Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Spectral Clustering, Bayesian Spanning Forest, and Forest Process

Leo L. Duan, Arkaprava Roy & For the Alzheimer's Disease Neuroimaging Initiative

To cite this article: Leo L. Duan, Arkaprava Roy & For the Alzheimer's Disease Neuroimaging Initiative (24 Aug 2023): Spectral Clustering, Bayesian Spanning Forest, and Forest Process, Journal of the American Statistical Association, DOI: 10.1080/01621459.2023.2250098

To link to this article: https://doi.org/10.1080/01621459.2023.2250098

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 00, NO. 0, 1–14: Theory and Methods
https://doi.org/10.1080/01621459.2023.2250098

Spectral Clustering, Bayesian Spanning Forest, and Forest Process

Leo L. Duan[a], Arkaprava Roy[b], and For the Alzheimer's Disease Neuroimaging Initiative

[a]Department of Statistics, University of Florida, Gainesville, FL; [b]Department of Biostatistics, University of Florida, Gainesville, FL

ABSTRACT

Spectral clustering views the similarity matrix as a weighted graph, and partitions the data by minimizing a graph-cut loss. Since it minimizes the across-cluster similarity, there is no need to model the distribution within each cluster. As a result, one reduces the chance of model misspecification, which is often a risk in mixture model-based clustering. Nevertheless, compared to the latter, spectral clustering has no direct ways of quantifying the clustering uncertainty (such as the assignment probability), or allowing easy model extensions for complicated data applications. To fill this gap, we propose the Bayesian forest model as a generative graphical model for spectral clustering. This is motivated by our discovery that the posterior connecting matrix in a forest model has almost the same leading eigenvectors, as the ones used by normalized spectral clustering. To induce a distribution for the forest, we develop a "forest process" as a graph extension to the urn process, while we carefully characterize the differences in the partition probability. We derive a simple Markov chain Monte Carlo algorithm for posterior estimation, and demonstrate superior performance compared to existing algorithms. We illustrate several model-based extensions useful for data applications, including high-dimensional and multi-view clustering for images. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received August 2022
Accepted August 2023

KEYWORDS
Graphical model clustering;
Model-based clustering;
Normalized graph-cut;
Partition probability function

1. Introduction

Clustering aims to partition data $y_1, \ldots, y_n$ into disjoint groups. There is a large literature ranging from various algorithms such as K-means and DBSCAN (MacQueen 1967; Ester et al. 1996; Frey and Dueck 2007) to mixture model-based approaches (reviewed by Fraley and Raftery 2002). In the Bayesian community, model-based approaches are especially popular. To roughly summarize the idea, we view each y_i as generated from a distribution $\mathcal{K}(\cdot \mid \theta_i)$, where $(\theta_1, \ldots, \theta_n)$ are drawn from a discrete distribution $\sum_{k=1}^{K} w_k \delta_{\theta_k^*}(\cdot)$, with w_k as the probability weight, and $\delta_{\theta_k^*}$ as a point mass at θ_k^*. With prior distributions, we could estimate all the unknown parameters (θ_k^*'s, w_k's, and K) from the posterior.

The model-based clustering has two important advantages. First, it allows important uncertainty quantification such as the probability for cluster assignment c_i, $\Pr(c_i = k \mid y_i)$, as a probabilistic estimate that y_i comes from the kth cluster ($c_i = k \Leftrightarrow \theta_i = \theta_k^*$). Different from commonly seen asymptotic results in statistical estimation, the clustering uncertainty does not always vanish even as $n \to \infty$. For example, in a two-component Gaussian mixture model with equal covariance, for a point y_i at nearly equal distances to two cluster centers, we would have both $\Pr(c_i = 1 \mid y_i)$ and $\Pr(c_i = 2 \mid y_i)$ close to 50% even as $n \to \infty$. For a recent discussion on this topic as well as how to quantify the partition uncertainty, see Wade and Ghahramani (2018) and the references within. Second, the model-based clustering can be easily extended to handle more complicated modeling tasks. Specifically, since there is a probabilistic process associated with the clustering, it is straightforward to modify it to include useful dependency structures. We list a few examples from a rich literature: Ng et al. (2006) used a mixture model with random effects to cluster correlated gene-expression data, Müller and Quintana (2010), Park and Dunson (2010), and Ren et al. (2011) allowed the partition to vary according to some covariates, Guha and Baladandayuthapani (2016) simultaneously clustered the predictors and use them in high-dimensional regression.

On the other hand, model-based clustering has its limitations. Primarily, one needs to carefully specify the density/mass function $\mathcal{K}$, otherwise, it will lead to unwanted results and difficult interpretation. For example, Coretto and Hennig (2016) demonstrated the sensitivity of the Gaussian mixture model to non-Gaussian contaminants, Miller and Dunson (2018) and Cai, Campbell, and Broderick (2021) showed that when the distribution family of $\mathcal{K}$ is misspecified, the number of clusters would be severely overestimated. It is natural to think of using more flexible parameterization for $\mathcal{K}$, in order to mitigate the risk of model misspecification. This has motivated many interesting works, such as modeling $\mathcal{K}$ via skewed distribution (Frühwirth-Schnatter and Pyne 2010; Lee and McLachlan 2016), unimodal distribution (Rodríguez and Walker 2014), copula (Kosmidis and Karlis 2016), mixture of mixtures (Malsiner-Walli, Frühwirth-Schnatter, and Grün 2017), among others. Nevertheless, as the flexibility of $\mathcal{K}$ increases, the modeling and computational burdens also increase dramatically.

CONTACT Leo L. Duan leo.duan.work@gmail.com Department of Statistics, University of Florida, Gainesville, FL
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

In parallel to the above advancements in model-based clustering, spectral clustering has become very popular in machine learning and statistics. Von Luxburg (2007) provided a useful tutorial on the algorithms and a review of recent works. On clustering point estimation, spectral clustering has shown good empirical performance for separating non-Gaussian and/or manifold data, without the need to directly specify the distribution for each cluster. Instead, one calculates a matrix of similarity scores between each pair of data, then uses a simple algorithm to find a partition that approximately minimizes the total loss of similarity scores across clusters (adjusted with respect to cluster sizes). This point estimate is found to be not very sensitive to the choice of similarity score, and empirical solutions have been proposed for tuning the similarity and choosing the number of clusters (Zelnik-Manor and Perona 2005; Shi, Belkin, and Yu 2009). There is a rapidly growing literature of frequentist methods on further improving the point estimate (Chi et al. 2007; Rohe, Chatterjee, and Yu 2011; Kumar, Rai, and Daume 2011; Lei and Rinaldo 2015; Han, Tong, and Fan 2021; Lei and Lin 2022; among others), although, in this article, we focus on the Bayesian perspective and aim to characterize the probability distribution.

Due to the algorithmic nature, spectral clustering cannot be directly used in model-based extension, or produce uncertainty quantification. This has motivated a large Bayesian literature. There have been several works trying to quantify the uncertainty around the spectral clustering point estimate. For example, since the spectral clustering algorithm can be used to estimate the community memberships in a stochastic block model, one could transform the data into a similarity matrix, then treat it as if generated from a Bayesian stochastic block model (Snijders and Nowicki 1997; Nowicki and Snijders 2001; McDaid et al. 2013; Geng et al. 2019). Similarly, one could take the Laplacian matrix (a transform of the similarity used in spectral clustering) or its spectral decomposition, and model it in a probabilistic framework (Socher, Maas, and Manning 2011; Duan, Michailidis, and Ding 2023).

Broadly speaking, we can view these works as following the recent trend of robust Bayesian methodology, in conditioning the parameter of interest (clustering) on an insufficient statistic (pairwise summary statistics) of the data. See Lewis, MacEachern, and Lee (2021) for recent discussions. Pertaining to Bayesian robust clustering, one gains model robustness by avoiding putting any parametric assumption on within-cluster distribution $\mathcal{K}(\cdot \mid \theta_k^*)$; instead, one models the pairwise information that often has an arguably simple distribution. Recent works include the distance-based Pólya urn process (Blei and Frazier 2011; Socher, Maas, and Manning 2011), Dirichlet process mixture model on Laplacian eigenmaps (Banerjee, Akbani, and Baladandayuthapani 2015), Bayesian distance clustering (Duan and Dunson 2021a), generalized Bayes extension of product partition model (Rigon, Herring, and Dunson 2023).

This article follows this trend. Instead of modeling y_i's as conditionally independent (or jointly dependent) from a certain within-cluster distribution $\mathcal{K}(\cdot \mid \theta_k^*)$, we choose to model y_i as dependent on another point y_j that is close by, provided y_i and y_j are from the same cluster. This leads to a Markov graphical model based on a spanning forest, a graph consisting of multiple disjoint spanning trees (each tree as a connected subgraph without cycles). The spanning forest itself is not new to statistics. There has been a large literature on using spanning trees and forests for graph estimation, such as Meila and Jordan (2000), Meilă and Jaakkola (2006), Edwards, De Abreu, and Labouriau (2010), Byrne and Dawid (2015), Duan and Dunson (2021b), and Luo, Sang, and Mallick (2021). Nevertheless, a key difference between graph estimation and graph-based clustering is that—the former aims to recover both the node partition and the edges characterizing dependencies, while the latter only focuses on estimating the node partition alone (equivalent to clustering). Therefore, a distinction of our study is that we will treat the edges as a nuisance parameter/latent variable, while we will characterize the node partition in the marginal distribution.

Importantly, we formally show that by marginalizing the randomness of edges, the point estimate on the node partition is provably close to the one from the normalized spectral clustering algorithm. As the result, the spanning forest model can serve as the probabilistic model for the spectral clustering algorithm—this relationship is analogous to the one between the Gaussian mixture model and the K-means algorithm (MacQueen 1967). Further, we show that treating the spanning forest as random, as opposed to a fixed parameter (i.e., unknown), leads to much less sensitivity in clustering performance, compared to cutting the minimum spanning tree algorithm (Gower and Ross 1969). On the distribution specification on the node and edges, we take a Bayesian nonparametric approach by considering the forest model as realized from a "forest process"—each cluster is initiated with a point from a root distribution, then gradually grown with new points from a leaf distribution. We characterize the key differences in the partition distribution between the forest and classic Pólya urn processes. This difference also reveals that extra care should be exerted during model specification when using graphical models for clustering. Lastly, by establishing the probabilistic model counterpart for spectral clustering, we show how such models can be easily extended to incorporate other dependency structures. We demonstrate several extensions, including a multi-subject clustering of the brain networks, and a high-dimensional clustering of photo images.

2. Method

2.1. Background on Spectral Clustering Algorithms

We first provide a brief review of spectral clustering algorithms. For data $y_1, \ldots, y_n$, let $A_{i,j} \geq 0$ be a similarity score between y_i and y_j, and denote the degree $D_{i,i} = \sum_{j \neq i} A_{i,j}$. To partition the data index $(1, \ldots, n)$ into K sets, $\mathcal{V} = (V_1, \ldots, V_K)$, we want to solve the following problem:

$$\min_{\mathcal{V}} \sum_{k=1}^{K} \frac{\sum_{i \in V_k, j \notin V_k} A_{i,j}}{\sum_{i \in V_k} D_{i,i}}. \tag{1}$$

This is known as the minimum normalized cut loss. The numerator above represents the across-cluster similarity due to cutting V_k off from the others; and the denominator prevents trivial solutions of forming tiny clusters with small $\sum_{i \in V_k} D_{i,i}$.

