Manifold Clustering for Latent Structure Block Models

John Koo
Department of YYY, University of XXX

April 13, 2022

Abstract

The text of your abstract. 200 or fewer words.

Keywords: block models, community detection, coordinate descent, latent structure models, manifold clustering, random dot product graph

1 Introduction

We define a *Bernoulli graph* as a random graph model for which edge probabilities are contained in an edge probability matrix $P \in [0,1]^{n \times n}$, and an edge occurs between vertices i and j with probability P_{ij}. Common random graph models then impose structure on P, based on various assumptions about the way in which the data are generated, or to allow P to be estimated. One example is the Erdös-Rényi model, in which all edge probabilities are fixed, i.e., $P_{ij} = p$.

One common analysis task for graph and network data is community detection, which assumes that each vertex of a graph has a hidden community label. The goal of the analysis is then to retrieve these labels. In order to perform this analysis as a statistical inference task is to define a probability model with inherent community structure. We call such models *block models*: First, each vertex is assigned a label $z_1, ..., z_n \in \{1, 2, ..., K\}$ where $K \ll n$. Then each edge probability P_{ij} is said to depend on the labels z_i and z_j, possibly along with some other parameters. For example, the stochastic block model (SBM) sets a fixed edge probability for each pair of communities, i.e., $P_{ij} = \omega_{z_i, z_j}$. The degree-corrected block model (DCBM) assigns an additional parameter θ_i to each vertex by which edge probabilities are scaled, i.e., $P_{ij} = \theta_i \theta_j \omega_{z_i, z_j}$. The popularity adjusted block model (PABM) assigns K parameters to each vertex $\lambda_{i1}, \lambda_{i2}, ..., \lambda_{iK}$ that describe that vertex's affinity toward each community; the edge probability between vertices i and j is then defined as the product of vertex i's affinity toward vertex j's community and vertex j's affinity toward vertex i's community, i.e., $P_{ij} = \lambda_{iz_j} \lambda_{jz_i}$.

The three block model types, as well as the Erdös-Rényi model, impose structure on P, including on the rank of P. P has rank 1 for the Erdös-Rényi model, rank K for the SBM and DCBM, and rank K^2 for the PABM. This provides the intuition behind another family of Bernoulli graphs called the *random dot product graph* (RDPG) and *generalized random dot product graph* (GRDPG). In the RDPG, each vertex has a corresponding latent vector in d-dimensional Euclidean space, where d is the rank of P and P is positive semidefinite. Then the edge probability between each pair of vertices is defined as the inner product between the corresponding latent vectors, i.e., $P_{ij} = x_i^\top x_j$. If the latent vectors are collected in a data matrix $X = \begin{bmatrix} x_1 \mid \cdots \mid x_n \end{bmatrix}^\top$, then the edge probability matrix for the RDPG is

$P = XX^\top$. Similarly, the edge probability between each pair of vertices for the GRDPG is defined as the indefinite inner product between the corresponding latent vectors, i.e., $P_{ij} = x_i^\top I_{p,q} x_j$, where $I_{p,q} = \text{blockdiag}(I_p, -I_q)$ and $p + q = d$. Then the edge probability matrix for the GRDPG is $P = XI_{p,q}X^\top$. This allows for a model similar to the RDPG for non-positive semidefinite P. While the RDPG and GRDPG do not necessarily have community structure, it has been shown that block models are specific cases of the RDPG or GRDPG in which latent vectors are organized by community. This includes the SBM, in which communities correspond to point masses, DCBM, in which communities correspond to line segments, and PABM, in which communities correspond to orthogonal subspaces. In this work, we extend this idea to communities organized into more general latent structures. In particular, we assume that each community corresponds to a manifold in the latent space.

2 Latent Structure Block Models

All block models are latent structure models.

Insert something about ASE and its consistency here.

3 Methods

Here, we provide two algorithms for MBM community detection.

First, we consider the case in which communities correspond to manifolds in the latent space that do not intersect and are separated by some finite distance. In this scenario, we use the convergence of the ASE to show that single linkage clustering on the latent space produces a clustering such that the total number of misclustered vertices goes to zero, with high probability.

Next, we consider the case in which communities correspond to one-dimensional manifolds in the latent space and may or may not intersect. In this scenario, we propose an alternating coordinate descent algorithm that alternates between estimating the structure of the manifolds and the community labels, which we call K-curves clustering. We again use the convergence of the ASE to show that under certain conditions, K-curves clustering produces a clustering such that the proportion of misclustered vertices goes to zero, with

high probability.

