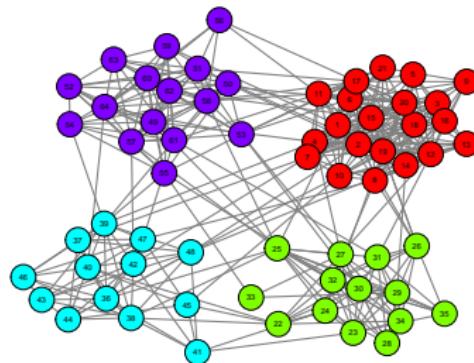


Minh Tang,
Assistant Professor,
NC State University



Michael W. Trosset,
Professor of Statistics,
Indiana University

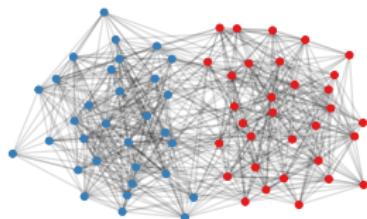
Community Detection for Networks



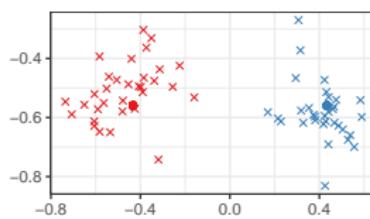
How might we cluster the nodes of a network?

Connecting Block Models to the GRDPG

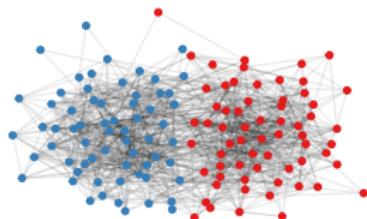
SBM



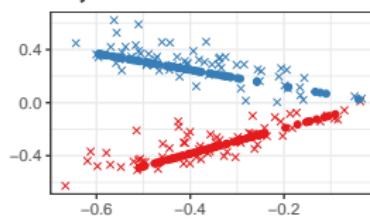
Point Masses



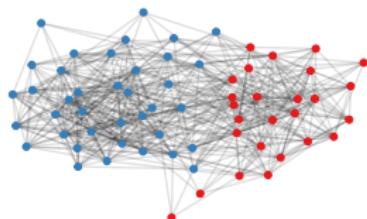
DCBM



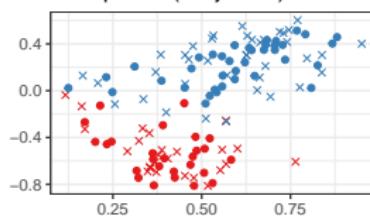
Rays



PABM

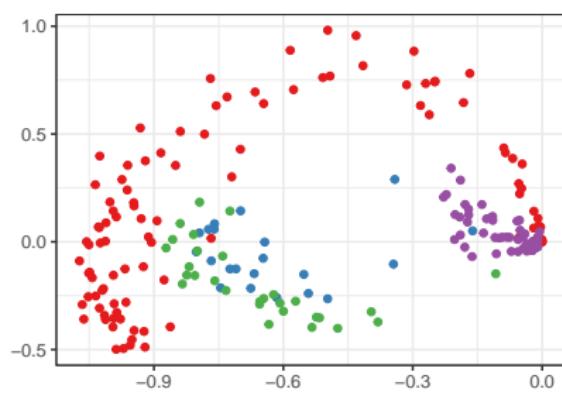
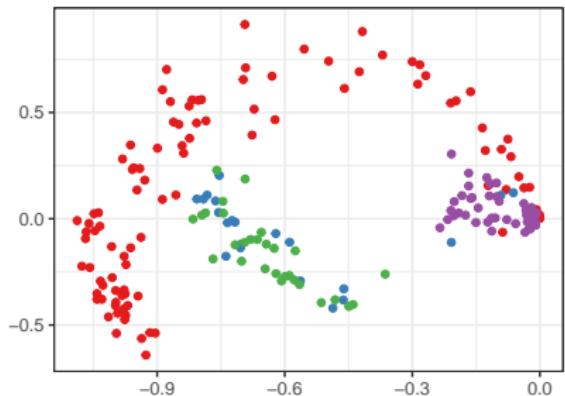
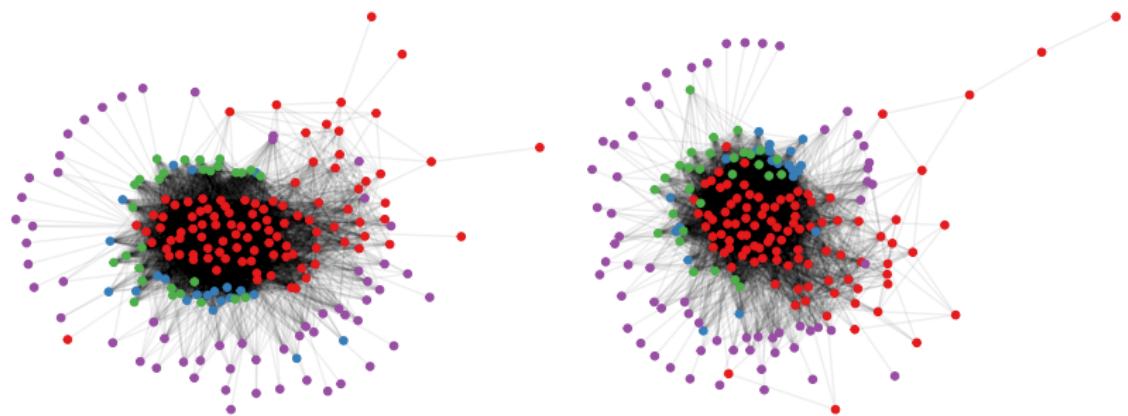