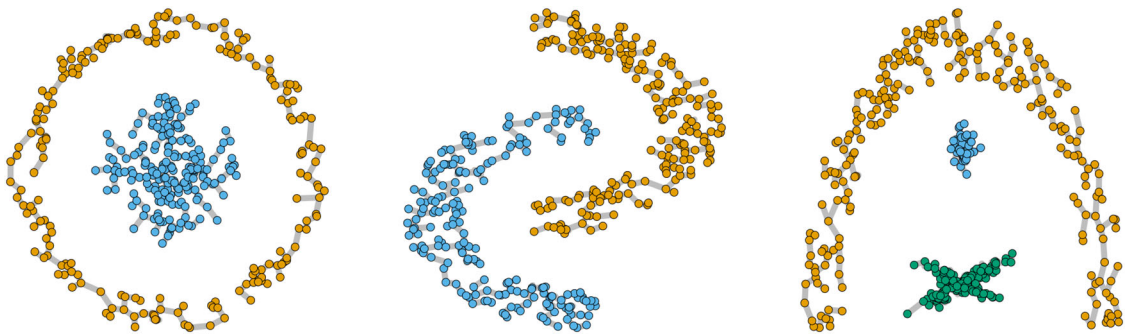


Figure 1. Three examples of clusters that can be represented by a spanning forest.

This optimization problem is a combinatorial problem, hence, has motivated approximate solutions such as spectral clustering. To start, using the Laplacian matrix $L = D - A$ with D the diagonal matrix of $D_{i,i}$'s, and the normalized Laplacian $N = D^{-1/2}LD^{-1/2}$, we can equivalently solve the above problem via:

$$\min_{\mathcal{V}} \operatorname{tr}(Z'_{\mathcal{V}} N Z_{\mathcal{V}}),$$

where $Z_{\mathcal{V}:i,k} = 1(i \in V_k)\sqrt{D_{i,i}}/\sqrt{\sum_{i\in V_k} D_{i,i}}$. It is not hard to verify that $Z'_{\mathcal{V}} Z_{\mathcal{V}} = I_K$. We can obtain a relaxed minimizer of $Z : Z'Z = I_K$, by simply taking $\hat{Z}$ as the bottom K eigenvectors of N (with the minimum loss equal to the sum of the smallest K eigenvalues). Afterward, we cluster the rows of $\hat{Z}$ into K groups (using algorithms such as the K-means), hence, producing an approximate solution to (1).

To clarify, there is more than one version of the spectral clustering algorithms. An alternative version to (1) is called "minimum ratio cut," which replaces the denominator $\sum_{i\in V_k} D_{i,i}$ by the size of cluster $|V_k|$. Similarly, continuous relaxation approximation can be obtained by following the same procedures above, except for clustering the eigenvectors of the unnormalized L. Details on comparing those two versions can be found in Von Luxburg (2007). In this article, we focus on the one based on (1) and the normalized Laplacian matrix N. This version is also commonly referred to as "normalized spectral clustering".

2.2. Probabilistic Model via Bayesian Spanning Forest

The next question is if there is some partition-based generative model for y, that has the maximum likelihood estimate (or, the posterior mode in the Bayesian framework) almost the same as the point estimate from the normalized spectral clustering.

We found an almost equivalence in the spanning forest model. A spanning forest model is a special Bayesian network that describes the conditional dependencies among $y_1, \ldots, y_n$. Given a partition $\mathcal{V} = (V_1, \ldots, V_K)$ of the data index $(1, \ldots, n)$, consider a forest graph $\mathcal{F}_{\mathcal{V}} = (T_1, \ldots, T_k)$, with each $T_k = (V_k, E_k)$ a component tree (a connected subgraph without cycles), V_k the set of nodes and E_k the set of edges among V_k. Using $\mathcal{F}_{\mathcal{V}}$ and a set of root nodes $\mathcal{R}_{\mathcal{V}} = (1^*, \ldots, K^*)$ with $k^* \in V_k$, we can form a graphical model with a conditional likelihood given the forest:

$$\mathcal{L}(y; \mathcal{V}, \mathcal{F}_{\mathcal{V}}, \mathcal{R}_{\mathcal{V}}, \theta) = \prod_{k=1}^{K} \left[r(y_{k^*}; \theta) \prod_{(i,j)\in T_k} f(y_i \mid y_j; \theta) \right], \quad (2)$$

where we refer to $r(\cdot; \theta)$ as a "root" distribution, and $f(\cdot \mid y_j; \theta)$ as a "leaf" distribution; and we use θ to denote the other parameter; and we use simplified notation $(i, j) \in G$ to mean that (i, j) is an edge of the graph G. Figure 1 illustrates the high flexibility of a spanning forest in representing clusters. It shows the sampled $\mathcal{F}$ based on three clustering benchmark datasets. Note that some clusters are not elliptical or convex in shape. Rather, each cluster can be imagined as if it were formed by connecting a point to another nearby. In the supplementary materials S4.8, we show two different realizations of spanning forest.

Remark 1. To clarify, the point estimation on a spanning forest (as some fixed and unknown graph) has been studied (Gower and Ross 1969). However, a distinction here is that we consider $\mathcal{V}$ as the parameter of interest, but the edges and roots $(\mathcal{F}_{\mathcal{V}}, \mathcal{R}_{\mathcal{V}})$ as latent variables. The performance differences are shown in the supplementary materials S4.6.

The stochastic view of $(\mathcal{F}_{\mathcal{V}}, \mathcal{R}_{\mathcal{V}})$ is important, as it allows us to incorporate the uncertainty of edges and avoids the sensitivity issue in the point graph estimate. Equivalently, our clustering model is based on the marginal likelihood that varies with the node partition $\mathcal{V}$:

$$\mathcal{L}(y; \mathcal{V}, \theta) = \sum_{\mathcal{F}_{\mathcal{V}}, \mathcal{R}_{\mathcal{V}}} \mathcal{L}(y; \mathcal{V}, \mathcal{F}_{\mathcal{V}}, \mathcal{R}_{\mathcal{V}}, \theta) \Pi(\mathcal{F}_{\mathcal{V}}, \mathcal{R}_{\mathcal{V}} \mid \mathcal{V}), \quad (3)$$

where $\Pi(\mathcal{F}_{\mathcal{V}}, \mathcal{R}_{\mathcal{V}} \mid \mathcal{V})$ is the latent variable distribution that we will specify in the next section. We can quantify the marginal connecting probability for each potential edge (i, j):

$$M_{i,j} := \Pr[F_{\mathcal{V}} \ni (i,j)] \propto \sum_{\mathcal{V}} \sum_{\mathcal{F}_{\mathcal{V}}, \mathcal{R}_{\mathcal{V}}} 1[(i,j) \in F_{\mathcal{V}}] \mathcal{L}(y; \mathcal{V}, \mathcal{F}_{\mathcal{V}}, \mathcal{R}_{\mathcal{V}}, \theta) \Pi(\mathcal{F}_{\mathcal{V}}, \mathcal{R}_{\mathcal{V}} \mid \mathcal{V}). \quad (4)$$

Similar to the normalized graph cut, there is no closed-form solution for directly maximizing (3). However, closed-form does exist for (4) (see Section 4). Therefore, an approximate maximizer of (3), $\hat{\mathcal{V}}$, can be obtained via computing the matrix M and searching for K diagonal blocks (after row and column index permutation) that contain the highest total values of $M_{i,j}$'s. Specifically, we can extract the top leading eigenvectors of M and cluster the rows into K groups.

This approximate marginal likelihood maximizer produces almost the same estimate as the normalized spectral clustering does. This is because the two sets of eigenvectors are almost the

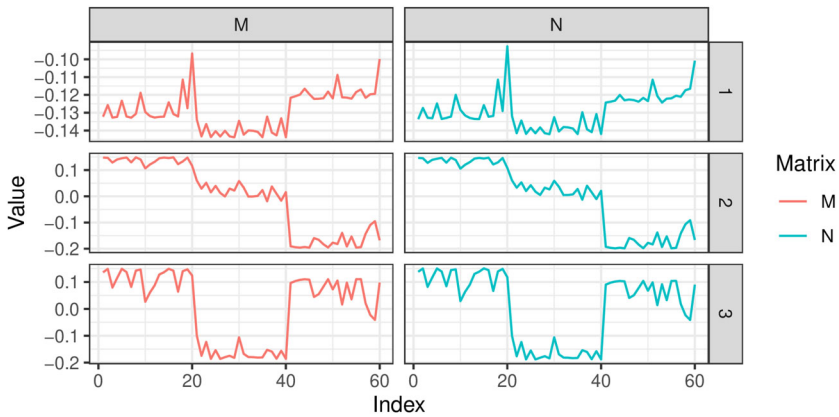


Figure 2. Comparing the eigenvectors of a marginal connecting probability matrix M and the ones of normalized Laplacian N.

same. Further, it is important to clarify that such closeness does not depend on how the data are really generated. Therefore, to provide some numerical evidence, for simplicity, we generate y_i from a simple three-component Gaussian mixture in $\mathbb{R}^2$ with means in $(0,0)$, $(2,2)$, $(4,4)$ and all variances equal to I_2. Figure 2 compares the eigenvectors of the matrix M and the normalized Laplacian N (that uses f and r to specify A, with details provided in Section 4). Clearly, these two are almost identical in values. Due to this connection, the clustering estimates from spectral clustering can be viewed as an approximate estimate for $\hat{\mathcal{V}}$ in (3).

We now fully specify the Bayesian forest model. For simplicity, we now focus on continuous $y_i \in \mathbb{R}^p$. For ease of computation, we recommend choosing f as a symmetric function $f(y_i \mid y_j;\theta) = f(y_j \mid y_i;\theta)$, so that the likelihood is invariant to the direction of each edge; and choose r as a diffuse density, so that the likelihood is less sensitive to the choice of a node as root. In this article, we choose a Gaussian density for f and Cauchy for r:

$$
\begin{aligned}
f(y_i \mid y_j;\theta) &= (2\pi\sigma_{i,j})^{-p/2} \exp\left\{-\frac{\|y_i - y_j\|_2^2}{2\sigma_{i,j}}\right\},\\
r(y_i;\theta) &= \frac{\Gamma[(1+p)/2]}{\gamma^p \pi^{(1+p)/2}} \frac{1}{(1+\|y_i-\mu\|_2^2/\gamma^2)^{(1+p)/2}},
\end{aligned}
\tag{5}
$$

where $\sigma_{ij} > 0$ and $\gamma > 0$ are scale parameters. As the magnitudes of distances between neighboring points may differ significantly from cluster to cluster, we use a local parameterization $\sigma_{i,j} = \tilde{\sigma}_i\tilde{\sigma}_j$, and will regularize $(\tilde{\sigma}_1,\ldots,\tilde{\sigma}_n)$ via a hyper-prior.

Remark 2. In (5), we effectively use Euclidean distances $\|y_i - y_j\|_2$. We focus on Euclidean distance in the main text, for the simplicity of presentation and to allow a complete specification of priors. One can replace Euclidean distance with some others, such as Mahalanobis distance and geodesic distance. We present a case of high-dimensional clustering based on geodesic distance on the unit-sphere in the supplementary materials S1.1.

