Spectral Clustering, Spanning Forest, and Bayesian Forest Process

Leo L. Duan* and Arkaprava Roy†

for the Alzheimer’s Disease Neuroimaging Initiative

Abstract: Spectral clustering algorithms are very popular. Starting from a pairwise similarity matrix, spectral clustering gives a partition of data that approximately minimizes the total similarity scores across clusters. Since there is no need to model how data are distributed within each cluster, such a method enjoys algorithmic simplicity and robustness in clustering non-Gaussian data such as those near manifolds. Nevertheless, several important questions are unaddressed, such as how to estimate the similarity scores and cluster assignment probabilities, as important uncertainty estimates in clustering. In this article, we propose to solve these problems with a discovered generative modeling counterpart. Our clustering model is based on a spanning forest graph that consists of several disjoint spanning trees, with each tree corresponding to a cluster. Taking a Bayesian approach, we assign proper densities on the root and leaf nodes, and we prove that the posterior mode is almost the same as spectral clustering estimates. Further, we show that the associated generative process, named “forest process”, is a continuous extension to the classic urn process, hence inheriting many nice properties such as having unbounded support for the number of clusters and being amenable to existing partition probability function; at the same time, we carefully characterize their differences. We demonstrate a novel application in joint clustering of multiple-subject functional magnetic resonance imaging scans of the human brain.

Keywords: Consistent Clustering, Model-based Clustering, Normalized graph-cut, Spectral Equivalence.

1 Introduction

Clustering algorithms partition data \(y_1, \ldots, y_n\) into disjoint groups. There is a great amount of related literature, ranging from various algorithms such as K-means, DBSCAN {Ester et al. (1996); Frey and Dueck (2007); among others} to mixture model-based approaches {reviewed by Fraley and Raftery (2002)}. In the statistical community, the model-based clustering approaches are especially popular, arguably, for two reasons: (i) one can easily estimate the parameters within and quantify clustering uncertainty (such as cluster assignment probability for each data point) using a likelihood-based framework; (ii) one can easily extend these clustering models to handle complex but common challenges arisen in statistical applications, such as incorporating random effects (Ng et al., 2006), covariate dependency (Müller and Quintana, 2010), time series structure (Maharaj et al., 2019), etc. Roughly speaking, one sees each cluster of data as independent deviations from a “center”, with deviations characterized by a probability distribution, with Gaussian distribution as one popular choice, and extensions to high-dimensional clustering (McLachlan et al., 2003), unimodal distribution (Rodríguez and Walker, 2014) and copula (Kosmidis and Karlis, 2016) for non-Gaussian clustering, etc.

On the other hand, for modern data applications, substantial limitations have been found for those model-based approaches. To be concrete, a cluster meaningful in the real world (such as photos of a person) may be formed based on some similarity between the data points, as opposed to modeling them as independent

---

*Department of Statistics, University of Florida, Gainesville, FL, email: li.duan@ufl.edu
†Department of Biostatistics, University of Florida, Gainesville, FL email: ark007@ufl.edu
deviations from a center. Forcing a simple model (such as Gaussian mixture) on these complex data often leads to undesirable results, such as obtaining too many clusters in practice {for theoretical studies related to mixture model misspecification, see Miller and Dunson (2018), and more recently, Cai et al. 2020 arXiv preprint 2007.04470} and difficult interpretation.

For such tasks, spectral clustering algorithms have shown impressive performances. The basic idea is that, based on a pre-computed similarity score matrix between all pairs of data, one tries to minimize a normalized graph cut loss (Von Luxburg, 2007) — the sum of similarity across different clusters (adjusted with respect to cluster sizes) that could be lost due to data partitioning. Since the loss function only involves across-cluster similarity, robustness is obtained on the “unmodeled” part within each cluster — for example, a cluster could be formed near a continuous manifold, which has one data point close to a neighbor, while still allowing multiple/most points to be far apart. Due to good empirical performances and algorithmic simplicity, spectral clustering has shown tremendous potentials in applications (Belkin and Niyogi, 2002) and received a series of rigorous theoretical studies {see Belkin and Niyogi (2004, 2008), among others}, including consistency in the sense of the sample-based eigenvectors converging to the population eigenfunctions (Von Luxburg et al., 2008). Further, due to a connection to stochastic block models in network data analysis, comparably rich literature has been developed on applications and theory as well {see Rohe et al. (2011); Lei and Rinaldo (2015), and the references within}.

Despite its success in producing clustering point estimates, spectral clustering suffers from a few major limitations. Namely, compared with model-based clustering, it falls short on several important aspects as described above — estimating the similarity scores and involved parameters, quantifying clustering uncertainty, and allowing flexible statistical extensions.

