Consistency of Graphical Model-based Clustering: Robust Clustering using Bayesian Spanning Forest

Yu Zheng
Department of Statistics, University of Florida, zheng.yu@ufl.edu
(Joint work with Leo L. Duan and Arkaprava Roy)

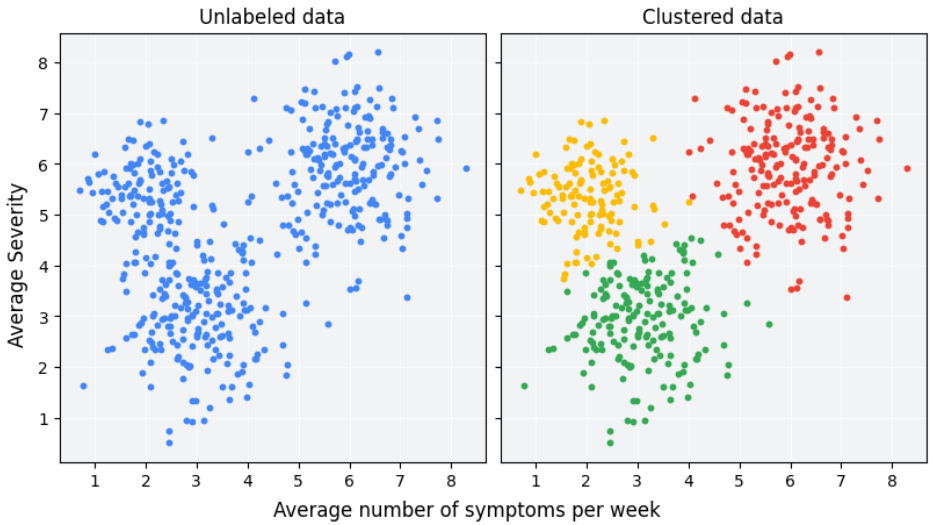
UF UNIVERSITY of FLORIDA

Table of Contents

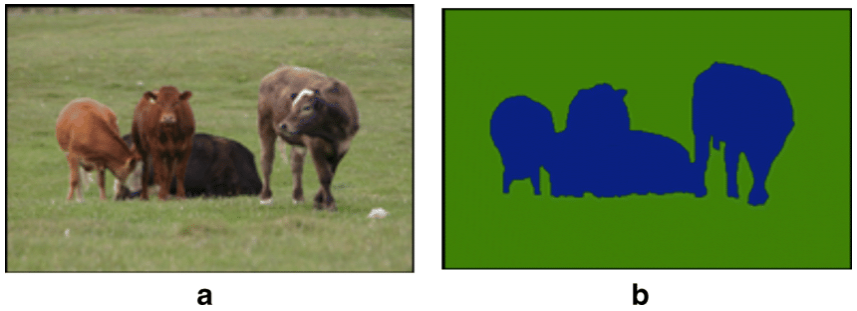
Introduction

What is Clustering?

- Clustering partitions data into groups where points in the same group are more similar to each other than to points in other groups.
- Applications include image segmentation, anomaly detection, and biological data analysis.
- Clustering is an unsupervised learning task and can be done using methods like k-means, hierarchical clustering, etc.



(a) Clustering Example



(b) Image Segmentation

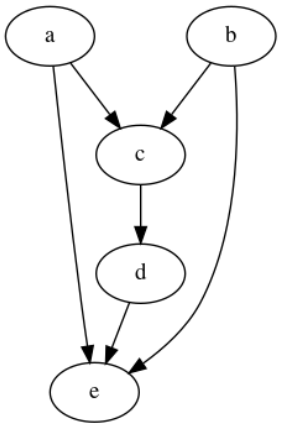
Clustering Notations

- Suppose we have data points $y^{(n)} = \{y_1, \ldots, y_n\}$.
- Let $[N]$ denote the index set $\{1, 2, \ldots, N\}$ for any positive integer N. The parameter of interest is a **partition** of $[n]$, $\mathcal{V}_n = (V_1, \ldots, V_K)$, representing K clusters such that:

$$\bigcup_{k=1}^{K} V_k = [n], \quad V_k \cap V_{k'} = \emptyset \quad \text{for} \quad k \neq k'.$$