2.3. Forest Process and Product Partition Prior

To simplify notations as well as to facilitate computation, we now introduce an auxiliary node 0 that connects to all roots $(1^*,\ldots,K^*)$. As the result, the model can be equivalently represented by a spanning tree rooted at 0:

$$
\begin{aligned}
&\mathcal{T} = (V_{\mathcal{T}}, E_{\mathcal{T}}),\\
&V_{\mathcal{T}} = \{0\} \cup V_1 \cup \cdots \cup V_K,\\
&E_{\mathcal{T}} = \{(0,1^*),\ldots,(0,K^*)\} \cup E_1 \cup \cdots \cup E_K.
\end{aligned}
$$

In this section, we focus on the distribution specification for $\mathcal{T}$. The distribution, denoted by $\Pi(\mathcal{T})$, $\Pi(\mathcal{T})$ can be factorized according to the following hierarchies: picking the number of partitions K, partitioning the nodes into $(V_1,\ldots,V_K)$, forming edges E_k and picking one root k^* for each V_k. To be clear on the nomenclature, we call $\Pi(\mathcal{F}_{\mathcal{V}},\mathcal{R}_{\mathcal{V}} \mid \mathcal{V})$ as the "latent variable distribution," $\Pi_0(\mathcal{V})$ as the "partition prior."

$$
\begin{aligned}
\Pi(\mathcal{T}) = &\underbrace{\left\{\Pi_0(K)\Pi_0(V_1,\ldots,V_K \mid K)\right\}}_{\Pi_0(\mathcal{V})}\\
&\underbrace{\prod_{k=1}^{K}\left\{\Pi(E_k \mid V_k)\Pi(k^* \mid E_k, V_k)\right\}}_{\Pi(\mathcal{F}_{\mathcal{V}},\mathcal{R}_{\mathcal{V}}\mid\mathcal{V})}.
\end{aligned}
\tag{6}
$$

Remark 3. In Bayesian nonparametric literature, $\Pi_0(K)$ $\Pi_0(V_1,\ldots,V_K \mid K)$ is known as the partition probability function, which plays the key role in controlling cluster sizes and cluster number in model-based clustering. However, when it comes to graphical model-based clustering (such as our forest model), it is important to note the difference—for each partition V_k, there is an additional probability $\Pi(E_k, k^* \mid V_k)$ due to the multiplicity of all possible subgraphs formed between the nodes in V_k.

For simplicity, we will use discrete uniform distribution for $\Pi(E_k, k^* \mid V_k)$. Since there are $n_k^{(n_k-2)_+}$ possible spanning trees for n_k nodes $[(x)_+ = x$ if $x > 0$, otherwise $0]$, and n_k possible choice of roots. We have $\Pi(E_k, k^* \mid V_k) = n_k^{-(n_k-1)}$.

We now discuss two different ways to complete the distribution specification. We first take a "ground-up" approach by viewing $\mathcal{T}$ as from a stochastic process where the node number n could grow indefinitely. Starting from the first edge $e_1 = (0,1)$,

we sequentially draw new edges and add to $\mathcal{T}$, from

$$\begin{aligned} e_i \mid e_1, \dots e_{i-1} &\sim \sum_{j=1}^{i-1} \pi_j^{[i]} \delta_{(j,i)}(\cdot) + \pi_i^{[i]} \delta_{(0,i)}(\cdot), \\ y_i \mid (j,i) &\sim 1(j \geq 1) f(\cdot \mid y_j) + 1(j=0) r(\cdot), \end{aligned} \tag{7}$$

with some probability vector $(\pi_1^{[i]}, \dots, \pi_i^{[i]})$ that adds up to one. We refer to (7) as a forest process. The forest process is a generalization of the Pólya urn process (Blackwell and MacQueen 1973). For the latter, $e_i = (j,i)$ would make node i take the same value as node j, $y_i = y_j$ [although in model-based clustering, one would use notation $\theta_i = \theta_j$, and $y_i \sim \mathcal{K}(\cdot \mid \theta_i)$]; $e_i = (0,i)$ would make node i draw a new value for y_i from the base distribution. Due to this relationship, we can borrow popular parameterization for $\pi_j^{[i]}$ from the urn process literature. For example, we can use the Chinese restaurant process parameterization $\pi_j^{[i]} = 1/(i-1+\alpha)$ for $j = 1, \dots, (i-1)$, and $\pi_i^{[i]} = \alpha/(i-1+\alpha)$ with some chosen $\alpha > 0$. After marginalizing over the order of i and partition index (see Miller 2019 for a simplified proof of the partition function), we obtain:

$$\Pi(\mathcal{T}) = \frac{\alpha^K \Gamma(\alpha)}{\Gamma(\alpha+n)} \prod_{k=1}^{K} \Gamma(n_k) n_k^{-(n_k-1)}. \tag{8}$$

Compared to the partition probability prior in the Chinese restaurant process, we have an additional $n_k^{-(n_k-1)}$ term that corresponds to the conditional prior weight of for each possible (k^*, E_k) given a partition V_k.

To help understand the effect of this additional term on the posterior, we can imagine two extreme possibilities in the conditional likelihood given a V_k. If the conditional $\mathcal{L}(y_i : i \in V_k \mid k^*, E_k)$ is skewed toward one particular choice of tree $(\hat{k}^*, \hat{E}_k)$ [that is, $\mathcal{L}(y_i : i \in V_k \mid k^*, E_k)$ is large when $(k^*, E_k) = (\hat{k}^*, \hat{E}_k)$, but is close to zero for other values of (k^*, E_k)], then $n_k^{-(n_k-1)}$ acts as a penalty for a lack of diversity in trees. On the other hand, if $\mathcal{L}(y_i : i \in V_k \mid k^*, E_k)$ is equal for all possible (k^*, E_k)'s, then we can simply marginalize over (k^*, E_k) and be not be subject to this penalty [since $\sum_{(k^*,E_k)} n_k^{-(n_k-1)} = 1$].

Therefore, we can form an intuition by interpolating those two extremes: if a set of data points (of size n_k) are "well-knit" such that they can be connected via many possible spanning trees (each with a high conditional likelihood), then it would have a higher posterior probability of being clustered together, compared to some other points (of the same size n_k) that have only a few trees with high conditional likelihood.

With the "ground-up" construction useful for understanding the difference from the classic urn process, the distribution (8) itself is not very convenient for posterior computation. Therefore, we also explore the alternative of a "top-down" approach. This is based on directly assigning a product partition probability (Hartigan 1990; Barry and Hartigan 1993; Crowley 1997; Quintana and Iglesias 2003) as

$$\Pi_0(V_1, \dots, V_K \mid K) = \frac{\prod_{k=1}^{K} n_k^{(n_k-1)}}{\sum_{\text{all } (V_1^*, \dots, V_K^*)} \prod_{k=1}^{K} |V_k^*|^{(|V_k^*|-1)}}, \tag{9}$$

where the cohesion function $n_k^{(n_k-1)}$ effectively cancels out the probability for each (k^*, E_k). To assign a prior for K, we assign a probability

$$\Pi_0(K) \propto \lambda^K \sum_{\text{all } (V_1^*, \dots, V_K^*)} \prod_{k=1}^{K} |V_k^*|^{(|V_k^*|-1)},$$

supported on $K \in \{1, \dots, n\}$ with $\lambda > 0$, with $\Pi(E_k, k^* \mid V_k) = n_k^{-(n_k-1)}$, multiplying the terms according to (6) leads to

$$\Pi(\mathcal{T}) \propto \lambda^K, \tag{10}$$

which is similar to a truncated geometric distribution and easy to handle in posterior computation, and we will use this from now on. In this article, we set $\lambda = 0.5$.

Remark 4. We now discuss the exchangeability of the sequence of random variables generated from the above forest process. The exchangeability is defined as the invariance of distribution $\Pi(X_1 = x_1, \dots X_n = x_n) = \Pi(X_1 = x_{\tilde{\pi}_1}, \dots X_n = x_{\tilde{\pi}_n})$ under any permutation $(\tilde{\pi}_1, \dots, \tilde{\pi}_n)$ (Diaconis 1977). For simplicity, we focus on the joint distribution with θ as given, and hence omit θ here. There are three categories of random variables associated with each node i: the first drawn edge (j,i) that points to a new node i (whose sequence forms $\mathcal{T} = (\mathcal{V}, \{E_k, k^*\}_{k=1}^K)$), the cluster assignment of a node c_i (whose sequence forms $\mathcal{V}$), and the data point y_i. It is not hard to see that, since each component tree encodes an order among $\{i : c_i = k\}$, the joint distribution of the data and the forest $\Pi(y_1, \dots, y_n, \mathcal{T})$ is not exchangeable. Nevertheless, as we marginalize out each (E_k, k^*) to form the clustering likelihood $\mathcal{L}(y; \mathcal{V})$ in (3), and all priors $\Pi_0(\mathcal{V})$ presented in this section only depend on the number and sizes of clusters, the joint distribution of the data and cluster labels $\Pi\{(y_1, c_1), \dots, (y_n, c_n)\} = \mathcal{L}(y; \mathcal{V}) \Pi_0(\mathcal{V})$ is exchangeable, with its form provided soon in (14). Lastly, we see that $\Pi(y_1, \dots, y_n)$ is exchangeable after marginalizing over $\mathcal{V}$.

2.4. Hyper-Priors for the Other Parameters

We now specify the hyper-priors for the parameters in the root and leaf densities. To avoid model sensitivities to scaling and shifting of the data, we assume that the data have been appropriately scaled and centered (e.g., via standardization), so that the marginally $\mathbb{E}y \approx 0$ and $\mathbb{E}\|y_{\cdot j} - \mathbb{E}y_{\cdot j}\|_2^2 \approx 1$ for $j = 1, \dots, p$. To make the root density $r(\cdot)$ close to a small constant in the support of the data, we set $\mu = 0$ and $\gamma^2 \sim$ Inverse-Gamma(2, 1).

For $\sigma_{i,j}$ in the leaf density $f(y_i \mid y_j; \sigma_{i,j})$, in order to likely pick an edge (i,j) with j as a close neighbors of i (that is, (i,j) with small $\|y_i - y_j\|_2$), we want most of $\sigma_{i,j} = \tilde{\sigma}_i \tilde{\sigma}_j$ to be small. We use the following hierarchical inverse-gamma prior that shrinks each $\tilde{\sigma}_i$, while using a common scale hyper-parameter β_σ to borrow strengths among $\tilde{\sigma}_i$'s,

$$\begin{aligned} \beta_\sigma &\sim \text{Exp}(\eta_\sigma), \qquad \eta_\sigma \sim \text{Inverse-Gamma}(a_\sigma, \xi_\sigma), \\ \tilde{\sigma}_i &\overset{\text{iid}}{\sim} \text{Inverse-Gamma}(b_\sigma, \beta_\sigma) \text{ for } i = 1, \dots, n, \end{aligned}$$

where η_σ is the scale parameter for the exponential. To induce a shrinkage effect a priori, we use $a_\sigma = 100$ and $\xi_\sigma = 1$ for a likely

small η_σ hence a small β_σ. Further, we note that the coefficient of variation $\sqrt{\mathrm{var}(\tilde{\sigma}_i \mid \beta_\sigma)}/\mathbb{E}(\tilde{\sigma}_i \mid \beta_\sigma) = 1/\sqrt{b_\sigma - 2}$; therefore, we set $b_\sigma = 10$ to have most of $\tilde{\sigma}_i$ near $\mathbb{E}(\tilde{\sigma}_i \mid \beta_\sigma) = \beta_\sigma/(b_\sigma - 1)$ in the prior. We use these hyper-prior settings in all the examples presented in this article.

In addition, Zelnik-Manor and Perona (2005) demonstrate good empirical performance in spectral clustering via setting $\tilde{\sigma}_i$ to a low order statistic of the distances to y_i. We show a model-based formalization with similar effects in the supplementary materials S5.

2.5. Model-based Extensions

Compared to algorithms, a major advantage of probabilistic models is the ease of building useful model-based extensions. We demonstrate three directions for extending the Bayesian forest model. Due to the page constraint, we defer the details and numeric results of these extensions in the supplementary materials S1.1, S1.2, and S1.3.