3.1 Nonintersecting Manifolds

In this section, we consider the following scenario: Suppose that each community is represented by a manifold $\mathcal{M}_k$, $k \in \{1, ..., K\}$ in the latent space of a RDPG or GDRPG, and the manifolds do not intersect each other. Define $\delta = \min_{k,\ell} \min_{x \in \mathcal{M}_k, y \in \mathcal{M}_\ell} \|x - y\|$, the minimum distance between two manifolds. We assume that $\delta > 0$.

First, we consider the case in which each $\mathcal{M}_k \subset \mathbb{R}^d$ is a one-dimensional manifold, which can be characterized by a function $g_k : [0, 1] \to \mathcal{M}_k$. Let F be a probability distribution with support $[0, 1]$. Then we define a mixture model as follows:

1. Draw $t_1, ..., t_n \overset{\text{iid}}{\sim} F$.
2. Draw $z_1, ..., z_n \overset{\text{iid}}{\sim} \text{Multinomial}(\alpha_1, ..., \alpha_K)$, the community labels.
3. Let each $x_i = g_{z_i}(t_i)$ be the latent vector for vertex v_i.
   Assume $\min_{k,\ell} \min_{t,s} \|g_k(t) - g_\ell(s)\| = \delta > 0$.
4. Draw $A \sim \text{RDPG}(X)$ or $A \sim \text{GRDPG}_{p,q}(X)$.

3.1.1 Preliminary Theory

Distributions of differences of order statistics Let $D_i = X_{(i+1)} - X_{(i)}$. Then if $\max_i D_i < \delta$, we have sufficient separation of points in $\mathcal{M}_1$. Then it is sufficient to quantify $P(\max_i D_i > \delta)$ as a function of n and δ and show that this converges to zero as n grows to ∞.

We denote $f(x)$ as the density of each X_i, $g_i(x)$ as the density of $X_{(i)}$, $g_{ij}(x, y)$ as the joint density of $X_{(i)}, X_{(j)}$, and $h_i(d)$ as the density of D_i (with corresponding capital letters for the cumulative distribution functions).

The following are taken as given[1]:

1. $g_i(x) = \frac{n!}{(n-i)!(i-1)!}(F(x))^{i-1}(1 - F(x))^{n-i} f(x)$.
2. $g_{ij}(x, y) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!}(F(x))^{i-1}(F(y) - F(x))^{j-i-1}(1 - F(y))^{n-j} f(x) f(y)$.
3. By convolution, $h_i(d) = \int_0^1 g_{i,i+1}(x, x + d)\mathrm{d}x$.

[1] `https://en.wikipedia.org/wiki/Order_statistic`

Lemma 1 (The probability density function of D_i)**.**

$$h_i(d) = \int_0^{1-d} \frac{n!}{(i-1)!(n-i-1)!}(F(x))^{i-1}(1-F(x+d))^{n-i-1}f(x)f(x+d)\mathrm{d}x \quad (1)$$

Proof. This is just a direct consequence of 2 and 3 under the given statements. We also note that because the support of X_i is $[0,1]$, the integral only needs to be evaluated from 0 to $1-d$ because of the $f(x+d)$ and $1-F(x+d)$ terms. □

Lemma 2 (The cumulative distribution function of D_i)**.**

$$P(D_i < \delta) = H_i(\delta) = 1 - \int_0^{1-\delta} \frac{n!}{(n-i)!(i-1)!}(F(x))^{i-1}(1-F(x+\delta))^{n-i}f(x)\mathrm{d}x \quad (2)$$

Proof.

$$\begin{aligned}
H_i(\delta) &= \int_x^{x+\delta} h_i(d)\mathrm{d}d \\
&= \int_x^{x+\delta}\int_0^1 \frac{n!}{(i-1)!(n-i-1)!}((F(x))^{i-1}(1-F(x+d))^{n-i-1}f(x)f(x+d)\mathrm{d}x\mathrm{d}d \\
&= \int_0^1 \frac{n!}{(i-1)!(n-i-1)!}(F(x))^{i-1}f(x)\int_x^{x+\delta}(1-F(x+d))^{n-i-1}f(x+d)\mathrm{d}d\mathrm{d}x \\
&= \int_0^1 \frac{n!}{(i-1)!(n-i-1)!}(F(x))^{i-1}f(x)\int_{F(x)}^{F(x+\delta)}(1-u)^{n-i-1}\mathrm{d}u\mathrm{d}x \\
&= \int_0^1 \frac{n!}{(i-1)!(n-i)!}(F(x))^{i-1}f(x)((1-F(x))^{n-i}-(1-F(x+\delta))^{n-i})\mathrm{d}x \\
&= \int_0^1 g_i(x)\mathrm{d}x - \int_0^1 \frac{n!}{(i-1)!(n-i)!}(F(x))^{i-1}(1-F(x+\delta))^{n-i}f(x)\mathrm{d}x \\
&= 1 - \int_0^1 \frac{n!}{(i-1)!(n-i)!}(F(x))^{i-1}(1-F(x+\delta))^{n-i}f(x)\mathrm{d}x
\end{aligned}$$