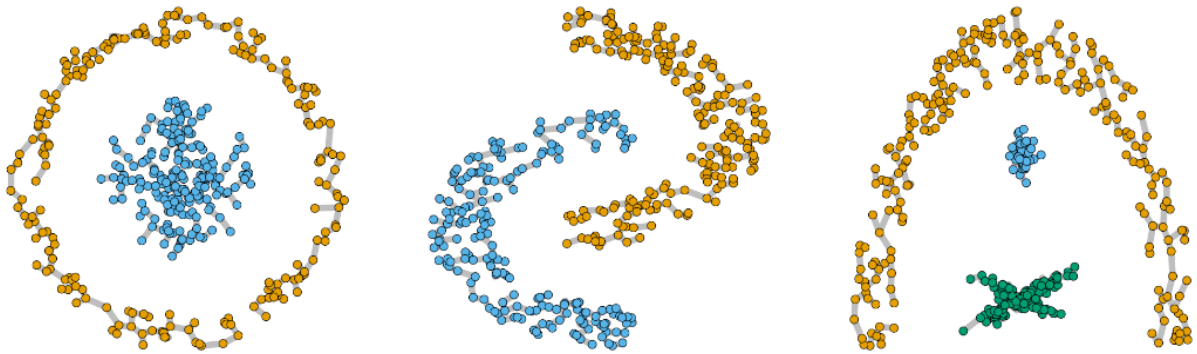
Graphical Model

- A graphical model, particularly based on **directed acyclic graphs (DAGs)**, offers a flexible way to specify the likelihood for clustering.
- **Generative Model Based on DAGs**:
  - Each cluster can be represented by a DAG-based likelihood.
  - The union of these DAGs, combined with a prior distribution on the disjoint union of DAGs, results in a generative model.
  - This model fits naturally into the Bayesian framework for statistical inference.



Forest-Based Graphical Model

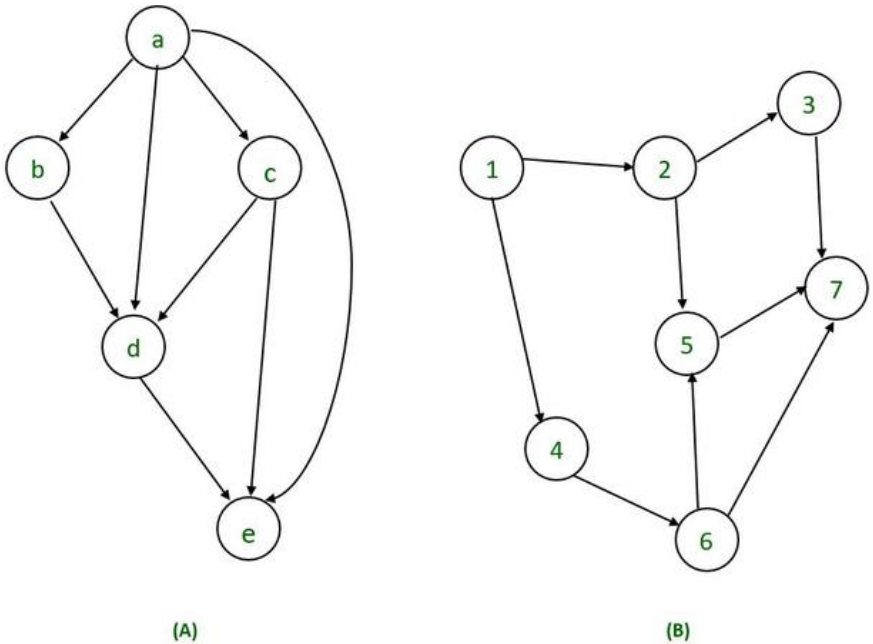
- A spanning forest (a union of spanning trees) offers a natural framework for clustering: Nodes connected by edges within the same tree are grouped into a cluster.
- Recent advancements in efficient algorithms for sampling and estimating trees have significantly enhanced the use of spanning forest-based models for clustering (Luo et al., 2021; Zhao Tang Luo and Mallick, 2024).