Latent Forest Model: First, one could use the realization of the forest process as latent variables in another model $\mathcal{M}$ for data $(y_1, \ldots, y_n)$,

$$z_1, \ldots, z_n \sim \text{Forest Model}(\mathcal{T}; \theta_z),$$

$$y_1, \ldots, y_n \sim \mathcal{M}(z_1, \ldots, z_n; \theta_y),$$

where θ_z and θ_y denote the other needed parameters. For example, for clustering high-dimensional data such as images, it is often necessary to represent each high-dimensional observation y_i by a low-dimensional coordinate z_i (Wu, Feng, and Zhou 2014; Chandra, Canale, and Dunson 2023). In the supplementary materials, we present a high-dimensional clustering model, using an autoregressive matrix Gaussian for $\mathcal{M}$ and a sparse von Mises-Fisher for the forest model.

Informative Prior–Latent Variable Distribution: Second, in applications it is sometimes desirable to have the clustering dependent on some external information x, such as covariates (Müller, Quintana, and Rosner 2011) or an existing partition (Paganin et al. 2021). From a Bayesian view, this can be achieved via taking an x-informative distribution:

$$\mathcal{T} \sim \Pi(\cdot \mid x), \qquad y_1, \ldots, y_n \sim \text{Forest Model}(\mathcal{T}; \theta).$$

In the supplementary materials, we illustrate an extension with a covariate-dependent product partition model [PPMx, Müller, Quintana, and Rosner (2011)] into the distribution of $\mathcal{T}$.

Hierarchical Multi-view Clustering: Third, for multi-subject data $(y_1^{(s)}, \ldots, y_n^{(s)})$ for $s = 1, \ldots, S$, we want to find a clustering for every s. At the same time, we can borrow strength among subjects, by letting subjects share some similar partition structure on a subset of nodes (while differing on the other nodes). This is known as multi-view clustering. On the other hand, a challenge is that a forest is a discrete object subject to combinatorial constraints, hence, it would be difficult to partition the nodes freely while accommodating the tree structure. To circumvent this issue, we propose a latent coordinate-based distribution that gives a continuous representation for $\mathcal{T}^{(s)}$. Consider a latent $z_i^{(s)} \in \mathbb{R}^d$ for each node $i = 1, \ldots, n$, we assign a joint prior–latent variable distribution for $z^{(s)}$ and $\mathcal{T}^{(s)}$:

$$\begin{aligned}
&\Pi[z^{(s)}, \mathcal{T}^{(s)}] \propto \\
&\lambda^{K[\mathcal{T}^{(s)}]}\Big[\prod_{(i,j)\in\mathcal{T}^{(s)}: i\geq 1, j\geq 1} \exp\Big(-\frac{\|z_i^{(s)} - z_j^{(s)}\|_2^2}{2\rho}\Big)\Big] \\
&\Big[\prod_{i=1}^{n}\Big\{\sum_{k=1}^{\tilde{\kappa}} v_{i,k} \exp\Big(-\frac{\|z_i^{(s)} - \eta_k^*\|_2^2}{2\sigma_z^2}\Big)\Big\}\Big], \\
&(v_{i,1}, \ldots, v_{i,\tilde{\kappa}}) \sim \mathrm{Dir}(1/\tilde{\kappa}, \ldots, 1/\tilde{\kappa}) \qquad \text{for } i = 1, \ldots n, \\
&\{y_1^{(s)}, \ldots, y_n^{(s)}\} \sim \text{Forest Model}(\mathcal{T}^{(s)}) \qquad \text{for } s = 1, \ldots S,
\end{aligned} \tag{11}$$

where $v_{i,1}, \ldots, v_{i,\tilde{\kappa}}$ are the weights that vary with i and $\sum_{k=1}^{\tilde{\kappa}} v_{i,k} = 1$, $\rho > 0$, and $z^{(s)} \in \mathbb{R}^{n\times d}$ is the matrix form. Equivalently, the above assigns each node a location parameter $\eta_i^{(s)}$, drawn from a hierarchical Dirichlet distribution with shared atoms $\{\eta_1^*, \ldots, \eta_{\tilde{\kappa}}^*\}$ and probability $(v_{.,1}, \ldots, v_{.,\tilde{\kappa}})$ (Teh et al. 2006). Further, one could let η_k^* vary over node according to some functional using a hybrid Dirichlet distribution (Petrone, Guindani, and Gelfand 2009).

Using a Gaussian mixture kernel on $z_i^{(s)}$, we can now separate $z_i^{(s)}$'s into several groups that are far apart. To make the parameters identifiable and have large separations between groups, we fix $\tilde{\eta}_k^*$'s on the d-dimensional integer lattice $\{0, 1, 2\}^d$ with $d = 2$ (hence, $\tilde{\kappa} = 9$); and we use $\sigma_z^2 = 0.01$ and $\rho = 0.001$ in this article.

Remark 5. To clarify, our goal is to induce between-subject similarity in the node partition, not the tree structure. For example, for two subjects s and s', when $z_i^{(s)}$ and $z_i^{(s')}$ are both near η_k^* for all $i \in C$, then both the spanning forest $\mathcal{T}^{(s)}$ and $\mathcal{T}^{(s')}$ will likely cluster the nodes in C together, even though $T_k^{(s)}$ and $T_k^{(s')}$ associated with $V_k \supset C$ may be different.

3. Posterior Computation

3.1. Gibbs Sampling Algorithm

We now describe the Markov chain Monte Carlo (MCMC) algorithm. For ease of notation, we use an $(n+1)\times(n+1)$ matrix S, with $S_{i,j} = \log f(y_i \mid y_j; \theta)$, $S_{0,i} = S_{i,0} = \log r(y_i; \theta) + \log\lambda$ (for convenience, we use 0 to index the last row/column), $S_{i,i} = 0$, and $A_{\mathcal{T}}$ to represent the adjacency matrix of $\mathcal{T}$. We have the posterior distribution

$$\Pi(\mathcal{T}, \theta \mid y) \propto \exp\big\{\mathrm{tr}[S(\theta)A_{\mathcal{T}}]/2\big\}\Pi_0(\theta). \tag{12}$$

Note the above form conveniently include the prior term for the number of clusters, λ^K, via the number of edges adjacent to node 0.

Our MCMC algorithm alternates in updating $\mathcal{T}$ and θ, hence is a Gibbs sampling algorithm. To sample $\mathcal{T}$ given θ, we take the random-walk covering algorithm for weighted spanning tree (Mosbah and Saheb 1999), as an extension of the Andrei–Broder algorithm for sampling uniform spanning tree (Broder 1989; Aldous 1990). For this article to be self-contained, we describe the algorithm below. The above algorithm produces a

random sample $\mathcal{T}$ following the full conditional $\Pi(\mathcal{T} \mid \theta, y)$ proportional to (12). It has an expected finish time of $O(n \log n)$. Although some faster algorithms have been developed (Schild 2018), we choose to present the random-walk covering algorithm for its simplicity.

Algorithm 1 Random-walk covering algorithm for sampling the augmented tree $\mathcal{T}$.

Start with $V_{\mathcal{T}} = \{0\}$ and $E_{\mathcal{T}} = \varnothing$, and set $i \leftarrow 0$:
while $|V_{\mathcal{T}}| \neq n+1$ **do**
  Take a random walk from i to j with probability $\Pr(j \mid i) = \frac{\exp[S_{i,j}(\theta)]}{\sum_{j:j\neq i}\exp[S_{i,j}(\theta)]}$.
  if $j \notin V_{\mathcal{T}}$ **then**
    Add j to $V_{\mathcal{T}}$. Add (i,j) to $E_{\mathcal{T}}$.
  Update $i \leftarrow j$.

We sample $\tilde{\sigma}_i$ using the following steps,

$$(\eta_\sigma \mid .) \sim \text{Inverse-Gamma}\big(1 + a_\sigma, \beta_\sigma + \xi_\sigma\big)$$

$$(\beta_\sigma \mid .) \sim \text{Gamma}\left\{1 + nb_\sigma, (\sum_{i=1}^{n} \frac{1}{\tilde{\sigma}_i} + \frac{1}{\eta_\sigma})^{-1}\right\}$$

$$(\tilde{\sigma}_i \mid .) \sim \text{Inverse-Gamma}\left[\frac{p\sum_j 1\{(i,j)\in\mathcal{T}\}}{2} + b_\sigma, \sum_{j:(i,j)\in\mathcal{T}} \frac{\|y_i - y_j\|_2^2}{2\tilde{\sigma}_j} + \beta_\sigma\right]$$

To update γ, we use the form of the multivariate Cauchy as a scale mixture of $\mathrm{N}(\mu, \gamma^2 u_{\gamma,i} I_p)$ over $u_{\gamma,i} \sim$ Inverse-Gamma$(1/2, 1/2)$. We can update via

$$u_{\gamma,i} \sim \text{Inverse-Gamma}(\frac{1+p}{2}, \frac{1}{2} + \frac{\|y_i - \mu\|_2^2}{2\gamma^2}),$$

$$\gamma^2 \sim \text{Inverse-Gamma}(2 + \frac{Kp}{2}, \hat{\sigma}_y^2 + \sum_{i:(0,i)\in\mathcal{T}} \frac{\|y_i - \mu\|_2^2}{2u_{\gamma,i}}).$$

We run the MCMC algorithm iteratively for many iterations. And we discard the first half of iterations as burn-in.

Remark 6. We want to emphasize that the Andrei–Broder random-walk covering algorithm (Broder 1989; Aldous 1990; Mosbah and Saheb 1999) is an exact algorithm for sampling a spanning tree $\mathcal{T}$. That is, if θ were fixed, each run of this algorithm would produce an *independent* Monte Carlo sample $\mathcal{T} \sim \Pi(\mathcal{T} \mid \theta, y)$. Removing the auxiliary node 0 from $\mathcal{T}$ will produce K disjoint spanning trees. This augmented graph technique is inspired by Boykov, Veksler, and Zabih (2001).

In our algorithm, since the scale parameters in θ are unknown, we use Markov chain Monte Carlo that updates two sets of parameters, (i) $(\theta_{[t+1]} \mid \mathcal{T}_{[t]})$ and (ii) $(\mathcal{T}_{[t+1]} \mid \theta_{[t+1]})$ from iteration $[t]$ to $[t+1]$. Therefore, rigorously speaking, there is a Markov chain dependency between $\mathcal{T}_{[t]}$ and $\mathcal{T}_{[t+1]}$ induced by $\theta_{[t+1]}$. Nevertheless, since we draw $\mathcal{T}$ in a block via the random-walk covering algorithm, we empirically find that $\mathcal{T}_{[t+1]}$ and $\mathcal{T}_{[t]}$ are substantially different. In the supplementary materials S4.4, we quantify the iteration-to-iteration graph changes, and provide diagnostics with multiple start points of $(\mathcal{T}_{[0]}, \theta_{[0]})$.

3.2. Posterior Point Estimate on Clustering

In the field of Bayesian clustering, for producing point estimate on the partition, it had been a long-time practice to simply track $\mathrm{pr}(c_i = k \mid y)$, then take the element-wise posterior mode over k as the point estimate for $\hat{c}_i$. Nevertheless, this was shown to be sub-optimal due to that: (i) label switching issue causes unreliable estimates on $\mathrm{pr}(c_i = k \mid y)$; (ii) the element-wise mode can be unrepresentative of the center of distribution for $(c_1, \ldots, c_n)$ (Wade and Ghahramani 2018). These weaknesses have motivated new methods of obtaining point estimate of clustering, that transform an $n \times n$ pairwise co-assignment matrix $\{\mathrm{pr}(c_i = c_j \mid y)\}_{\text{all }(i,j)}$ into an $n \times K$ assignment matrix (Medvedovic and Sivaganesan 2002; Rasmussen et al. 2008; Molitor et al. 2010; Wade and Ghahramani 2018). More broadly speaking, minimizing a loss function based on the posterior sample (via some estimator or algorithm) is common for producing a point estimate under some decision theory criterion. For example, the posterior mean comes as the minimizer of the squared error loss; in Bayesian factor modeling, an orthogonal Procrustes-based loss function is used for producing the posterior summary of the loading matrix from the generated MCMC samples (Aßmann, Boysen-Hogrefe, and Pape 2016).