Because of the $x+\delta$ term, we can't actually evaluate this integral all the way up to 1, and so we are left with

$$= 1 - \int_0^{1-\delta} \frac{n!}{(i-1)!(n-i)!}(F(x))^{i-1}(1-F(x+\delta))^{n-i}f(x)\mathrm{d}x.$$

□

Uniform case

Lemma 3 (Differences between order statistics of a uniform distribution)**.** *If* $X_1, ..., X_n \overset{\text{iid}}{\sim} \text{Uniform}(0,1)$*, then each* $D_i \sim \text{Beta}(1,n)$*.*

Proof. We begin with Eq. (1), plugging in $f(x) = 1$ and $F(x) = x$:

$$h_i(d) = \int_0^{1-d} \frac{n!}{(i-1)!(n-i-1)!} x^{i-1}(1-x-d)^{n-i-1}\mathrm{d}x$$

Then we proceed with integration by parts, setting $u = x^{i-1} \implies du = (i-1)x^{i-2}$ and $dv = (1-x-d)^{n-i-1}dx \implies v = -\frac{1}{n-i}(1-x-d)^{n-i-1}$. Note that $uv|_0^{1-d} = 0$ in this case. This yields

$$= \frac{n!}{(i-1)!(n-i-1)!} \int \frac{i-1}{n-i} x^{i-2}(1-x-d)^{n-i}\mathrm{d}x$$

Then applying integration by parts again until the x^p term disappears, we get:

$$\begin{aligned}
&= \frac{n!}{(i-1)!(n-i-1)!} \frac{(i-1)!}{(n-i)\cdots(n-2)} \int_0^{1-d} (1-x-d)^{n-2}\mathrm{d}x \\
&= -\frac{n(n-1)}{n-1}(1-x-d)^{n-1}\Big|_0^{1-d} \\
&= n(1-d)^{n-1}
\end{aligned}$$

This the density function for $\mathrm{Beta}(1, n)$, completing the proof. □

Theorem 1. *Let* $X_1, ..., X_n \overset{\text{iid}}{\sim} \mathrm{Uniform}(0, 1)$. *Then for any* ϵ *and* $\delta > 0$, *there exists an* $N = O\big(\frac{-\log \epsilon}{\delta}\big)$ *such that* $P(\max_i X_{(i+1)} - X_{(i)} < \delta) \geq 1 - \epsilon$ *when* $n > N$.

Proof (sketch). Since $X_{(i+1)} - X_{(i)} = D_i \sim \mathrm{Beta}(1, n)$, $P(X_{(i+1)} - X_{(i)} < \delta) = 1 - (1-\delta)^n$. This yields

$$\begin{aligned}
P(\max_i D_i < \delta) &\geq (P(D_i < \delta))^{n-1} \\
&= (1 - (1-\delta)^n)^{n-1} \\
&\approx e^{-n\exp(-n\delta)}.
\end{aligned}$$

In the limit $n \to \infty$, this goes to 1. □

General case for one-dimensional manifolds

Theorem 2. *Let* $X_1, ..., X_n \overset{\text{iid}}{\sim} F$ *with support* $[0, 1]$, *and suppose* $f(x)$ *is continuous and* $f(x) \geq a > 0$ *everywhere on the support. Let* $D_i = X_{(i+1)} - X_{(i)}$. *Then for any* $\epsilon > 0$, *there exists* $N > 0$ *such that* $P(\max_i D_i < \delta) \geq 1 - \epsilon$ *when* $n > N$.