To address these issues, we propose to use a generative model that has its (conditional) posterior mode on clustering almost equivalent to the estimate from spectral clustering algorithms using normalized Laplacian. This model is based on a spanning forest graph that consists of several disjoint spanning trees, with each tree corresponding to a cluster. From a generative perspective, each tree is initiated with a data point drawn from a root distribution, and is gradually grown by a new data point drawn from a leaf distribution conditioned on an existing one. We show that the infinite extension leads to a continuous extension to the classic Polya urn process, while we carefully characterize the differences.

We are certainly not the first ones trying to address uncertainty quantification problems for spectral clustering and related algorithms. Early works include distance-based Polya urn process (Blei and Frazier, 2011; Socher et al., 2011), distance clustering (Duan and Dunson, 2021), Dirichlet process mixture model on Laplacian eigenmaps (Banerjee et al 2015, arXiv preprint 1509.07535), mixture of finite mixtures clustering for stochastic block model (Geng et al., 2019), probabilistic decomposition of Laplacian (Duan et al. 2021, arXiv preprint 1910.02471), among others. However, a fundamental difference is that these existing works view the similarity, distance or Laplacian matrix as given or as if the source data; whereas our method models $y_i$ directly, hence allowing a fully Bayesian treatment and straightforward model extension in the observed domain.

2 Method

2.1 Clustering Likelihood based on Spanning Forest

For data $y = (y_1, y_2, \ldots, y_n)$, imagine that they are generated based on a likelihood $\mathcal{L}(y; \mathcal{F}, \theta)$, that depends on a graph $\mathcal{F}$ known as “rooted spanning forest” (or simply, “forest”) and some other parameter $\theta$. To introduce the necessary background, a spanning forest is the union of $K$ disjoint spanning trees, and each tree is a connected graph $T_k = (V_k, E_k)$ with $V_k$ the set of nodes and $E_k$ the set of edges. Importantly, (i) each tree has no cycles hence it is the smallest graph connecting $|V_k|$ nodes, and (ii) those trees give a partition of the data index $(1, \ldots, n) = (V_1 \cup \ldots \cup V_K)$. Further, in each tree $T_k$, there is one node $k^* \in V_k$ chosen as the root. Therefore, we have the graph $\mathcal{F} = (T_1, \ldots, T_k, 1^*, \ldots, k^*)$.

Consider a likelihood that can be factorized over trees in the following way,

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

(1)
where for ease of modeling and computation, we require $f$ to be 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.

In this article, for simplicity, we focus on continuous $y_i \in \mathbb{R}^p$. Due to the graph representation, we will refer to $r(\cdot; \theta)$ as a “root” density, and $f(\cdot \mid y_j; \theta)$ as a “leaf” density. With some notation overloading, we will refer to $r$ and $f$ as distributions as well.

![Figure 1: Three examples of clusters that can be represented by a spanning forest.](image)

For now, we first illustrate the high flexibility of a spanning forest in representing clusters. Figure 1 shows point estimates of $F$ from three simulated examples. Note that some clusters are not in an elliptical or convex shape, and some are not distributed around a center. Rather, each cluster is formed by connecting a point to another one nearby.

Next, we discuss our choice of densities for $r$ and $f$. To favor edges formed between $y_i$ and $y_j$ that are close by, we choose a Gaussian density for $y_i$, with $y_j$ as its location parameter and a small scale. For the root nodes, we choose an isotropic Cauchy as a diffuse distribution, so that the likelihood is less sensitive to the choice of a node as root.

$$f(y_i \mid y_j; \theta) = (2\pi \sigma_{i,j})^{-p/2} \exp \left\{ -\frac{\|y_i - y_j\|^2}{2\sigma_{i,j}} \right\},$$

$$r(y_i; \theta) = (\pi \gamma)^{-p} \prod_{l=1}^{p} \frac{1}{1 + (y_{i,l} - \mu)^2 / \gamma^2}.$$  

(2)

where $y_{i,l}$ is the $l$th element of $y_i$, $\sigma_{ij} > 0$ and $\gamma > 0$ are scale parameters. As we can imagine, the magnitudes of distances between neighboring points may differ significantly from cluster to cluster (and even from location to location), 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 prior as described soon.

**Remark 1** In (2), we are effectively selecting edges based on Euclidean distances $\|y_i - y_j\|_2$. This can be easily generalized to other forms, for example, using subspace distances for high-dimensional image clustering (Vidal, 2011). In this article, we focus on the form of (2) for simplicity.