UF UNIVERSITY of FLORIDA

BSF Model

- For each cluster V_k, we associate a rooted spanning tree, $G_k = (V_k, E_k, k^*)$, where E_k represents the set of edges, and k^* is the root node within V_k.
- We define $\mathcal{E}_\mathcal{V} = \{E_1, \ldots, E_K\}$ as the collection of edge sets and $\mathcal{R}_\mathcal{V} = \{1^*, \ldots, K^*\}$ as the collection of root nodes for all clusters.
- Thus, $(\mathcal{V}_n, \mathcal{E}_{\mathcal{V}_n}, \mathcal{R}_{\mathcal{V}_n})$ forms a (rooted) **spanning forest**, which is the disjoint union of K component rooted spanning trees, each corresponding to a cluster.



BSF Model

Consider the likelihood based on the spanning forest for data $y^{(n)} = \{y_1, \ldots, y_n\}$:

$$P(y^{(n)} \mid \mathcal{V}_n, \mathcal{E}_{\mathcal{V}_n}, \mathcal{R}_{\mathcal{V}_n}, \theta) = \prod_{k=1}^{K} \left[r(y_{k^*}; \theta) \prod_{(i,j) \in G_k} f(y_i \mid y_j; \theta) \right],$$

where the model is associated with a generative process:

- $r(\cdot; \theta)$ is the probability kernel of the **root distribution**, responsible for generating the first data point in each cluster.
- $f(\cdot \mid y_j; \theta)$ is the probability kernel of the **leaf distribution**, generating subsequent data points in a cluster given an existing data point.

UF UNIVERSITY of FLORIDA

Bayesian Spanning Forest (BSF) Model

- **Challenge**: Considering all possible spanning forests in a graph with n nodes leads to a **vast parameter space**, which can result in computational inefficiency and poor estimation.
- In response to this challenge, Duan and Roy (2024) propose a novel **Bayesian spanning forest (BSF) model** based on the key idea that the primary interest in clustering is the **partition of nodes** $\mathcal{V}_n$, NOT the directed edges within each DAG.
- Their model treats the edges in each DAG as latent variables and focuses on the **integrated posterior**, where the edges are marginalized out.

BSF Model

Let $\Pi_0(K, \mathcal{V}_n)$ be a partition probability function serving as the prior, $\Pi_0(\mathcal{E}_{\mathcal{V}_n}, \mathcal{R}_{\mathcal{V}_n} \mid \mathcal{V}_n)$ as the conditional prior for edges and roots. We derive the posterior distribution of $\mathcal{V}_n$ given the data $y^{(n)}$ as follows:

$$\Pi(\mathcal{V}_n \mid y^{(n)}) = \frac{\sum_{\mathcal{E}_{\mathcal{V}_n}, \mathcal{R}_{\mathcal{V}_n}} P(y^{(n)} | \mathcal{V}_n, \mathcal{E}_{\mathcal{V}_n}, \mathcal{R}_{\mathcal{V}_n}) \Pi_0(K, \mathcal{V}_n) \Pi_0(\mathcal{E}_{\mathcal{V}_n}, \mathcal{R}_{\mathcal{V}_n} | \mathcal{V}_n)}{\sum_{\mathcal{V}'_n, \mathcal{E}_{\mathcal{V}'_n}, \mathcal{R}_{\mathcal{V}'_n}} P(y^{(n)} | \mathcal{V}'_n, \mathcal{E}_{\mathcal{V}'_n}, \mathcal{R}_{\mathcal{V}'_n}) \Pi_0(K, \mathcal{V}'_n) \Pi_0(\mathcal{E}_{\mathcal{V}'_n}, \mathcal{R}_{\mathcal{V}'_n} | \mathcal{V}'_n)}.$$