Subspaces (Projected)



- K-means clustering
- Gaussian mixture models
- K-means with cosine similarity
- GMM on angles
- Orthogonal Spectral Clustering
- Sparse Subspace Clustering

Nonlinear Community Structure



Manifold Block Model

Let $p, q \geq 0$, $d = p + q \geq 1$, $1 \leq r < d$, $K \geq 2$, and $n > K$ be integers. Define manifolds $\mathcal{M}_1, \dots, \mathcal{M}_K \in \mathcal{X}$ for

$\mathcal{X} = \{x, y \in \mathbb{R}^d : x^\top I_{p,q} y \in [0, 1]\}$ each by continuous function $g_k : [0, 1]^r \rightarrow \mathcal{X}$. Define probability distribution F with support $[0, 1]^r$. Then the following mixture model is a *manifold block model*:

1. Draw labels $z_1, \dots, z_n \stackrel{\text{iid}}{\sim} \text{Categorical}(\alpha_1, \dots, \alpha_K)$.
2. Draw latent vectors by first taking $t_1, \dots, t_n \stackrel{\text{iid}}{\sim} F$ and then computing each $x_i = g_{z_i}(t_i)$.
3. Compile the latent vectors into data matrix $X = [x_1 \mid \dots \mid x_n]^\top$ and define the adjacency matrix as $A \sim \text{GRDPG}_{p,q}(X)$.

Manifold Block Model

1. $z_1, \dots, z_n \stackrel{\text{iid}}{\sim} \text{Categorical}(1/2, 1/2)$
2. $t_1, \dots, t_n \stackrel{\text{iid}}{\sim} \text{Uniform}(0, 1)$
3. $x_i = g_{z_i}(t_i)$
 - $g_1(t) = [t^2, 2t(1-t)]^\top$
 - $g_2(t) = [2t(1-t), (1-t)^2]^\top$
4. $A \sim \text{GRDPG}_{2,0}(X)$

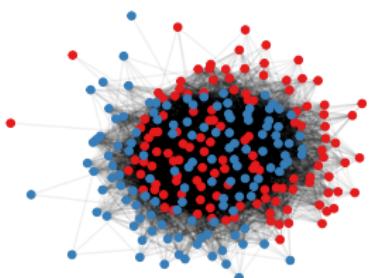
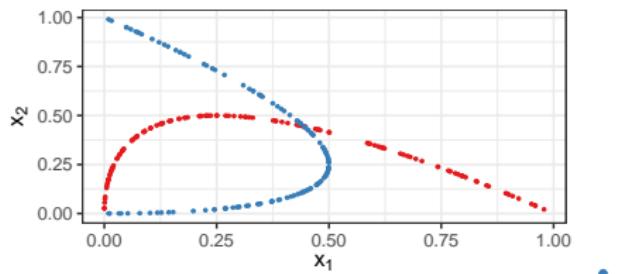


Figure 1: Latent vectors on intersecting curves (left), along with an RDGP drawn from this configuration (center) and its ASE (right).

K-Curves Clustering

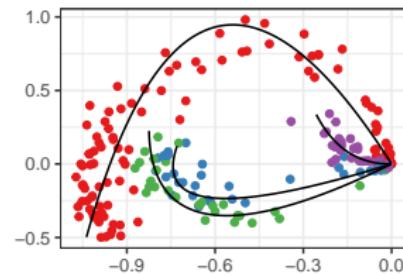
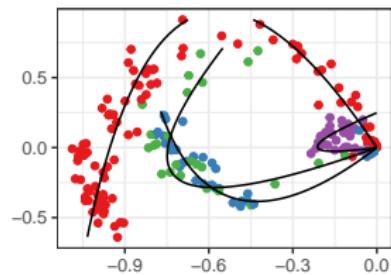
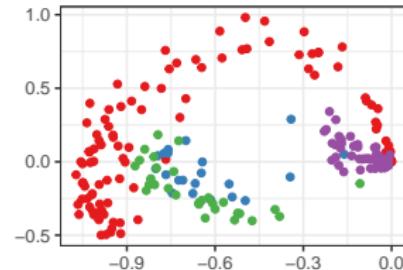
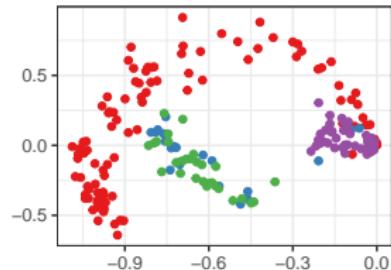
Algorithm 1: *K*-curves clustering.