Using (1) as a model for the data, we can estimate $F$ and extract partition $(V_1, \ldots, V_K)$ for clustering. To record the clustering membership for each data point, we use $c_i = k$ if $i \in V_k$. We refer to this approach as forest model clustering, and will show the connection of (1) to spectral clustering algorithms in Section 4.

### 2.2 Prior on Forest

With the likelihood depending on $F$, we now assign it to a prior distribution. We specify prior according to a factorization of $\Pi_0(F)$: 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 set of $V_k$.

$$\Pi_0(F) = \Pi_0(K) \Pi_0(V_1, \ldots, V_K \mid K) \prod_{k=1}^{K} \left\{ \Pi_0(E_k \mid V_k) \Pi_0(k^* \mid V_k) \right\},$$
in which, we make edges $E_k$ as well as root $k^*$ conditionally independent over different $k$ given $V_k$.

We now describe a specific set of priors, based on a “top-down” approach aiming to create a simple form on $\Pi_0(\mathcal{F})$ for convenient posterior computation (we will discuss alternative form in the next section). For the shape of each tree (as in $E_k$), we use uniform prior $\Pi(E_k \mid V_k) = n_k^{-{(n_k - 2) +}}$, where $n_k = |V_k|$, and $n_k^{-{(n_k - 2) +}}$ is the total number of possible spanning trees for $n_k$ nodes; $(x)_+ = x$ if $x > 0$, otherwise 0. Similarly, we use uniform for root $\Pi_0(k^* \mid V_k) = 1/n_k$. To cancel out these complicated terms, we further assign:

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

$$\Pi_0(K) \propto \frac{\lambda^K}{K!} \sum_{\{V_1^* \ldots, V_K^*\}} \prod_{k=1}^{K} |V_k^*|^{(|V_k^*| - 1)}, \quad K = 1, \ldots, n,$$

where for notational ease, we omit the normalizing constant in $\Pi_0(K)$. We choose the above forms for two reasons: (i) multiplying those terms together yields a joint

$$\Pi_0(\mathcal{F}) \propto \lambda^K / (K!),$$

which is easy to handle in posterior computation; (ii) the marginal probability $\Pi_0(K)$ can be interpreted as proportional to the multiplicity of forming rooted spanning forest with $K$ roots, adjusted by a truncated Poisson from $K \in (1, \ldots, n)$. Therefore, we can use a small $\lambda$ impose a prior preference on smaller $K$.

### 2.3 Forest Process

We now discuss the above generative process in the prior and likelihood from a point of view of simulating a stochastic process where $n$ grows indefinitely. Omitting parameter $\theta$, it is not hard to see an equivalence to

$$y_1 \sim r(\cdot),$$

$$y_i \mid y_1, \ldots, y_{i-1} \sim \sum_{j=1}^{i-1} \pi_j^{[i]} f(\cdot \mid y_j) + \pi_i^{[i]} r(\cdot) \text{ for } i = 2, 3, \ldots,$$

for some probability vector $(\pi_1^{[i]}, \ldots, \pi_i^{[i]})$ that sums up to one. In this process, each time we draw from $r(\cdot)$, we start a new tree at an isolated node $i$; each time we draw from $f(\cdot \mid y_j)$, we grow an existing tree by adding an edge $(j, i)$.

Note that if we used $f(\cdot \mid y_j) = \delta_{y_j}(\cdot)$ representing a point mass at $y_j$, then the above would be the same as the famous Pólya urn process (Blackwell and MacQueen, 1973) with $r$ as its base measure. Therefore, the forest process is a generalization. Importantly, by replacing the point mass with a non-atomic distribution, we obtain a continuous joint distribution of $(y_1, \ldots, y_n)$ for finite $n$, while still encoding the clustering information via edge $(i, j)$’s.

Naturally, this connection may suggest an alternative “bottom-up” approach for prior specification by assigning forms on $(\pi_1^{[i]}, \ldots, \pi_i^{[i]})$ directly. For example, one can follow the Chinese resturant process and set $\pi_j^{[i]} = 1/(i - 1 + \alpha)$ for $j = 1, \ldots, (i - 1)$, and $\pi_i^{[i]} = \alpha / (i - 1 + \alpha)$. This would lead to a partition probability function $\Pi_0(K, V_1, \ldots, V_K) = \alpha^K \Gamma(\alpha) \Gamma(n + \alpha) / \prod_{k=1}^{K} \Gamma(n_k)$ (see Miller (2019) for a recent proof). However, when it comes to the prior probability for the forest, assuming uniform $\Pi_0(E_k \mid V_k)$ and $\Pi_0(k^* \mid V_k)$, we would have $\Pi_0(\mathcal{F}) = \alpha^K \Gamma(\alpha) \Gamma(n + \alpha) / \prod_{k=1}^{K} \Gamma(n_k) n_k^{-(n_k - 1)}$, which is less convenient than (4) for computation; therefore, we will use (4) in the rest of article.