We follow this strategy. There have been many algorithms that one could use. For a recent survey, see Dahl, Johnson, and Müller (2022). In this article, we use a simple solution of first finding the mode of K from the posterior sample, then doing a $\hat{K}$-rank symmetric matrix factorization on $\{\mathrm{pr}(c_i = c_j \mid y)\}_{\text{all}(i,j)}$ and clustering into $\hat{K}$ groups, provided by `RcppML` package (DeBruine, Melcher, and Triche Jr 2021).

4. Theoretical Properties

4.1. Convergence of Eigenvectors

We now formalize the closeness of the eigenvectors of matrices N and M (shown in Section 2.2), by establishing the convergence of the two sets of eigenvectors as n increases.

To be specific, we focus on the normalized spectral clustering algorithm using the similarity $A_{i,j} = \exp(S_{i,j})$, with $S_{i,j} = \log f(y_i \mid y_j; \theta)$, $S_{0,i} = S_{i,0} = \log r(y_i; \theta) + \log \lambda$. On the other hand, for the specific form, $f(y_i \mid y_j)$ can be any density satisfying $f(y_i \mid y_j, \theta) = f(y_j \mid y_i, \theta)$, $r(y_i; \theta)$ can be any density satisfying $r(y_i; \theta) > 0$. For the associated normalized Laplacian N, we denote the first K bottom eigenvectors by $\phi_1, \ldots, \phi_K$, which correspond to the smallest K eigenvalues.

Let M be the matrix with $M_{i,j} = \mathrm{pr}[\mathcal{T} \ni (i,j) \mid y, \theta]$ for $i \neq j$ and $M_{i,i} = 0$. The Kirchhoff's tree theorem (Chaiken and Kleitman 1978) gives an enumeration of all $\mathcal{T} \in \mathbb{T}$,

$$\sum_{\mathcal{T}\in\mathbb{T}} \prod_{(i,j)\in\mathcal{T}} \exp(S_{i,j}) = (n+1)^{-1} \prod_{h=2}^{n+1} \lambda_{(h)}(L) \tag{13}$$

where L is the Laplacian matrix transform of the similarity matrix A; $\lambda_{(h)}$ denotes the hth smallest eigenvalue. Differentiating its logarithmic transform with respect to $S_{i,j}$,

$$M_{i,j} = \Pr[\mathcal{T} \ni (i,j) \mid y]$$

$$= \frac{\sum_{\mathcal{T}\in\mathbb{T},(i,j)\in\mathcal{T}} \prod_{(i',j')\in\mathcal{T}} \exp(S_{i',j'})}{\sum_{\mathcal{T}\in\mathbb{T}} \prod_{(i',j')\in\mathcal{T}} \exp(S_{i',j'})}$$

$$= \frac{\partial \sum_{i=2}^{n+1} \log \lambda_{(i)}(L)}{\partial S_{i,j}}.$$

Let $\Psi_1, \ldots, \Psi_K$ be the top K eigenvectors of M, associated with eigenvalues $\xi_1 \geq \xi_2 \geq \cdots \geq \xi_K$, and $\xi_K > \xi_{K+1} \geq \xi_{K+2} \geq \cdots \geq \xi_{n+1}$. And we can compare with the K leading eigenvectors of $(-N) \in \mathbb{R}^{n\times n}$, $\phi_1, \ldots, \phi_K$. Using $\Psi_{1:K}$ and $\phi_{1:K}$ to denote two $(n+1)\times K$ matrices, we now show they are close to each other.

Theorem 1. There exists an orthonormal matrix $R \in \mathbb{R}^{K\times K}$ and a finite constant $\epsilon > 0$,

$$\|\Psi_{1:K} - \phi_{1:K}R\|_F \leq \frac{40\sqrt{K(n+1)}}{\xi_K - \xi_{K+1}}$$

$$\max_{i,j} \left\{ (1+\epsilon)(D_i^{-1/2} - D_j^{-1/2})^2 A_{i,j} \right\},$$

with probability at least $1 - \exp(-n)$.

Remark 7. To make the right-hand side go to zero, a sufficient condition is to have all $A_{i,j}/D_{i,i} = O(n^{-\kappa})$ with $\kappa > 1/2$. We provide a detailed definition of the bound constant ϵ in the supplementary materials S2.

To explain the intuition behind this theorem, our starting point is the close relationship between Laplacian and spanning tree models—multiplying both sides of (13) by $(n+1)^{-(n-1)}$ shows that the nonzero eigenvalue product of the graph Laplacian L is proportional to the marginal probability of n data points from a spanning forest-mixture model. Starting from this equality, we can write the marginal inclusion probability matrix of $\mathcal{T}$ as a mildly perturbed form of the normalized Laplacian matrix. Intuitively, when two matrices are close, their eigenvectors will be close as well (Yu, Wang, and Samworth 2015).

Therefore, under mild conditions, as $n \to \infty$, the two sets of leading eigenvectors converge. In the supplementary materials S4.7, we show that the convergence is very fast, with the two sets of leading eigenvectors becoming almost indistinguishable starting around $n \geq 50$.

Besides the eigenvector convergence, we can examine the marginal posterior $\Pi(\mathcal{V} \mid \theta, y)$, which is proportional to

$$\mathcal{L}(y;\mathcal{V},\theta)\Pi_0(\mathcal{V}) = \Pi_0(K, V_1, \ldots, V_K)\left\{\prod_{k=1}^{K}[\sum_{i\in V_k} r(y_i)]\right\} \prod_{k=1}^{K}\left\{n_k^{-1}\prod_{h=2}^{n_k}\lambda_{(h)}(L_k)\right\}, \tag{14}$$

where L_k is the unnormalized Laplacian matrix associated with matrix $\{A_{i,j}\}_{i\in V_k, j\in V_k}$. Imagine that if we put all indices in one partition $V_1 = (1, \ldots, n)$, then $\Pi(\mathcal{V} \mid \theta, y)$ would be very small due to those close-to-zero eigenvalues. Applying this deduction recursively on subsets of data, it is not hard to see that a high-valued $\Pi(\mathcal{V} \mid \theta, y)$ would correspond to a partition, wherein each V_k has $\lambda_{(h)}(L_k)$ away from 0 for any $h \geq 2$. Further, since $\{n_k^{-1}\prod_{h=2}^{n_k}\lambda_{(h)}(L_k)\} = |L_k + J/n_k^2|$, a permutation in $(1, \ldots, n)$ corresponds to congruent and simultaneous permutations of rows and columns of each L_k, which does not change each determinant. Therefore, the joint distribution of $\Pi\{(y_1, c_1), \ldots, (y_n, c_n)\}$ is exchangeable.

4.2. Consistent Clustering of Separable Sets

We show that clustering consistency is possible, under some separability assumptions when the data-generating distribution follows a forest process. Specifically, we establish posterior ratio consistency, as the ratio between the maximum posterior probability assigned to other possible clustering assignments to the posterior probability assigned to the true clustering assignments converges to zero almost surely under the true model (Cao, Khare, and Ghosh 2019).

To formalize the above, we denote the true cluster label for generating y_i by c_i^0 (subject to label permutation among clusters), and we define the enclosing region for all possible $y_i : c_i^0 = k$ as R_k^0 for $k = 1, \ldots, K_0$ for some true finite K_0. And we refer to $R^0 = (R_1^0, \ldots, R_{K_0}^0)$ as the "null partition". By separability, we mean the scenario that $(R_1^0, \ldots, R_{K_0}^0)$ are disjoint and there is a lower-bounded distance between each pair of sets. As alternatives, regions $R = (R_1, \ldots, R_K)$ could be induced by $\{c_1, \ldots, c_n\}$ from the posterior estimate of $\mathcal{T}$. For simplicity, we assume the scale parameter in f is known and all equal $\sigma_{i,j} = \sigma^{0,n}$.

Number of clusters is known. We first start with a simple case when we have fixed $K = K_0$. For regularities, we consider data as supported in a compact region $\mathcal{X}$, and satisfying the following assumptions:

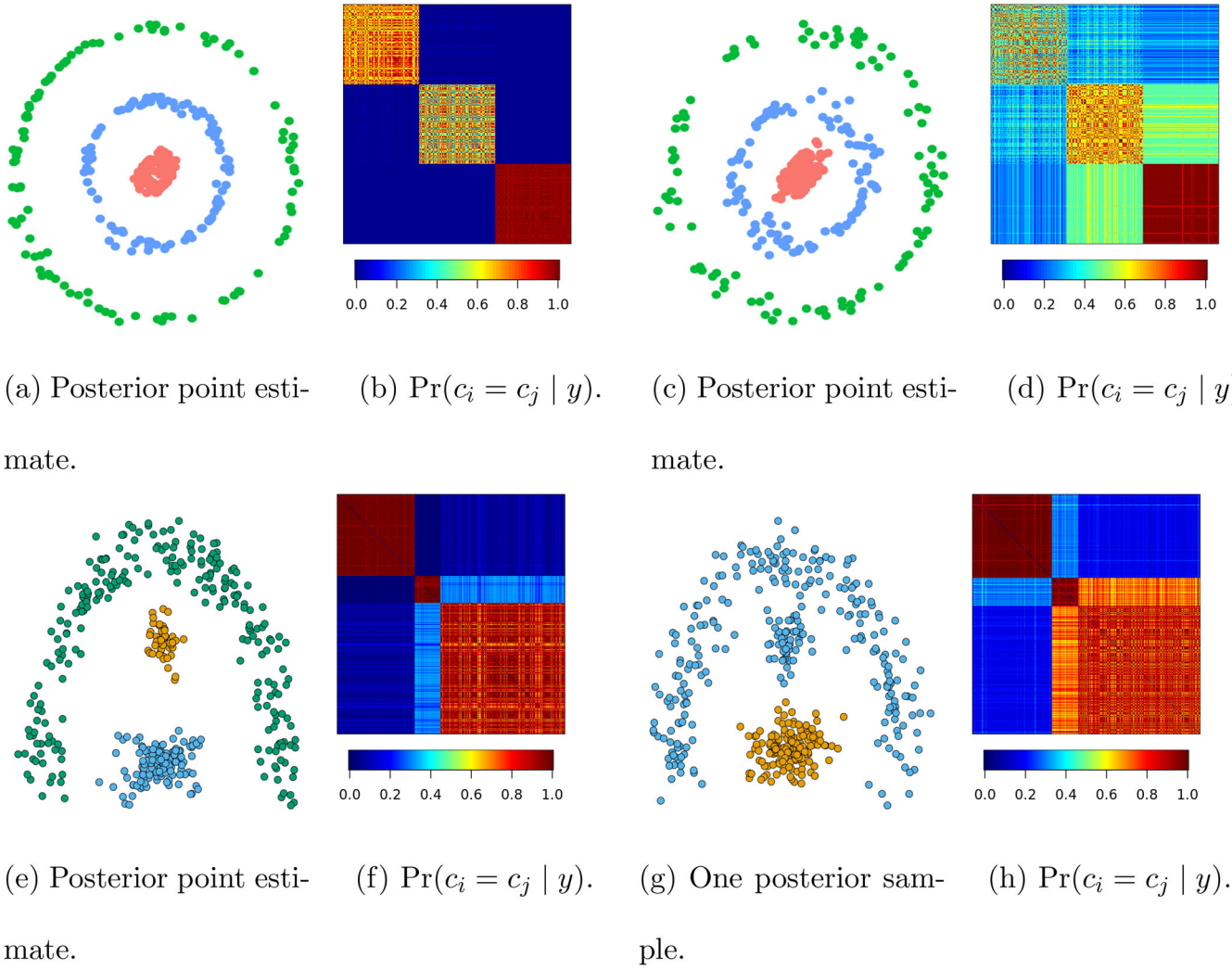
- (A1, diminishing scale) $\sigma^{0,n} = C'(1/\log n)^{1+\iota}$ for some $\iota > 0$ and $C' > 0$.
- (A2, minimum separation) $\inf_{x\in R_k^0, y\in R_{k'}^0} \|x - y\|_2 > M_n$, for all $k \neq k'$ with some positive constant $M_n > 0$ such that $M_n^2/\sigma^{0,n} = 8\tilde{m}_0 \log(n)$ for all (i,j) and is known for some constant $\tilde{m}_0 > p/2 + 2$.
- (A3, near-flatness of root density) For any n, $\epsilon_1 < r(y) < \epsilon_2$ for all $y \in \mathcal{X}$.