Proof (sketch). We start with Eq. (2):

$$P(D_i \leq \delta) = 1 - \int_0^{1-\delta} \frac{n!}{(n-i)!(i-1)!}(F(x))^{i-1}(1-F(x+\delta))^{n-i}f(x)\mathrm{d}x.$$

Making the approximation $F(x+\delta) \approx F(x) + \delta f(x)$ and bounding $f(x) \geq a$, we get:

$$P(D_i \leq \delta) \geq 1 - \int_0^{1-\delta} \frac{n!}{(n-i)!(i-1)!}(F(x))^{i-1}(1-F(x)-a\delta)^{n-i}f(x)\mathrm{d}x.$$

Then making the substitution $u = F(x) \implies du = f(x)\mathrm{d}x$, we obtain

$$1 - \int_0^{F(1-\delta)} \frac{n!}{(n-i)!(i-1)!}u^{i-1}(1-u-a\delta)^{n-i}\mathrm{d}u$$

Evaluating the integral yields

$$P(D_i < \delta) = 1 - (1-a\delta)^n + (1 - F(1-\delta) - a\delta)^n.$$

Then as before,

$$\begin{aligned} P(\max_i D_i < \delta) &= P(\text{all } D_i < \delta) \\ &= 1 - P(\text{some } D_i > \delta) \\ &\geq 1 - \sum_i^{n-1} P(D_i > \delta) \\ &= 1 - (n-1)(1-a\delta)^n + (n-1)(1 - F(1-\delta) - a\delta)^n \end{aligned}$$

This converges to 1 in the limit $n \to \infty$.

We can also approximate $F(1-\delta) \approx 1 - a\delta$, which yields $1 - (n-1)(1-a\delta)^n$. Setting this $\geq 1 - \epsilon$ ... □

Extension to multidimensional manifolds Here, we extend the results on the line to the unit hypercube. The following theorem is a direct consequence of Lemma 2 of **?**].

Theorem 3. *Let $X_1, ..., X_n \overset{\text{iid}}{\sim} F$ with support $[0,1]^r$, and $f(x) \geq a > 0$ everywhere on the support. Define E_n as the event that an η-neighborhood graph constructed from the sample is connected. Then for any $\epsilon > 0$, there exists $N = O\left(\frac{\log \epsilon\eta^r/r^{r/2}}{\log(1-\frac{a\eta^r}{r^{r/2}})}\right)$ such that $P(E_n) > 1-\epsilon$ when $n \geq N$.*

Proof (sketch). Divide the hypercube $[0,1]^r$ into a grid of sub-hypercubes of side length at most $\eta/\sqrt{r}$. E_n is satisfied if each sub-hypercube contains at least one X_i from the sample.

$$\begin{aligned} P(E_n) &= 1 - P(\text{some cells don't contain } X_i) \\ &\geq 1 - \sum_{k}^{\lceil\sqrt{r}/\eta\rceil^r} \prod_{i}^{n} P(X_i \text{ is not in the } k^{th} \text{ hypercube}) \\ &\geq 1 - \lceil\sqrt{r}/\eta\rceil^r (1 - a\eta^r/r^{r/2})^n \end{aligned}$$

Setting this $\geq 1 - \epsilon$ and solving for n yields the desired rate. □

Algorithm 1: ASE clustering for nonintersecting communities.

Data: Adjacency matrix A, number of communities K, embedding dimensions p and q.

Result: Community assignments $1, ..., K$.

1 Compute X, the ASE of A using the p most positive and q most negative eigenvalues and their corresponding eigenvectors.

2 Apply single linkage clustering with K communities on the ASE.

3.2 Intersecting Manifolds

In this section, we again consider the setting for the RDPG or GRDPG in which each community lies on a manifold in the latent space. However, this time, we do not assume that the manifolds are nonintersecting.

Theorem 4. *Let each g_k be smooth. Then K-curves clustering converges to a stationary point of the objective, $\sum_k \sum_{i \in C_k} \|x_i - g_k(t_i)\|^2$.*

Proof. K-curves clustering is a batch coordinate descent algorithm. Thus, in order to show that it converges to a stationary point, it is sufficient to show that each descent step decreases the objective function. □

K-curves clustering assumes that the functional form of g_k is known. The choice of g_k affects the difficulty of the algorithm. As a balance between flexibility and ease of

Algorithm 2: K-curves clustering.

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

Result: Community assignments $1, ..., K$, curves $g_1, ..., g_K$

Compute X, the ASE of A using the p most positive and q most negative eigenvalues and their corresponding eigenvectors.

Initialize community labels $z_1, ..., z_n$.

repeat

  for $k = 1, ..., K$ **do**

    Define X_k as the rows of X for which $z_i = k$.