**Remark 2** Compared to the typical partition probability function, the extra $\Pi_0(E_k, k^* \mid V_k) = n_k^{-(n_k - 1)}$ can be understood as a term that accommodates the rooted tree graph multiplicity given each partition $V_k$. In the case of $f(\cdot \mid y_j) = \delta_{y_j}(\cdot)$, all points in $V_k$ coincide on a single location, and we would have a multiplicity reduced to one.
2.4 Calibrated Prior for Scale Parameters

In (2), the scale parameters $\sigma_{i,j}$ and $\gamma$ play an important role in influencing the clustering uncertainty — as one can imagine, if the root scale $\gamma$ is too small compared to the range of data, then we would effectively force the roots to be near $\mu$; on the other hand, if the leaf scale $\sigma_{i,j}$ is much larger than $\gamma$, then we would have leaf density to be too diffuse $f(y_i | y_j) \ll r(y_i)$ for most of $y_i$, leading to too many clusters. Therefore, we need a way to calibrate the scale priors using some information from the data.

We are inspired by the classic high-density clustering algorithms {Hartigan (1981); Ester et al. (1996); Steinwart (2011), among others}. To briefly review their ideas, high-density clustering views the data as independently generated based on some density function $\tilde{f}$, we can exclude those data points with $\tilde{f}(y_i) \leq \epsilon$ using some small threshold $\epsilon > 0$. As $n \rightarrow \infty$, it is not hard to see that those $\{y_i : \tilde{f}(y_i) > \epsilon\}$ in the high-density region would form several dense sets that are disjoint — each set is now a cluster, and the sets are separated by the low-density region. This gives a very flexible representation for the clusters, as illustrated in Figure 2(a). On the other hand, the challenges are that one does not know how to appropriately choose $\epsilon$, and how to quantify the clustering uncertainty.

We now show how our spanning forest clustering can exploit the key idea of high-density clustering, while addressing the above challenges — we consider $r(\cdot | \mu, \lambda)$ as a diffuse distribution that can generate points in the low-density region, where $f(y_i | y_j; \sigma_{i,j})$ gives a dependent density estimation in the high-density region.

Specifically, for $r(\cdot)$, we first standardize each vector $(y_{1,j}, \ldots, y_{n,j})$ to have mean 0 and standard deviation 1 for each $j = 1, \ldots, p$, which enables us to simply fix $\mu = 0$ and $\lambda = 1$. The diffuseness of the standard Cauchy density makes $r(\cdot)$ close to a small constant in the support of the data. For each tree to more likely be formed over small distances via $f(y_i | y_j; \sigma_{i,j})$ with $\sigma_{i,j} = \tilde{\sigma}_i \tilde{\sigma}_j$, while allowing borrowing of information among $\tilde{\sigma}_i$’s, we use a non-parametric discrete prior $\tilde{\sigma}_i \sim \sum_{k=1}^{H} w_h \delta_{d_h}(\cdot)$, with $(w_1, \ldots, w_H) \sim \text{Dir}(0.1, \ldots, 0.1)$ (so that only a few $w_h$ are away from zero, in this article, we use $H = 20$), while having $d_h$ from the uniform distribution based on maximum shortest distance (adjusted by $p$) $d_h \sim \text{Uniform}(0, \max_{i=1, \ldots, n} \min_{j \neq i} ||y_i - y_j||_2 / \sqrt{p})$. As shown in Figure 2(b), the leaf density $f(y_i | y_j, \sigma_{i,j})$ dominates over the root density $r(\cdot)$, as long as $||y_i - y_j||_2$ is smaller than roughly $2\sqrt{\sigma_{i,j}}$. Therefore, a point within $2\sqrt{\sigma_{i,j}}$ of another point is more likely to be considered as in the high density region.

![Figure 2: Illustration on how spanning forest clustering can lead to dependent extension of high density clustering, with a calibrated choice of $\sigma_{i,j}$ and $\gamma$.](image)

(a) Illustration of the classic high density clustering algorithm by forming disjoint sets (clusters, shown in blue) via thresholding the density $\{y_i : \tilde{f}(y_i) \geq \epsilon\}$, some points (red) are left out in the low density region. The spanning forest likelihood mimics this by drawing roots in the red, while forming short edges in the blue.