Under the null partition, $\Pi(\mathcal{T}|y)$ is maximized at $\mathcal{T} = \mathcal{T}_{\mathrm{MST},R^0}$, which contains K_0 trees with each T_k being the minimum spanning tree (denoted by subscript "MST") within region R_k^0. Similarly, for any alternative R, $\Pi(\mathcal{T}|y)$ is maximized at the $\mathcal{T} = \mathcal{T}_{\mathrm{MST},R}$.

Theorem 2. Under (A1,A2,A3), we have $\Pi(\mathcal{T}_{\mathrm{MST},R}|y)/\Pi(\mathcal{T}_{\mathrm{MST},R^0}|y) \to 0$ almost surely, unless $R_i^0 \subseteq R_{\xi(i)}$ for some permutation map $\xi(\cdot)$.

Number of clusters is unknown: Next, we relax the condition by having a K not necessarily equal to K_0. We show the consistency in two parts for 1)$K < K_0$, and 2) $K > K_0$ separately. In order to show posterior ratio consistency in the second part, we need some finer control on $r(y)$:

- (A3') The root density satisfies $\tilde{m}_1 e^{-M/2\sigma^{0,n}} \leq r(y) \leq \tilde{m}_2 e^{-M/2\sigma^{0,n}}$ for some $\tilde{m}_1 < \tilde{m}_2$.



(a) Posterior point estimate. (b) $\Pr(c_i = c_j \mid y)$. (c) Posterior point estimate. (d) $\Pr(c_i = c_j \mid y)$.

(e) Posterior point estimate. (f) $\Pr(c_i = c_j \mid y)$. (g) One posterior sample. (h) $\Pr(c_i = c_j \mid y)$.

Figure 3. Uncertainty quantification in clustering data generated near three manifolds. When data are close to the manifolds (Panels (a) and (e)), there is very little uncertainty on clustering in low $\Pr(c_i = c_j \mid j)$ between points from different clusters (Panels (b) and (f)). As data deviate more from the manifolds (Panel (c) and (g)), the uncertainty increases (Panels (d) and (h)). And in Panel (g), the point estimate shows a two-cluster partitioning, while there is about 20% of probability for three-cluster partitioning.

In this assumption, we essentially assume the root distribution to be flatter with a larger n. Then we have the following results.

Theorem 3. 1) If $K < K_0$, under the assumptions (A1,A2,A3), we have
$\Pi(\mathcal{T}_{\mathrm{MST},R}|y)/\Pi(\mathcal{T}_{\mathrm{MST},R^0}|y) \to 0$ almost surely.

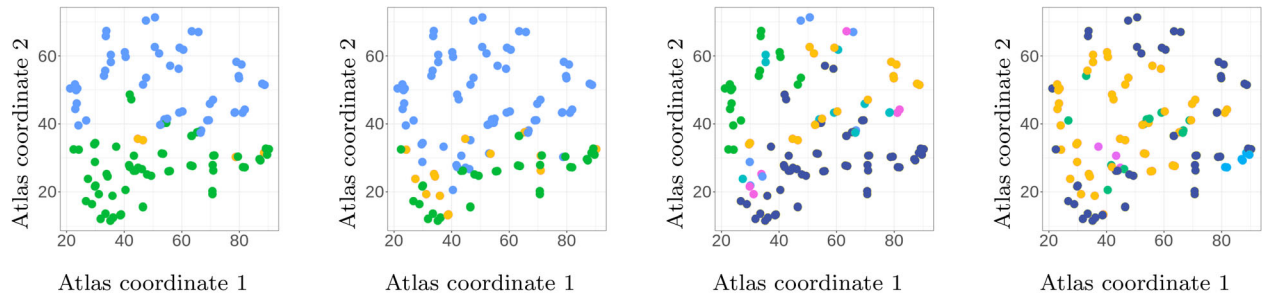
2) If $K > K_0$, under the assumptions (A1,A2,A3'), we have $\Pi(\mathcal{T}_{\mathrm{MST},R}|y)/\Pi(\mathcal{T}_{\mathrm{MST},R^0}|y) \to 0$ almost surely.

The above results show posterior ratio consistency. Furthermore, when the true of clusters is known, the ratio consistency result can be further extended to show clustering consistency, which is proved in the supplementary materials S3.

5. Numerical Experiments

To illustrate the capability of uncertainty quantification, we carry out clustering tasks on those near-manifold data commonly used for benchmarking clustering algorithms. In the first simulation, we start with 300 points drawn from three rings of radii 0.2, 1 and 2, with 100 points from each ring. Then we add some Gaussian noise to each point to create a coordinate near a ring manifold. We present two experiments, one with noises from $\mathrm{N}(0, 0.05^2 I_2)$, and one with noises $\mathrm{N}(0, 0.1^2 I_2)$. As shown in Figure 3, when these data are well separated (Panel (a), showing posterior point estimate), there is very little uncertainty on the clustering (Panel (b)), with the posterior co-assignment $\Pr(c_i = c_j \mid y)$ close to zero for any two data points near different rings. As noises increase, these data become more difficult to separate. There is a considerable amount of uncertainty for those red and blue points: these two sets of points are assigned into one cluster with a probability close to 40% (Panel (d)). We conduct another simulation based on an arc manifold and two point clouds (Panels (e)–(h)), and find similar results. Additional experiments are described in the supplementary materials S4.2.

In the supplementary materials S4.1 and S4.3, we present some uncertainty quantification results, for clustering data that are from mixture models. We compare the estimates with the ones from Gaussian mixture models, which could correspond to correctly/erroneously specified component distribution. Empirically, we find that the uncertainty estimates on $\Pr(c_i = c_j \mid y)$



(a) Clustering of the nodes for one healthy subject. (b) Clustering of the nodes for another healthy subject. (c) Clustering of the nodes for one diseased subject. (d) Clustering of the nodes for another diseased subject.

Figure 4. Results of brain region clustering (lateral view) for four subjects taken from the healthy and diseased groups. The multi-view clustering model allows subjects to have similar partition structures on a subset of nodes, while having subtle differences on the others (Panels (a) and (b), Panels (c) and (d)). At the same time, the healthy subjects show less degree of variability in the brain clustering than the diseased subjects.

and $\Pr(K \mid y)$ from the forest model are close to the ones based on the true data-generating distribution; whereas the Gaussian mixture models suffer from sensitivity in model specification, especially when K is not known.

6. Application: Clustering in Multi-subject Functional Magnetic Resonance Imaging Data

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to predict the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). In this application, we conduct a neuroscience study for finding connected brain regions under a varying degree of impact from Alzheimer's disease. The source dataset is resting-state functional magnetic resonance imaging (rs-fMRI) scan data, collected from $S = 166$ subjects at different stages of Alzheimer's disease. Each subject has scans over $n = 116$ regions of interest using the Automated Anatomical Labeling (AAL) atlas (Rolls et al. 2020; Shi et al. 2021) and over $p = 120$ time points. We denote the observation for the sth subject in the ith region by $y_i^{(s)} \in \mathbb{R}^p$.

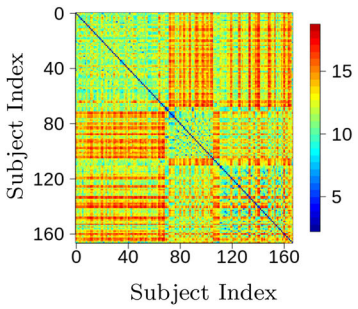
The rs-fMRI data are known for their high variability, often characterized by a low intraclass correlation coefficient (ICC), $(1 - \hat{\sigma}^2_{\text{within-group}}/\hat{\sigma}^2_{\text{total}})$, as the estimate for the proportion of total variance that can be attributed to variability between groups (Noble, Scheinost, and Constable 2021). Therefore, our goal is to use the multi-view clustering to divide the regions of interest for each subject, while improving our understanding of the source of high variability.

We fit the multi-view clustering model to the data, by running MCMC for 5000 iterations and discarding the first 2500 as burn-in. As shown in Figure 4, the hierarchical Dirichlet distribution on the latent coordinates induces similarity between the clustering of brain regions among subjects on a subset of nodes, while showing subtle differences on the other nodes. On the other hand, some major differences can be seen in the clusterings between the healthy and diseased subjects. Using the latent coordinates (at the posterior mean), we quantify the distances between $z^{(s)}$ and $z^{(s')}$ for each pair of subjects $s \neq s'$. As shown in Figure 5(a), there is a clear two-group structure in the pairwise distance matrix formed by $\|z^{(s)} - z^{(s')}\|_F$, and the separation corresponds to the first 64 subjects being healthy (denoted by $s \in g_1$) and the latter 102 being diseased (denoted by $s \in g_2$).

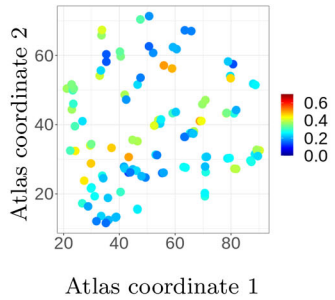
Next, we compute the within-group variances for these two groups, using $\sum_{s \in g_l} \|z_i^{(s)} - (\sum_{s \in g_l} z_i^{(s)}/|g_l|)\|_F^2/|g_l|$ for $l = 1$ and 2, and plot the variance over each region of interest i on the spatial coordinate of the atlas. Figure 5(b) and (c) show that, although both groups show some degree of variability, the diseased group shows clearly higher variances in some regions of the brain. Specifically, the paracentral lobule (PCL) and superior parietal gyrus (SPG), dorsolateral superior frontal gyrus (SFGdor), and supplementary motor area (SMA) in the frontal lobe show the highest amount of variability. Indeed, those regions are also associated with very low ICC scores [Figure 5(e)] calculated based on the variance of $z_i^{(s)}$, with pooled estimates $\hat{\sigma}^2_{\text{total},i} = \sum_s \|z_i^{(s)} - (\sum_s z_i^{(s)}/S)\|_F^2/S$ and $\hat{\sigma}^2_{\text{within-group},i} = \sum_{l=1}^2 \sum_{s \in g_l} \|z_i^{(s)} - (\sum_{s \in g_l} z_i^{(s)}/|g_l|)\|_F^2/S$. On the other hand, some regions such as the hippocampus (HIP), parahippocampal gyrus (PHG), and superior occipital gyrus (SOG) show relatively lower variances within each group, hence, higher ICC scores.