Bayesian Spanning Forest Model: Empirical and Theoretical Performance

- The empirical performance of the Bayesian spanning forest model has been extensively studied in Duan and Roy (2024).
- Theoretically, the model's good performance is attributed to:
  - **Asymptotic equivalence** between the posterior mode (given a number of clusters) and the estimate from the **normalized spectral clustering algorithm** (Ng et al., 2001).
  - **Clustering consistency** when data are generated from a forest graphical model.
- However, it remains unknown whether the integrated posterior of the node partition is **robustly consistent**.
  - Specifically, if the data-generating mechanism **differs** from the specified graphical model, can the posterior still concentrate on the ground-truth partition for separable data points?

UF UNIVERSITY of FLORIDA

Problem of Interest

Clustering Consistency

Definition (Posterior Consistency for Clustering)

The posterior $\Pi(\mathcal{V}_n|y^{(n)})$ is said to be *weakly consistent* at $(V_1^{0,n}, \ldots, V_{K_0}^{0,n})$ if:

$$\Pi(\mathcal{V}_n \neq (V_1^{0,n}, \ldots, V_{K_0}^{0,n})|y^{(n)}) \overset{n\to\infty}{\to} 0$$

in $P_0^{(\infty)}$-probability. It is *strongly consistent* if this convergence occurs almost surely.

- $\Pi(\mathcal{V}_n|y^{(n)})$: The posterior distribution of the partition $\mathcal{V}_n$ given the data.
- $(V_1^{0,n}, \ldots, V_{K_0}^{0,n})$: The oracle clustering (true partition).
- $P_0^{(\infty)}$: The probability space of $y^{(\infty)}$, representing the oracle data-generating mechanism.

UF UNIVERSITY of FLORIDA

Problem of Interest

- **Oracle Clustering:** $(V_1^{0,n}, \ldots, V_{K_0}^{0,n})$ is the true partition of the data.
- **Oracle Data-Generating Mechanism:**

$$y_i \overset{indep}{\sim} G^0_{z_i^*}, \quad i = 1, \ldots, n.$$

- **Posterior of the Partition $\mathcal{V}_n$ under the BSF Model:**

$$\Pi(\mathcal{V}_n = (V_1, \ldots, V_K) \mid y^{(n)}) = C_n \cdot (\delta\lambda)^K \prod_{k=1}^{K} \left| L_{V_k} + \frac{1}{n_k} J \right|$$

where L_{V_k} is the Laplacian matrix of cluster V_k, and n_k is the size of cluster V_k.

- We seek to determine the mild assumptions required to achieve **strong clustering consistency**:

$$\Pi\left(\mathcal{V}_n \neq (V_1^{0,n}, \ldots, V_{K_0}^{0,n}) \mid y^{(n)}\right) \to 0 \quad \text{as} \quad n \to \infty \quad P_0^{(\infty)}\text{-almost surely}$$

UF UNIVERSITY of FLORIDA

Main Results

Two Mild Assumptions

- **Assumption 1:** The oracle partition $(V_1^{0,\infty}, \ldots, V_{K_0}^{0,\infty})$ satisfies $|V_k^{0,\infty}| > 0$ for $k \in [K_0]$.
- **Assumption 2:** There exist positive constants $C_1, C_2 > 0$ such that, for sufficiently large n, $C_1\delta_n \leq \min_{i\in[n]} r(y_i) \leq \max_{i\in[n]} r(y_i) \leq C_2\delta_n$.

Notations

We define $f_{st}^{(n)} = f(y_t \mid y_s; \theta_n)$ as the conditional probability kernel between two nodes y_s and y_t.