Data: Adjacency matrix A , number of communities K , embedding dimensions p , q , stopping criterion ϵ

Result: Community assignments $1, \dots, K$, curves g_1, \dots, g_K

```
1 Compute  $X$ , the ASE of  $A$  using the  $p$  most positive and  $q$  most negative
   eigenvalues and their corresponding eigenvectors.
2 Initialize community labels  $z_1, \dots, z_n$ .
3 repeat
4   for  $k = 1, \dots, K$  do
5     Define  $X_k$  as the rows of  $X$  for which  $z_i = k$ .
6     Fit curve  $g_k$  and positions  $t_{k_i}$  to  $X_k$  by minimizing
          $\sum_{k_i} \|x_{k_i} - g_k(t_{k_i})\|^2$ .
7   end
8   for  $k = 1, \dots, K$  do
9     Assign  $z_i \leftarrow \arg \min_\ell \|x_i - g_\ell(t_i)\|^2$ .
10  end
11 until the change in  $\sum_k \sum_{i \in C_k} \|x_i - g_k(t_i)\|^2$  is less than  $\epsilon$ 
```

Example



Asymptotic Results

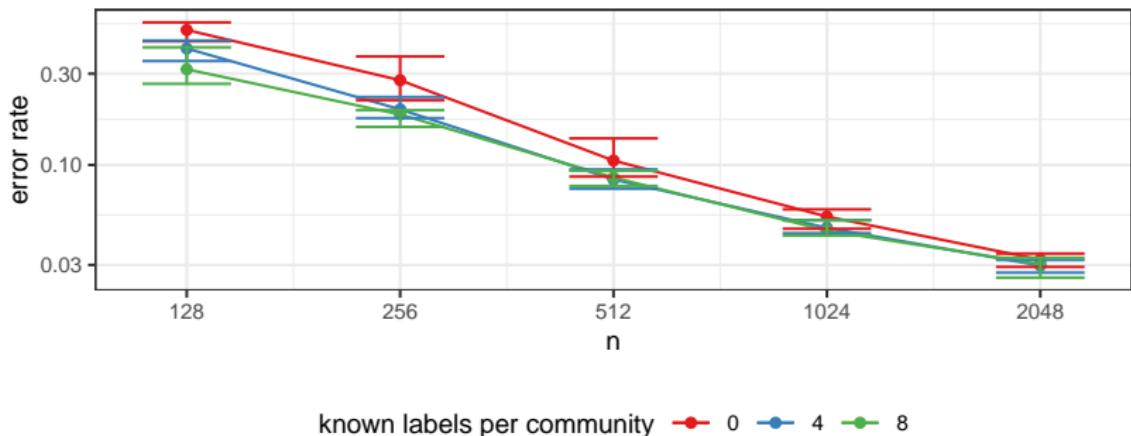
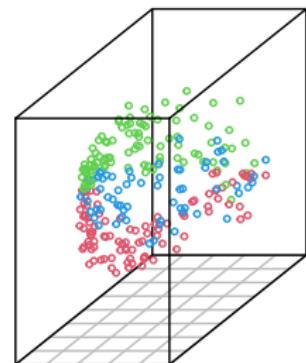
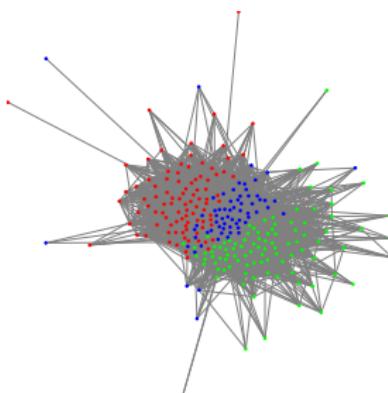
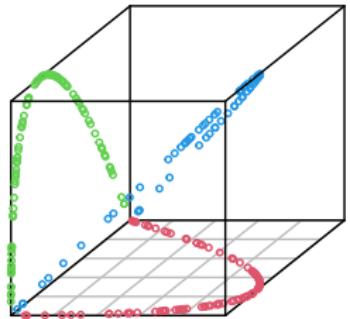
Theorem. Let an MBM be such that the manifolds are described by functions $g_1(t), \dots, g_K(t)$ which are polynomial curves of order R . Define the loss of K -curves clustering as:

$$L(\hat{z}_1, \dots, \hat{z}_n, \hat{g}_1, \dots, \hat{g}_K; A) = \sum_k \sum_{i: \hat{z}_i=k} \|\hat{x}_i - \hat{g}_k(t_i)\|^2,$$

where \hat{x}_i are the embedding vectors of A . Suppose that for each community k , we have labels for at least $R + 1$ vertices. Then as $n \rightarrow \infty$, K -curves clustering outputs estimators such that

$$L(\hat{z}_1, \dots, \hat{z}_n, \hat{g}_1, \dots, \hat{g}_K; A) \xrightarrow{p} 0.$$

Simulation



Thank you!

Code and drafts available at

<https://github.com/johneverettkoo/manifold-block-models>