    Fit curve g_k and positions t_{k_i} to X_k by minimizing $\sum_{k_i} \|x_{k_i} - g_k(t_{k_i})\|^2$.

  end

  for $k = 1, ..., K$ **do**

    Assign $z_i \leftarrow \arg\min_\ell \|x_i - g_\ell(t_i)\|^2$.

  end

until *the change in* $\sum_k \sum_{i \in C_k} \|x_i - g_k(t_i)\|^2$ *is less than* ϵ

estimation, we consider the case where each g_k is a Bezier polynomial of degree R with coefficients p_k. Then we have $g_k(t) = g(t; p_k) = \sum_{r=0}^{R} p_k^{(r)} \binom{R}{r} (1-t)^{R-r} t^r$.

Given $\{t_i\}$ and $\{z_i\}$, it is straightforward to obtain $\hat{p}_k = \arg\min_p \sum_{k_i} \|x_{k_i} - g_k(t_{k_i}; p)\|^2$

$$\hat{p}_k = (T_k^\top T_k)^{-1} T_k^\top X_k,$$

where T_k is an $n_k \times (R+1)$ matrix with rows $\begin{bmatrix} (1-t_{k_i})^R & (1-t_{k_i})^{R-1} t_{k_i} & \cdots & (1-t_{k_i}) t_{k_i}^{R-1} & t_{k_i}^R \end{bmatrix}$. Estimation of $\{t_i\}$ given $\{z_i\}$ and $\{p_k\}$ is more difficult. Each t_i can be estimated separately:

$$\hat{t}_i = \arg\min_t \|x_i - g(t; p_{z_i})\|^2. \tag{3}$$

This is equivalent to solving $0 = (x_i - g(t; p_{z_i}))^\top (\dot{g}(t; p_{z_i}))$. Setting $c^{(s)} = \sum_{r=0}^{s} (-1)^{s-r} \binom{R}{r} p_{z_i}^{(r)}$ for $s \neq 0$ and $c^{(0)} = p_{z_i}^{(0)} - x_i$, let $c = \begin{bmatrix} c^{(0)} & \cdots & c^{(R)} \end{bmatrix}^\top$. Then solving Eq. 3 is equivalent to finding the real roots of a polynomial with coefficients that are the sums of the reverse diagonals of $CD^\top$, where $C_{ij} = c_{ij}(-1)^i \binom{R}{i}$ and $D_{ij} = c_{i-1,j}(-1)^{i-1} \binom{R-1}{i-1}$.

Algorithm 3: Semi-supervised K-curves clustering.

Data: Adjacency matrix A, number of communities K, embedding dimensions p, q, stopping criterion ϵ, $m_k \leq n_k$ known community assignments for each community

Result: Community assignments $1, ..., K$, curves $g_1, ..., g_K$

Compute X, the ASE of A using the p most positive and q most negative eigenvalues and their corresponding eigenvectors.

Fit curves $g_1, ..., g_K$ using each of the $m_1, ..., m_K$ points with known community labels by minimizing $\sum_{j=1}^{m_i} \|x_j - g_k(t_j)\|^2$.

Assign labels $z_1, ..., z_n$ to each $x_1, ..., x_n$ by minimizing $\|x_i - g_k(t_i)\|^2$ for k, holding the initial known labels constant.

repeat

  for $k = 1, ..., K$ **do**

    Define X_k as the rows of X for which $z_i = k$.

    Fit curve g_k and positions t_{k_i} to X_k by minimizing $\sum_{k_i} \|x_{k_i} - g_k(t_{k_i})\|^2$.

  end

  for $k = 1, ..., K$ **do**

    Assign $z_i \leftarrow \arg\min_\ell \|x_i - g_\ell(t_i)\|^2$, holding the known initial labels constant.

  end

until *the change in* $\sum_k \sum_{i \in C_k} \|x_i - g_k(t_i)\|^2$ *is less than* ϵ

Theorem 5. *Let each* $g(\cdot; p_k)$ *be a nonintersecting Bezier polynomial of order* R*, and a GRDPG is drawn from vectors that lie on the curves. Suppose we observe the true labels of* m_k *vertices from each community, and each* $m_k > R + 1$*. Suppose further that latent vectors* $x_j = g(t_i; p_{z_j})$ *that correspond to vertices with observed labels are such that Then as* $n \to \infty$*, the proportion of misclustered vertices from* K*-curves clustering approaches* 0 *with probability* 1*.*

4 Examples

Example 1. Here, $K = 2$ with $g_1(t) = \begin{bmatrix} t^2 & 2t(1-t) \end{bmatrix}^\top$ and $g_2(t) = \begin{bmatrix} 2t(1-t) & (1-t)^2 \end{bmatrix}^\top$. We draw $n_1 = n_2 = 2^8$ points uniformly from each curve.