- The magnitude of $f_{st}^{(n)}$ quantifies the probabilistic closeness or association between the nodes.
- The larger $f_{st}^{(n)}$, the more likely two nodes arise from the same cluster.
- A key step in our analysis is controlling these conditional kernels efficiently using θ_n.

Main Results

Let

$$\mathcal{D}_{\phi}^{(n)} := \Bigg\{ y^{(n)} : \frac{\max_{s \not\sim t; s,t \in [n]} f_{st}^{(n)}}{\delta_n \lambda_n} \leq c_1 (K_0 - 1 + \iota_1)^{-n}, \quad \frac{\delta_n \lambda_n}{\min_{s' \sim t'; s',t' \in [n]} f_{s't'}^{(n)}} \leq c_2 (K_0 + 1 + \iota_2)^{-n} \Bigg\}.$$

Theorem (General Strong Clustering Consistency under BSF)

$$\Pi \left(\mathcal{V}_n \neq (V_1^{0,n}, \ldots, V_{K_0}^{0,n}) \mid y^{(n)} \right) \to 0 \quad \textit{as} \quad n \to \infty \quad P_0^{(\infty)}\textit{-almost surely},$$

if $\sum_{n=1}^{\infty} P_0^{(\infty)}(y^{(n)} \notin \mathcal{D}_{\phi}^{(n)}) < \infty$ *for a fixed constant* ϕ.

Gaussian-BSF

- For simplicity, we denote $d_{st} := d(y_s, y_t)$.
- Consider $f_{st}^{(n)} = \zeta(\sigma^{0,n}) \exp\left\{-\frac{d_{st}^2}{2(\sigma^{0,n})^2}\right\}$.
- Plugging this specific form of $f_{st}^{(n)}$ for Gaussian-BSF into the conditions of $\mathcal{D}_{\phi}^{(n)}$ and rearranging terms, we have:

$$\mathcal{D}_{\phi}^{(n)} = \left\{ y^{(n)} : \min_{s \not\sim t} d_{st}^2 \geq a_n, \quad \max_{s' \sim t'} d_{s't'}^2 \leq b_n \right\},$$

where:

$$\begin{cases} a_n = 2(\sigma^{0,n})^2 \left[n \log(K_0 - 1 + \iota_1) - \log(\delta_n \lambda_n) + \log(\zeta(\sigma^{0,n})) - \log(c_1) \right], \\ b_n = 2(\sigma^{0,n})^2 \left[-n \log(K_0 + 1 + \iota_2) - \log(\delta_n \lambda_n) + \log(\zeta(\sigma^{0,n})) + \log(c_2) \right]. \end{cases}$$

- For Euclidean distance, $\log(\zeta(\sigma^{0,n})) = -p \log(\sqrt{2\pi}\sigma^{0,n})$. Said et al. (2022) give expressions for $\zeta(\sigma^{0,n})$ for a wide range of homogenous Riemannian manifolds.

Interpretation of the Conditions on the Oracle

$$\mathcal{D}_{\phi}^{(n)} = \left\{ y^{(n)} : \min_{s \not\sim t} d_{st}^2 \geq a_n, \quad \max_{s' \sim t'} d_{s't'}^2 \leq b_n \right\}$$

- The **minimum distance** between any two points from different oracle clusters must be bounded below by a sequence a_n, ensuring clear separation between clusters.
- The **maximum distance** between any two points within the same oracle cluster must be bounded above by b_n, ensuring that points within a cluster remain closely grouped.

These conditions do not need to hold for $y^{(n)}$ at every n, but the probability that they hold should approach one as $n \to \infty$.

