Manifold Clustering in the Setting of Generalized Random Dot Product Graphs

John Koo[1], Minh Tang[2], Michael W. Trosset[1]

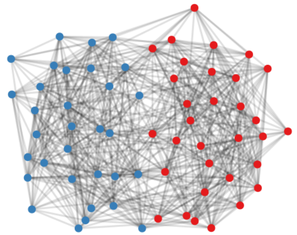
[1] Indiana University

[2] North Carolina State University

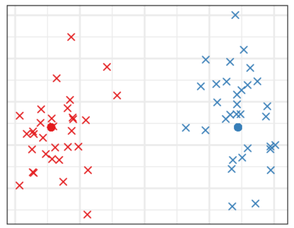
Block Models as Linear GRDPGs

SBM, DCBM, and PABM are generalized random dot product graphs in which the communities correspond to linear structures in the latent space.

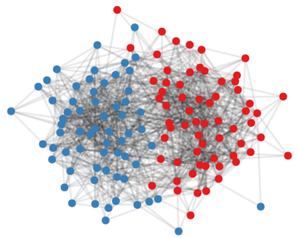
SBM



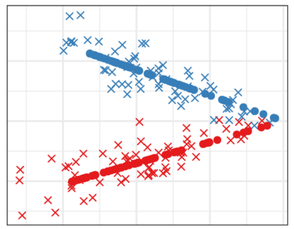
Point Masses



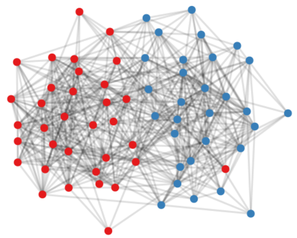
DCBM



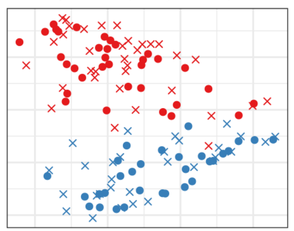
Rays



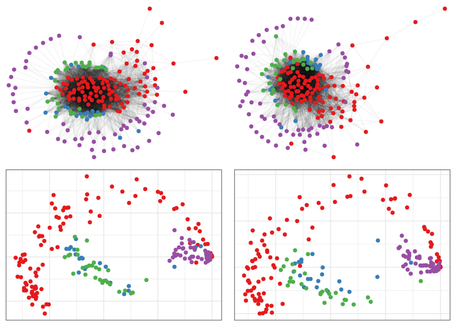
PABM



Subspaces (Projected)



GRDPGs with Nonlinear Latent Structure



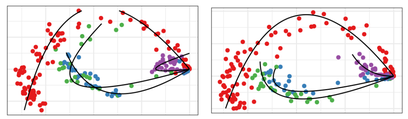
Manifold Block Model

Let $p, q \geq 0$, $d = p + q \geq 1$, $1 \leq r < d$, $K \geq 2$, and $n > K$ be integers. Let $\mathcal{X} = \{x, y \in \mathbb{R}^d : x^\top I_{p,q} y \in [0,1]\}$ and define manifolds $\mathcal{M}_1, \ldots, \mathcal{M}_K \subset \mathcal{X}$ each by continuous function $g_k : [0,1]^r \to \mathcal{X}$. Define probability distribution F with support $[0,1]^r$. Then the following mixture model is a *manifold block model*:

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

K-Curves Clustering

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 $z_1, \ldots, z_n$.
3. While change in the loss function $L = \sum_k \sum_{i \in C_k} \|x_i - g_k(t_i)\|^2$ is less than ϵ:
   i. For $k = 1, \ldots, K$:
      a. Define X_k as the rows of X for which $z_i = k$.
      b. Fit curve g_k and positions t_i by minimizing $\sum_{i \in C_k} \|x_i - g_k(t_i)\|^2$.
   ii. For $i = 1, \ldots, n$:
      a. Assign $z_i \leftarrow \arg\min_k \|x_i - g_k(t_i)\|^2$.

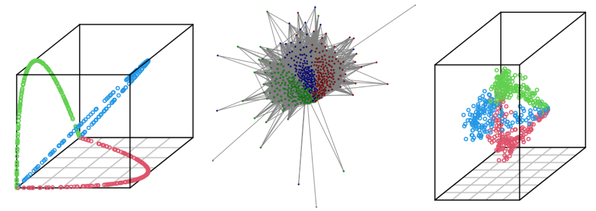


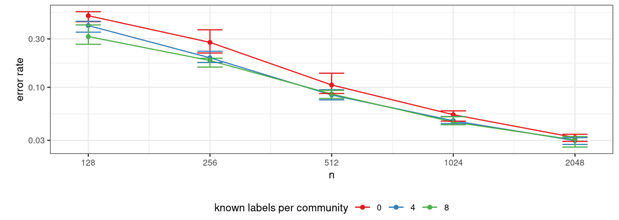
Theorem. Let an MBM be such that the manifolds are described as polynomial curves of order R. Suppose that for each community k, we have labels for at least $R+1$ vertices. Then K-curves clustering outputs estimators such that

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

Simulation

Latent vectors were drawn uniformly on three intersecting quadratic curves in $\mathbb{R}^3$ (left) to construct a GRDPG (middle). Curves were then fitted to the ASE (right) and embedding vectors were assigned labels based on proximity to the curves.





Conclusion

Block models can be expressed as GRDPGs in which the communities are linear structures in the latent space. We propose the manifold block model to extend this to nonlinear latent structures and the K-curves clustering algorithm to estimate these structures for community detection.