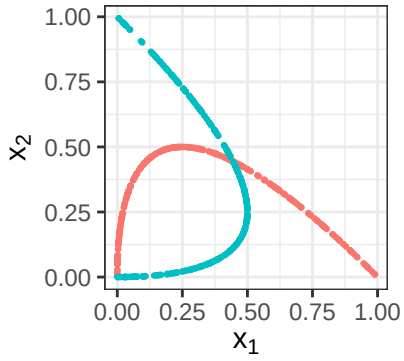


Figure 1: Latent positions, labeled by curve/community.

We draw $A \sim \text{RDPG}(X)$ and obtain the following ASE:

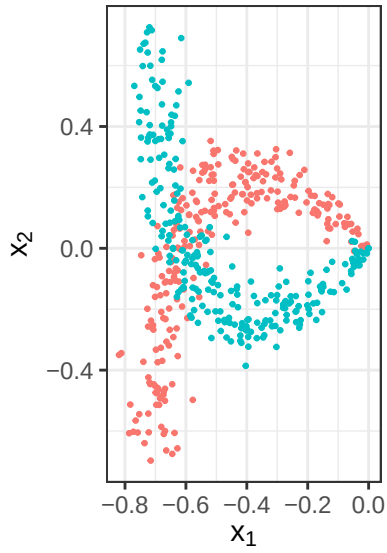


Figure 2: ASE of an RDPG drawn from the latent positions, labeled by curve/community.

We then try applying K-curves clustering to this graph. The first three are with random initial labels, forcing the intercept to be zero. The fourth initializes the labels randomly but allows the intercept to be nonzero. The fifth initializes the labels by spectral clustering with the normalized Laplacian, again forcing the intercept to be zero. The sixth also initializes

via spectral clustering but allows the intercept to be nonzero.

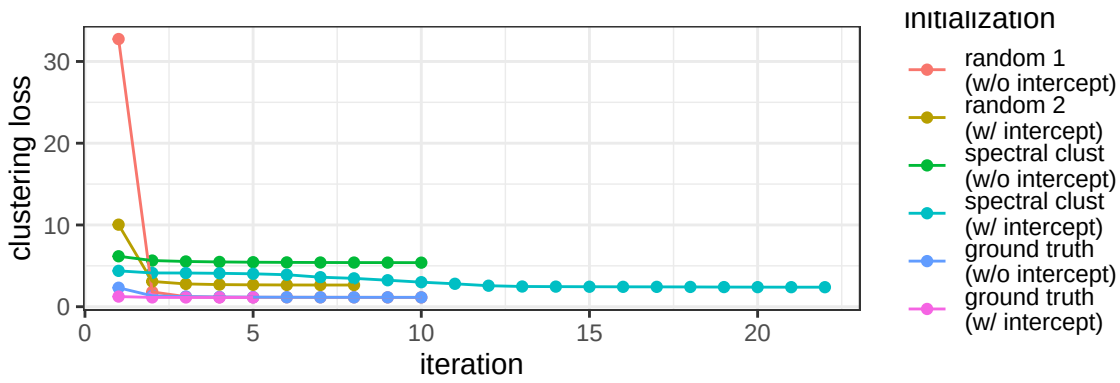


Figure 3: Clustering loss vs. iteration for each run of K-curve clustering.