Main Results on Gaussian-BSF

Theorem (Clustering consistency under Gaussian-BSF model)

For $f_{st}^{(n)} = \zeta(\sigma_n)\exp\left\{-d_{st}^2/[2\sigma_n^2]\right\}$,
$\Pi(\mathcal{V}_n \neq (V_1^{0,n},\ldots,V_{K_0}^{0,n})|y^{(n)}) \to 0$ *as* $n\to\infty$ $\quad P_0^{(\infty)} -$ *almost surely, if there exists* $\phi = (c_1, c_2, \iota_1, \iota_2) \in \mathbb{R}_+^4$ *such that*

$$\sum_{n=1}^{\infty} n^2 \max_{k\neq\ell; k,\ell\in[K_0]} P(D_{k\ell}^2 < a_n) < \infty,$$

$$\sum_{n=1}^{\infty} n^2 \max_{k'\in[K_0]} P(D_{k'k'}^2 > b_n) < \infty.$$

Two Concrete Examples

Concrete Example I

Theorem (Consistency when using Gaussian-BSF for clustering data from Gaussian distributions)

Suppose $(V_1^{0,\infty}, \ldots, V_{K_0}^{0,\infty})$ *is the oracle clustering for* $y^{(\infty)} = \{y_1, y_2, \ldots\}$, *and* $f_{st}^{(n)} = (\sqrt{2\pi}\sigma_n)^{-p} \exp\left\{-\|y_s - y_t\|_2^2/[2\sigma_n^2]\right\}$. *Suppose* $y_i \overset{indep}{\sim} N(\mu_k, \Sigma_k)$ *if* $y_i \in V_k^{0,\infty}$. *Set* $\Lambda_{\max} := \max_{k\in[K_0]} \lambda_{\max}(\Sigma_k)$ *and* $D_{\mu,\min} := \min_{k,\ell\in[K_0],k\neq\ell} \|\mu_k - \mu_\ell\|_2$. *Assume that*

(i) *Assumptions 1 and 2 hold;*

(ii) $\delta_n\lambda_n \asymp (K_0 + 1 + \iota)^{-n}$ *for a fixed constant* $\iota > 0$;

(iii) $\log(\sigma_n) = o(n)$;

(iv) $n\sigma_n^2/D_{\mu,\min}^2 = o(1)$;

(v) $\Lambda_{\max}\log(n)/n\sigma_n^2 = o(1)$.

Then we have $\Pi(\mathcal{V}_n \neq (V_1^{0,1}, \ldots, V_{K_0}^{0,n})|y^{(n)}) \to 0$ *as* $n \to \infty \quad P_0^{(\infty)} -$ *almost surely.*

Discussion of the Theorem

Corollary

Define the signal-to-noise ratio (SNR) as $SNR := D_{\mu,\min}/\sqrt{\Lambda_{\max}}$. *If* $SNR/\sqrt{\log(n)} \to \infty$ *as* $n \to \infty$, *there exists* σ_n^2 *that leads to the clustering consistency under the BSF model. Specially, one can take* $n\sigma_n^2 = (D_{\mu,\min}^2)^{\alpha}(\Lambda_{\max}\log(n))^{1-\alpha}$ *for any* $\alpha \in (0,1)$.

UF UNIVERSITY of FLORIDA

Concrete Example II

- We can generalize to the cases when the oracle is a non-Gaussian mixture.
- The robustness of BSF model further allows us to present this result for a more general setting where the truth is assumed to be a mixture of **object-valued** distributions, supported on a metric space (Ω, d).

Concrete Example II

Theorem (Consistency when using Gaussian-BSF for clustering data from general mixture oracle)

Suppose $(V_1^{0,\infty}, \ldots, V_{K_0}^{0,\infty})$ *is the oracle clustering for* $y^{(\infty)} = \{y_1, y_2, \ldots\}$ *and* $f_{st}^{(n)} = \zeta(\sigma_n)\exp\{-d_g^2(y_s, y_t)/[2\sigma_n^2]\}$. *Suppose* $y_i \overset{indep}{\sim} G_k^0$ *if* $y_i \in V_k^{0,\infty}$ *with* $\{G_k^0\}_{k\in[K_0]}$ *standing for the family of probability measures on* Ω. *Assume that*

(i) *Let* $\mu_k := \arg\min_z \mathbb{E}_{x\sim G_k^0} d^2(z, x)$ *be the unique Fréchet mean under the density* G_k^0 *and* $D_{\mu,\min} := \min_{k,\ell\in[K_0],k\neq\ell} d(\mu_k, \mu_\ell)$.