To show more details on the heterogeneity, we plot the latent coordinates associated with those ROIs using boxplots. Since each $z_i^{(s)}$ is in two-dimensional space, we plot the linear transform $\tilde{z}_i^{(s)} = z_{i,1}^{(s)} + z_{i,2}^{(s)}$. Interestingly, those 8 ROIs with high variability still seem quite informative for distinguishing the two groups (Figure 5(f)). To verify, we concatenate those latent coordinates and form an $S \times 16$ matrix, and fit them in a logistic regression model for classifying the healthy versus diseased states. The Area Under the Curve (AUC) of the Receiver Operating Characteristic is 86.6%. On the other hand, when we fit the 6 ROIs with low variability in logistic regression, the AUC increases to 96.1%.

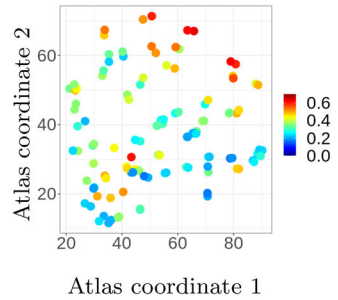
An explanation for the above results is that Alzheimer's disease does different degrees of damage in the frontal and parietal



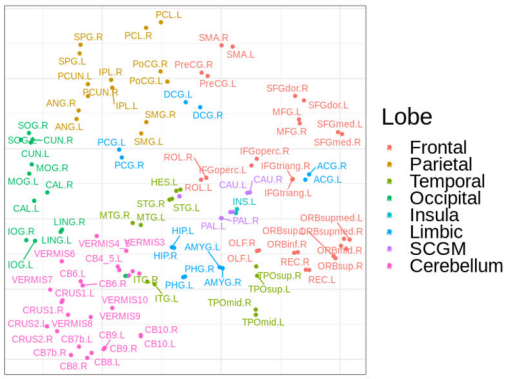
(a) Pairwise distance of the latent coordinates $\|z^{(s)} - z^{(s')}\|_F$ between subjects, the first 64 subjects are healthy and the latter 102 are diseased.



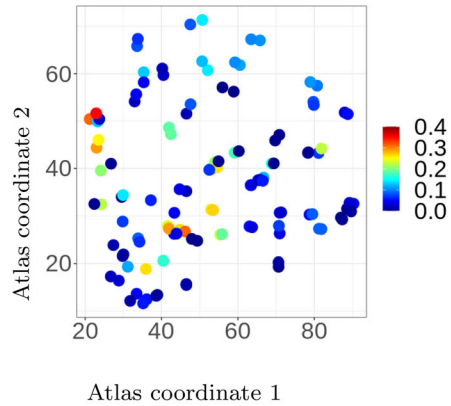
(b) Within-group variances of $z_i^{(\cdot)}$ for subjects in the healthy group, plotted over the automated anatomical labeling atlas.



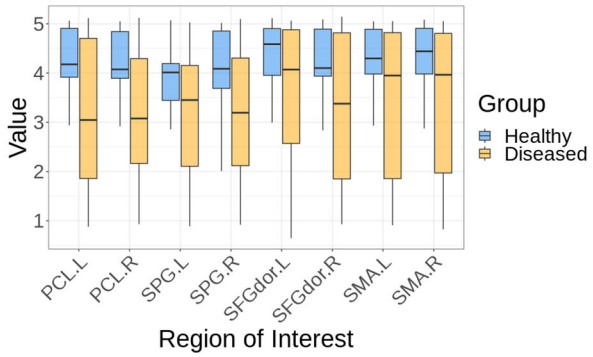
(c) Within-group variances of $z_i^{(\cdot)}$ for subjects in the diseased group, plotted over the automated anatomical labeling atlas.



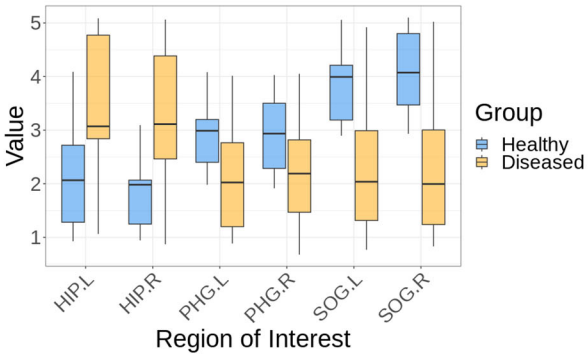
(d) The regions of interest (ROIs) colored by the associated lobe names, under the automated anatomical labeling atlas.



(e) Intraclass correlation coefficients for the regions of interest $(1 - \hat{\sigma}^2_{\text{within-group},i}/\hat{\sigma}^2_{\text{total},i})$.



(f) Boxplot visualization of the latent coordinates for the regions with high variability in the diseased group.



(g) Boxplot visualization of the latent coordinates for the regions with low variability in the diseased group.

Figure 5. Using the latent coordinates to characterize the heterogeneity within the subjects.

lobes (see the two distinct clusterings in Figure 4(c) and (d)), and the severity of the damage can vary from person to person. On the other hand, the hippocampus region (HIP and PHG), important for memory consolidation, is known to be commonly affected by Alzheimer's disease (Braak and Braak 1991; Klimova et al. 2015), which explains the low heterogeneity in the diseased group. Further, to our best knowledge, the high discriminability of the superior occipital gyrus (SOG) is a new quantitative finding, that could be meaningful for a further clinical study.

For validation, without using any group information, we concatenate those $z_i^{(s)}$'s over all $i = 1, \ldots, 116$ and form an $S \times 232$ matrix and use lasso logistic regression to classify the two

groups. When 12 predictors are selected (as a similar-size model to the one above using six ROIs), the AUC is 96.4%. Since $z_i^{(s)}$'s are obtained in an unsupervised way, this validation result shows that the multi-view clustering model produces meaningful representation for the nodes in this Alzheimer's disease data. We provide further details on the clusterings, including the number of clusters, and the posterior co-assignment probability matrices in the supplementary materials S4.5.

7. Discussion

In this article, we present our discovery of a probabilistic model for popular spectral clustering algorithms. This enables straightforward uncertainty quantification and model-based extensions through the Bayesian framework. There are several directions worth exploring. First, our consistency theory is conducted under the condition of separable sets, similar to Ascolani et al. (2022). For general cases with non-separable sets, clustering consistency (especially on estimating K) is challenging to achieve; to our best knowledge, existing consistency theory only applies to data generated independently from a mixture model (Miller and Harrison 2018; Zeng, Miller, and Duan 2023). For data generated dependently via a graph, this is still an unsolved problem. Second, in all of our forest models, we have been careful in choosing densities with tractable normalizing constants. One could relax this constraint by using densities $f(y_i \mid y_j, \theta) = \alpha_f g_f(y_i \mid y_j; \theta)$ and $r(y_i; \theta) = \alpha_r g_r(y_i; \theta)$, with g some similarity function, and (α_f, α_r) potentially intractable. In these cases, the forest posterior becomes $\Pi(\mathcal{T} \mid .) \propto (\lambda\alpha_r/\alpha_f)^K \prod_{(0,i)\in\mathcal{T}} g_r(y_i; \theta) \prod_{(i,j)\in\mathcal{T}} g_r(y_i \mid y_j; \theta)$. Therefore, one could choose an appropriate $\tilde{\lambda} = \lambda\alpha_r/\alpha_f$ (equivalent to choosing some value of λ), without knowing the value of α_f or α_r; nevertheless, how to calibrate $\tilde{\lambda}$ still requires further study. Third, a related idea is the Dirichlet Diffusion Tree (Neal 2003), which considers a particle starting at the origin, following the path of previous particles, and diverging at a random time. The data are collected as the locations of particles at the end of a time period. Compared to the forest process, the diffusion tree process has the conditional likelihood given the tree invariant to the ordering of the data index, which is a stronger property compared to the marginal exchangeability of the data points. Therefore, it is interesting to further explore the relationship between those two processes.

Supplementary Materials

The supplementary materials include additional details on the model-based extensions, proofs of theorems, numerical experiments and a link to the R source code.

Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada.

Disclosure Statement

The authors report that there are no competing interests to declare.

Funding

Private sector contributions are facilitated by the Foundation for the National Institutes of Health (*www.fnih.org*). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see *www.adni-info.org*.

ORCID

Arkaprava Roy http://orcid.org/0000-0001-6976-9967

References

Aldous, D. J. (1990), "The Random Walk Construction of Uniform Spanning Trees and Uniform Labelled Trees," *SIAM Journal on Discrete Mathematics*, 3, 450–465. [6,7]

Ascolani, F., Lijoi, A., Rebaudo, G., and Zanella, G. (2022), "Clustering Consistency With Dirichlet Process Mixtures," arXiv preprint arXiv:2205.12924. [12]

Aßmann, C., Boysen-Hogrefe, J., and Pape, M. (2016), "Bayesian Analysis of Static and Dynamic Factor Models: An Ex-Post Approach Towards the Rotation Problem," *Journal of Econometrics*, 192, 190–206. [7]

Banerjee, S., Akbani, R., and Baladandayuthapani, V. (2015), "Bayesian Nonparametric Graph Clustering," arXiv preprint arXiv:1509.07535. [2]

Barry, D., and Hartigan, J. A. (1993), "A Bayesian Analysis for Change Point Problems," *Journal of the American Statistical Association*,88, 309–319. [5]

Blackwell, D., and MacQueen, J. B. (1973), "Ferguson Distributions via Pólya Urn Schemes," *The Annals of Statistics*, 1, 353–355. [5]

Blei, D. M., and Frazier, P. I. (2011), "Distance Dependent Chinese Restaurant Processes," *Journal of Machine Learning Research*, 12, 2461–2488. [2]

Boykov, Y., Veksler, O., and Zabih, R., (2001), "Fast Approximate Energy Minimization via Graph Cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23, 1222–1239. [7]

Braak, H., and Braak, E. (1991), "Neuropathological Stageing of Alzheimer-Related Changes," *Acta Neuropathologica*, 82, 239–259. [11]

Broder, A. Z. (1989), "Generating Random Spanning Trees," in *Annual Symposium on Foundations of Computer Science* (Vol. 89), pp. 442–447. [6,7]

Byrne, S., and Dawid, A. P. (2015), "Structural Markov Graph Laws for Bayesian Model Uncertainty," *The Annals of Statistics*, 43, 1647–1681. [2]

Cai, D., Campbell, T., and Broderick, T. (2021), "Finite Mixture Models Do Not Reliably Learn the Number of Components," in *International Conference on Machine Learning*, pp. 1158–1169. PMLR. [1]

Cao, X., Khare, K., and Ghosh, M. (2019), "Posterior Graph Selection and Estimation Consistency for High-Dimensional Bayesian DAG Models," *The Annals of Statistics*, 47, 319–348. [8]

Chaiken, S., and Kleitman, D. J. (1978), "Matrix Tree Theorems," *Journal of Combinatorial Theory*, Series A, 24, 377–381. [7]

Chandra, N. K., Canale, A., and Dunson, D. B. (2023), "Escaping the Curse of Dimensionality in Bayesian Model Based Clustering," *Journal of Machine Learning Research*, 24, 1–42. [6]

Chi, Y., Song, X., Zhou, D., Hino, K., and Tseng, B. L. (2007), "Evolutionary Spectral Clustering by Incorporating Temporal Smoothness," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 153–162. [2]

Coretto, P., and Hennig, C. (2016), "Robust Improper Maximum Likelihood: Tuning, Computation, and a Comparison With Other Methods for Robust Gaussian Clustering," *Journal of the American Statistical Association*, 111, 1648–1659. [1]

Crowley, E. M. (1997), "Product Partition Models for Normal Means," *Journal of the American Statistical Association*, 92, 192–198. [5]