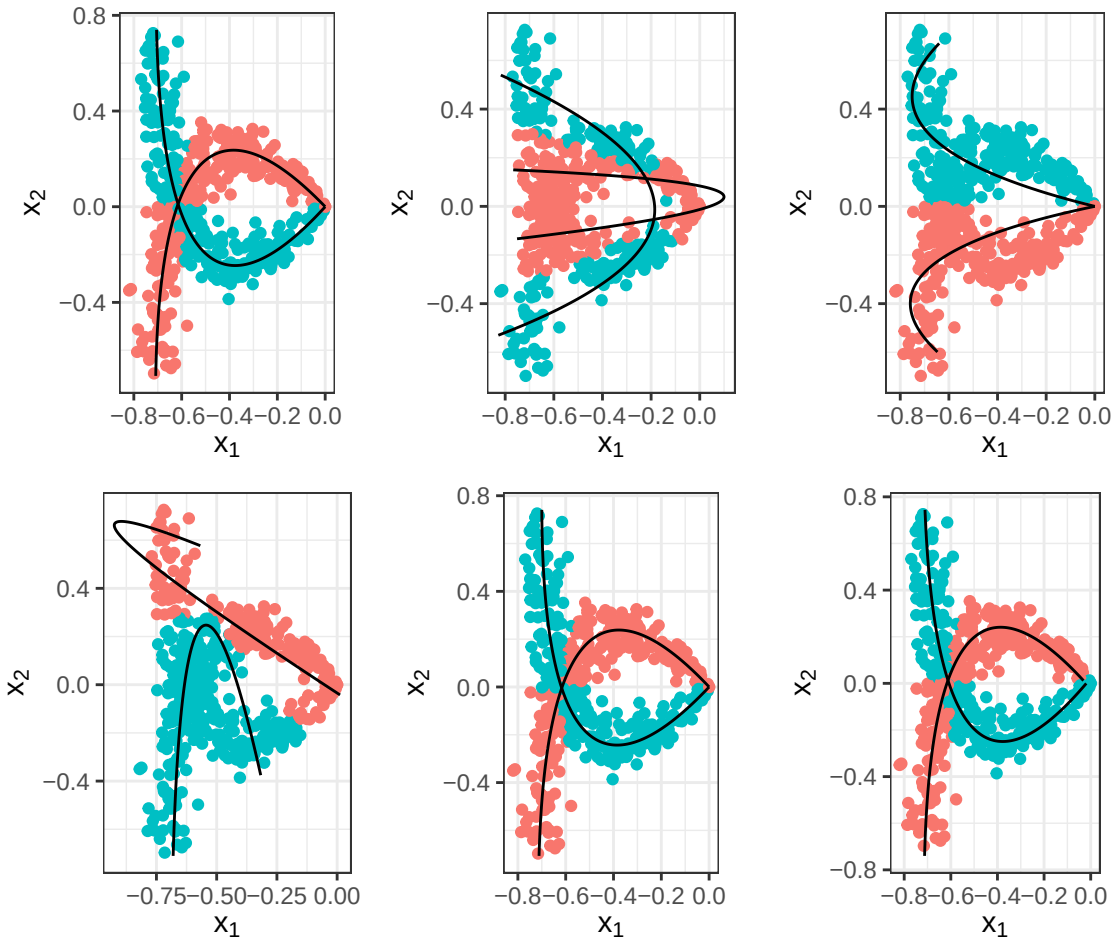
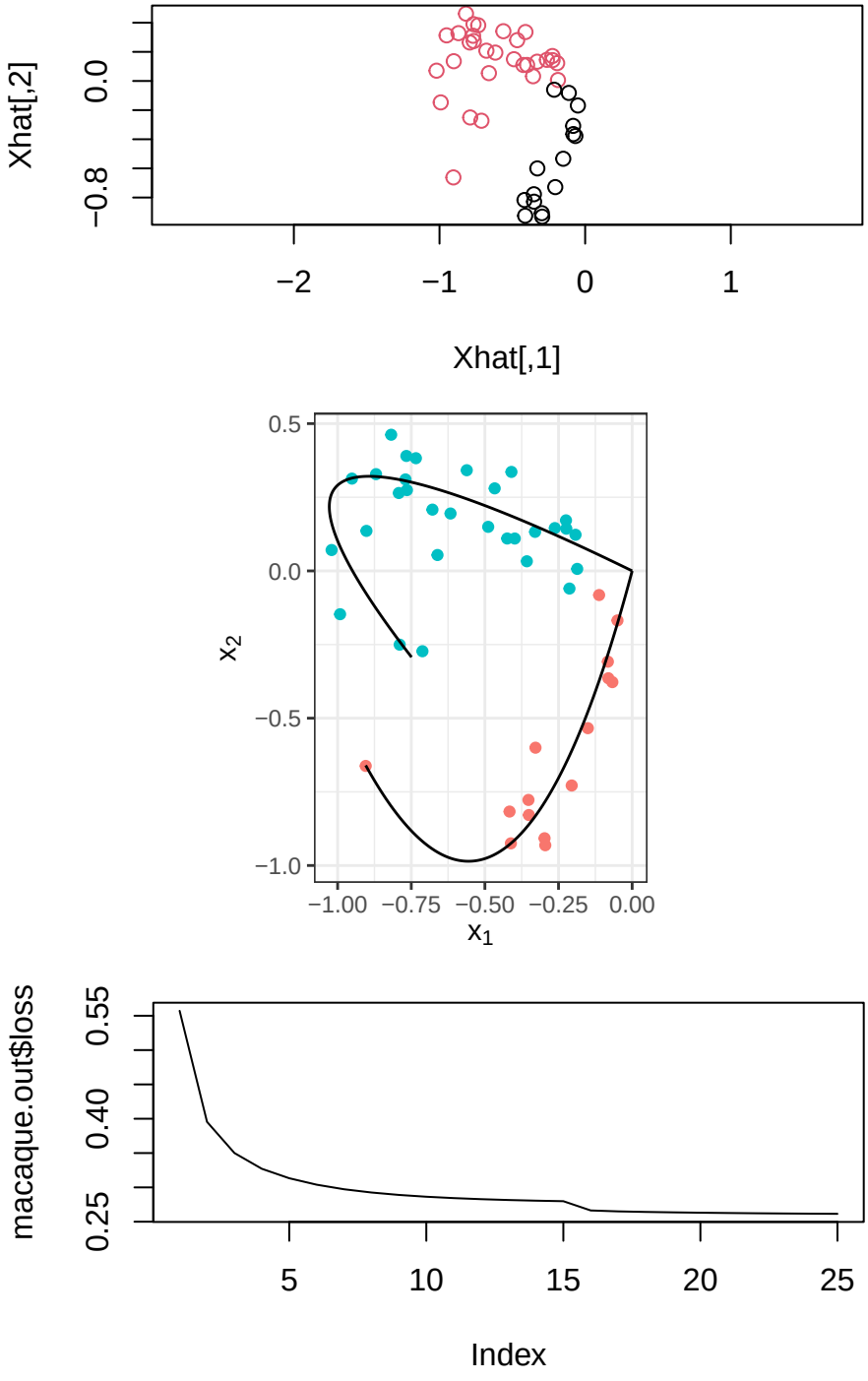
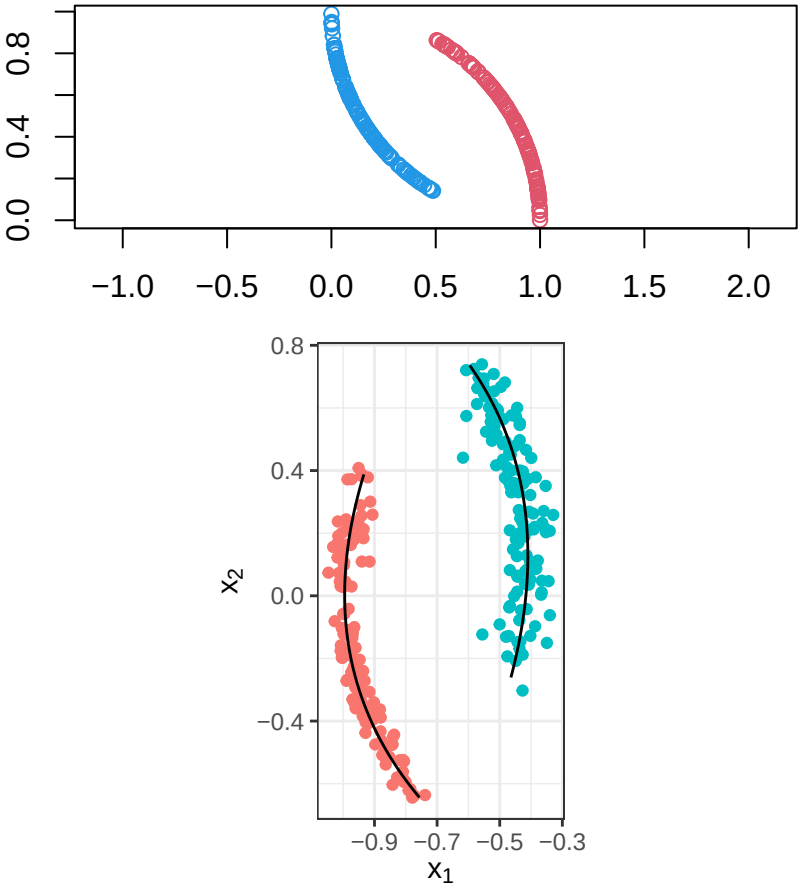


Figure 4: ASE labeled by estimated community labels for each initialization strategy.

Example 2 (Macaque visuotactile brain areas and connections)**.** [1]



Example 3 (Non-intersecting curves).



5 Simulation Study

References

[1] Négyessy, L., Nepusz, T., Kocsis, L. and Bazsó, F. [2006], 'Prediction of the main cortical areas and connections involved in the tactile function of the visual cortex by network analysis', *European Journal of Neuroscience* **23**(7), 1919–1930.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-9568.2006.04678.x*