(ii) *Assumptions 1 and 2 hold;*

(iii) $\delta_n\lambda_n \asymp (K_0 + 1 + \iota)^{-n}$ *for a fixed constant* $\iota > 0$;

(iv) $P_{G_k^0}(d(X, \mu_k) > R) \leq \exp(-CR^\nu)$ *for fixed constants* $C, \nu > 0$, *any* $k \in [K_0]$ *and* $R \geq 0$;

(v) $(\log(n))^{2/\nu}/n\sigma_n^2 = o(1)$ *with* $\log(\zeta(\sigma_n)) = o(n)$;

(vi) $n\sigma_n^2/D_{\mu,\min}^2 = o(1)$,

then we have $\Pi(\mathcal{V}_n \neq (V_1^{0,1}, \ldots, V_{K_0}^{0,n})|y^{(n)}) \to 0$ *as* $n \to \infty$ $P_0^{(\infty)} -$ *almost surely.*

Discussion of the Theorem

In this case, the impact of the variances of the oracle distributions is incorporated in (iv) as the control on the tail probabilities. Similar to the discussion following Theorem 4, under some mild conditions, if the minimum separation satisfies $D_{\mu,\min}/(\log(n))^{1/\nu} \to \infty$, then there exists σ_n^2 that leads to clustering consistency.

Wrap-up and Discussions

Key Contributions

- **Robust model specification:** Our findings present a practical approach to bypassing the need for an entirely correct specification of the mixture component distribution.
- **Simultaneous recovery:** We demonstrate that the posterior achieves **strong consistency**, recovering both the number of clusters and the true clustering labels simultaneously. In contrast, previous work on mixture models often requires additional conditions, such as restricting the family of distributions or assuming the number of clusters is known.

UF UNIVERSITY of FLORIDA

Discussion (Part 1)

- For consistency theory, we focus on the case where oracle clustering is asymptotically identifiable via n-dependent separation conditions, as similarly posited in recent clustering consistency theory on infinite mixture (Ascolani et al., 2022).
- One interesting extension would be to explore cases where the oracle clustering is only partially identifiable, allowing for Bayes misclustering error. However, reaching the Bayes error would likely require stronger assumptions than those used in this work.

Discussion (Part 2)

- We chose the spanning forest graph for graphical model-based clustering due to its strong empirical performance and mathematical tractability.
- Future work could explore a broader class of graphs, including those where edge formation is influenced by external covariates. Expanding the theory to these new model-based clustering methods would be an interesting direction.

Thank You!

Feel free to reach out for any further questions.

Email: zheng.yu@ufl.edu

UF UNIVERSITY of FLORIDA

Reference I

Ascolani, F., A. Lijoi, G. Rebaudo, and G. Zanella (2022, Sep). Clustering consistency with dirichlet process mixtures. *Biometrika 110*(3), 551–558.

Duan, L. L. and A. Roy (2024). Spectral clustering, bayesian spanning forest, and forest process. *Journal of the American Statistical Association 119*(547), 2140–2153.

Luo, Z. T., H. Sang, and B. Mallick (2021). A bayesian contiguous partitioning method for learning clustered latent variables. *Journal of Machine Learning Research 22*(37), 1–52.

Ng, A., M. Jordan, and Y. Weiss (2001). On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems 14*, 849–856.

Said, S., C. Mostajeran, and S. Heuveline (2022). Gaussian distributions on riemannian symmetric spaces of nonpositive curvature. In *Handbook of Statistics*, Volume 46, pp. 357–400. Elsevier.

Zhao Tang Luo, H. S. and B. Mallick (2024). A nonstationary soft partitioned gaussian process model via random spanning trees. *Journal of the American Statistical Association 119*(547), 2105–2116.