Dahl, D. B., Johnson, D. J., and Müller, P. (2022), "Search Algorithms and Loss Functions for Bayesian Clustering," *Journal of Computational and Graphical Statistics*, 31, 1189–1201. [7]

DeBruine, Z. J., Melcher, K., and Triche Jr, T. J. (2021), "Fast and Robust Non-Negative Matrix Factorization for Single-Cell Experiments," bioRxiv, 2021–09. [7]

Diaconis, P. (1977), "Finite Forms of de Finetti's Theorem on Exchangeability," *Synthese*, 36, 271–281. [5]

Duan, L. L., and Dunson, D. B. (2021a), "Bayesian Distance Clustering," *Journal of Machine Learning Research*, 22, 1–27. [2]

——— (2021b), "Bayesian Spanning Tree: Estimating the Backbone of the Dependence Graph," arXiv preprint arXiv:2106.16120. [2]

Duan, L. L., Michailidis, G., and Ding, M. (2023), "Bayesian Spiked Laplacian Graphs," *Journal of Machine Learning Research*, 24, 1–35. [2]

Edwards, D., De Abreu, G. C., and Labouriau, R. (2010), "Selecting High-Dimensional Mixed Graphical Models Using Minimal AIC or BIC Forests," *BMC Bioinformatics*, 11, 1–13. [2]

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996), "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–231, AAAI Press. [1]

Fraley, C., and Raftery, A. E. (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631. [1]

Frey, B. J., and Dueck, D. (2007), "Clustering by Passing Messages Between Data Points," *Science*, 315, 972–976. [1]

Frühwirth-Schnatter, S., and Pyne, S. (2010), "Bayesian Inference for Finite Mixtures of Univariate and Multivariate Skew-Normal and Skew-t Distributions," *Biostatistics*, 11, 317–336. [1]

Geng, J., Bhattacharya, A., and Pati, D. (2019), "Probabilistic Community Detection With Unknown Number of Communities," *Journal of the American Statistical Association*, 114, 893–905. [2]

Gower, J. C., and Ross, G. J. (1969), "Minimum Spanning Trees and Single Linkage Cluster Analysis," *Journal of the Royal Statistical Society*, Series C, 18, 54–64. [2,3]

Guha, S., and Baladandayuthapani, V. (2016), "A Nonparametric Bayesian Technique for High-Dimensional Regression," *Electronic Journal of Statistics*, 10, 3374–3424. [1]

Han, X., Tong, X., and Fan, Y. (2021), "Eigen Selection in Spectral Clustering: A Theory-Guided Practice," *Journal of the American Statistical Association*, 118, 109–121. [2]

Hartigan, J. A. (1990), "Partition Models," *Communications in Statistics-Theory and Methods*, 19, 2745–2756. [5]

Klimova, B., Maresova, P., Valis, M., Hort, J., and Kuca, K. (2015), "Alzheimer's Disease and Language Impairments: Social Intervention and Medical Treatment," *Clinical Interventions in Aging*, 10, 1401–1408. [11]

Kosmidis, I., and Karlis, D. (2016), "Model-Based Clustering Using Copulas With Applications," *Statistics and Computing*, 26, 1079–1099. [1]

Kumar, A., Rai, P., and Daume, H. (2011), "Co-regularized Multi-view Spectral Clustering," in *Advances in Neural Information Processing Systems* (Vol. 24), eds. J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Curran Associates, Inc. [2]

Lee, S. X., and McLachlan, G. J. (2016), "Finite Mixtures of Canonical Fundamental Skew t-Distributions," *Statistics and Computing*, 26, 573–589. [1]

Lei, J., and Lin, K. Z. (2022), "Bias-Adjusted Spectral Clustering in Multi-Layer Stochastic Block Models," *Journal of the American Statistical Association*, 1–13. [2]

Lei, J., and Rinaldo, A. (2015), "Consistency of Spectral Clustering in Stochastic Block Models," *The Annals of Statistics*, 43, 215–237. [2]

Lewis, J. R., MacEachern, S. N., and Lee, Y. (2021), "Bayesian Restricted Likelihood Methods: Conditioning on Insufficient Statistics in Bayesian Regression," *Bayesian Analysis*, 16, 1393–1462. [2]

Luo, Z., Sang, H., and Mallick, B. (2021), "A Bayesian Contiguous Partitioning Method for Learning Clustered Latent Variables," *Journal of Machine Learning Research*, 22, 1–52. [2]

MacQueen, J. (1967), "Classification and Analysis of Multivariate Observations," in *5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297. [1,2]

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2017), "Identifying Mixtures of Mixtures Using Bayesian Estimation," *Journal of Computational and Graphical Statistics*, 26, 285–295. [1]

McDaid, A. F., Murphy, T. B., Friel, N., and Hurley, N. J. (2013), "Improved Bayesian Inference for the Stochastic Block Model With Application to Large Networks," *Computational Statistics & Data Analysis*, 60, 12–31. [2]

Medvedovic, M., and Sivaganesan, S. (2002), "Bayesian Infinite Mixture Model Based Clustering of Gene Expression Profiles," *Bioinformatics*, 18, 1194–1206. [7]

Meilă, M., and Jaakkola, T. (2006), "Tractable Bayesian Learning of Tree Belief Networks," *Statistics and Computing*, 16, 77–92. [2]

Meila, M., and Jordan, M. I. (2000), "Learning With Mixtures of Trees," *Journal of Machine Learning Research*, 1, 1–48. [2]

Miller, J. W. (2019), "An Elementary Derivation of the Chinese Restaurant Process From Sethuraman's Stick-Breaking Process," *Statistics & Probability Letters*, 146, 112–117. [5]

Miller, J. W., and Dunson, D. B. (2018), "Robust Bayesian Inference via Coarsening," *Journal of the American Statistical Association*, 114, 1113–1125. [1]

Miller, J. W., and Harrison, M. T. (2018), "Mixture Models With a Prior on the Number of Components," *Journal of the American Statistical Association*, 113, 340–356. [12]

Molitor, J., Papathomas, M., Jerrett, M., and Richardson, S. (2010), "Bayesian Profile Regression With an Application to the National Survey of Children's Health," *Biostatistics*, 11, 484–498. [7]

Mosbah, M., and Saheb, N. (1999), "Non-Uniform Random Spanning Trees on Weighted Graphs," *Theoretical Computer Science*, 218, 263–271. [6,7]

Müller, P., and Quintana, F. (2010), "Random Partition Models With Regression on Covariates," *Journal of Statistical Planning and Inference*, 140, 2801–2808. [1]

Müller, P., Quintana, F., and Rosner, G. L. (2011), "A Product Partition Model With Regression on Covariates," *Journal of Computational and Graphical Statistics*, 20, 260–278. [6]

Neal, R. M. (2003), "Density Modeling and Clustering Using Dirichlet Diffusion Trees," *Bayesian Statistics*, 7, 619–629. [12]

Ng, S.-K., McLachlan, G. J., Wang, K., Ben-Tovim Jones, L., and Ng, S.-W. (2006), "A Mixture Model With Random-Effects Components for Clustering Correlated Gene-Expression Profiles," *Bioinformatics*, 22, 1745–1752. [1]

Noble, S., Scheinost, D., and Constable, R. T. (2021), "A Guide to the Measurement and Interpretation of fMRI Test-Retest Reliability," *Current Opinion in Behavioral Sciences*, 40, 27–32. [10]

Nowicki, K., and Snijders, T. A. B. (2001), "Estimation and Prediction for Stochastic Blockstructures," *Journal of the American Statistical Association*, 96, 1077–1087. [2]

Paganin, S., Herring, A. H., Olshan, A. F., and Dunson, D. B. (2021), "Centered Partition Processes: Informative Priors for Clustering," *Bayesian Analysis*, 16, 301–370. [6]

Park, J.-H., and Dunson, D. B. (2010), "Bayesian Generalized Product Partition Model," *Statistica Sinica*, 20, 1203–1226. [1]

Petrone, S., Guindani, M., and Gelfand, A. E. (2009), "Hybrid Dirichlet Mixture Models for Functional Data," *Journal of the Royal Statistical Society*, Series B, 71s, 755–782. [6]

Quintana, F. A., and Iglesias, P. L. (2003), "Bayesian Clustering and Product Partition Models," *Journal of the Royal Statistical Society*, Series B, 65, 557–574. [5]

Rasmussen, C., Bernard, J., Ghahramani, Z., and Wild, D. L. (2008), "Modeling and Visualizing Uncertainty in Gene Expression Clusters Using Dirichlet Process Mixtures," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6, 615–628. [7]

Ren, L., Du, L., Carin, L., and Dunson, D. B. (2011), "Logistic Stick-Breaking Process," *Journal of Machine Learning Research*, 12, 203–239. [1]

Rigon, T., Herring, A. H., and Dunson, D. B. (2023), "A Generalized Bayes Framework for Probabilistic Clustering," *Biometrika*, 110, 559–578. [2]

Rodríguez, C. E., and Walker, S. G. (2014), "Univariate Bayesian Nonparametric Mixture Modeling With Unimodal Kernels," *Statistics and Computing*, 24, 35–49. [1]

Rohe, K., Chatterjee, S., and Yu, B. (2011), "Spectral Clustering and the High-Dimensional Stochastic Blockmodel," *The Annals of Statistics*, 39, 1878–1915. [2]

Rolls, E. T., Huang, C.-C., Lin, C.-P., Feng, J., and Joliot, M. (2020), "Automated Anatomical Labelling Atlas 3," *Neuroimage*, 206, 116189. [10]

Schild, A. (2018), "An Almost-Linear Time Algorithm for Uniform Random Spanning Tree Generation," in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 214–227. [7]

Shi, D., Zhang, H., Wang, S., Wang, G., and Ren, K. (2021), "Application of Functional Magnetic Resonance Imaging in the Diagnosis of Parkinson's Disease: A Histogram Analysis," *Frontiers in Aging Neuroscience*, 13, 624731. [10]

Shi, T., Belkin, M., and Yu, B. (2009), "Data Spectroscopy: Eigenspaces of Convolution Operators and Clustering," *The Annals of Statistics*, 37, 3960–3984. [2]

Snijders, T. A., and Nowicki, K. (1997), "Estimation and Prediction for Stochastic Blockmodels for Graphs With Latent Block Structure," *Journal of Classification*, 14, 75–100. [2]

Socher, R., Maas, A., and Manning, C. (2011), "Spectral Chinese Restaurant Processes: Nonparametric Clustering Based on Similarities," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 698–706. JMLR Workshop and Conference Proceedings. [2]

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006), "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, 101, 1566–1581. [6]

Von Luxburg, U. (2007), "A Tutorial on Spectral Clustering," *Statistics and Computing*, 17, 395–416. [2,3]

Wade, S., and Ghahramani, Z. (2018), "Bayesian Cluster Analysis: Point Estimation and Credible Balls," *Bayesian Analysis*, 13, 559–626. [1,7]

Wu, S., Feng, X., and Zhou, W. (2014), "Spectral Clustering of High-Dimensional Data Exploiting Sparse Representation Vectors," *Neurocomputing*, 135, 229–239. [6]

Yu, Y., Wang, T., and Samworth, R. J. (2015), "A Useful Variant of the Davis–Kahan Theorem for Statisticians," *Biometrika*, 102, 315–323. [8]

Zelnik-Manor, L., and Perona, P. (2005), "Self-Tuning Spectral Clustering," in *Advances in Neural Information Processing Systems* (Vol. 17). [2,6]

Zeng, C., Miller, J. W., and Duan, L. L. (2023), "Consistent Model-based Clustering Using the Quasi-Bernoulli Stick-Breaking Process," *Journal of Machine Learning Research*, 24, 1–32. [12]