(b) The Gaussian leaf density $f(\cdot)$ (dashed lines, corresponding to $\sqrt{\sigma_{i,j}} = 0.1, 0.2$ or 0.5) is larger than the root Cauchy density $r(\cdot)$ (red solid line) when the distance $||y_i - y_j||_2$ is smaller than roughly $2\sqrt{\sigma_{i,j}}$; therefore, a point within $2\sqrt{\sigma_{i,j}}$ of another point is more likely to be considered as in the high density region.
Remark 3 To be rigorous, we want to acknowledge that assigning $\sigma_{i,j}$ with a minimum distance-based prior breaks the “projectivity” of the forest process, under which, one would obtain the marginal distribution of $(y_1,\ldots,y_n)$ by integrating over $y_{n+1}$. There are alternative priors for $\sigma_{i,j}$’s that satisfy projectivity. For example, conditioned on the set distances between the high density regions $R_k, R_k^\prime$, $\text{dist}(k,k^\prime) = \inf_{x \in R_k, y \in R_k^\prime} \|x - y\|_2$, we could set $\sigma_{i,j}$ to be slightly smaller than $\text{dist}(k,k^\prime)$ if $c_i = k, c_j = k^\prime \neq k$, while set $\sigma_{i,j}$ close to zero if $c_i = c_j$. Obviously, this would make $\sigma_{i,j}$ dependent on $c_i$ and increase the complexity of the posterior computation. As our focus is on clustering $n$ data points (instead of prediction), we choose our simple empirical prior and justify it by showing its clustering consistency in Section 4.

3 Posterior computation

To facilitate posterior computation, we now introduce a much simpler representation for $\mathcal{F}$. Consider an auxiliary node 0 that connects to all roots $(1^*,\ldots,K^*)$, we obtain a single spanning tree with only one root at 0. We denote this augmented graph by $T = (V, E)$ with node set $V = \{0\} \cup V_1 \cup \ldots \cup V_K$ and edges $E = \{(0,1^*),\ldots,(0,K^*)\} \cup E_1 \cup \ldots \cup E_K$. By sampling this augmented spanning tree, we effectively avoid challenges such as updating the choice for root nodes, changing $K$, etc.

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

$$
\Pi(T, \tilde{\sigma}_1, \ldots, \tilde{\sigma}_n \mid y_1, \ldots, y_n) \propto \exp\{\text{tr}(SA_T)/2\} (\lambda^K / K!) \prod_{i=1}^n \Pi_0(\tilde{\sigma}_i).
$$

where $K = |\{k : n_k \geq 1\}|$ is the number of non-empty clusters.

To update the enclosing $T$, we are inspired by the cut-and-reconnect algorithm for sampling a spanning tree (Duan and Dunson (2021), arXiv preprint 2106.16120). To briefly summarize their algorithm, for a spanning tree $T$, cutting an edge $(i, j) \in T$ will create two disjoint subgraphs $G_a = (V_a, E_a)$ and $G_b = (V_b, E_b)$, with each of $G_a$ and $G_b$ being connected; then one can sample a new edge $(i', j')$ using the discrete distribution supported at $\{(i', j') : i' \in V_a, j' \in V_b\}$, which leads to an updated $T$. This algorithm retains the combinatorial constraints for a tree, while enjoying excellent mixing performance. On the other hand, for the forest model, a new challenge is how to efficiently track the changes in the clustering membership and cluster size.

To solve this problem, we further extend the above into a “detach-and-reattach” algorithm. As illustrated in Figure 3, in the “detach” step, cutting an edge will detach a subgraph $G_b$ from the rest of $T$ containing node 0 — it is not hard to see that all nodes in $G_b$ must be in the same cluster. In the “reattach” step, if the new edge attaches $G_b$ to a node $i \neq 0$, then all the memberships in $G_b$ will be updated to $c_i$, which corresponds to growing an existing cluster; while if it attaches $G_b$ to node 0, then we effectively add a new cluster consisting of nodes in $G_b$. In our implementation code (provided in supplementary materials), we use the R package “igraph” to find out the component nodes $V_a$ and $V_b$.

For the associated probability, using $K^*$ as the number of clusters within $G_a$, we can see that sampling $(i, j)$ with $i \in V_a$ and $j \in V_b$ has:

$$
\text{pr}\{(i, j) \mid \} = \frac{\lambda^{(K^* + 1)(i = 0)}}{\sum_{(i', j')} \lambda^{(K^* + 1)(i' = 0)}} \frac{1}{\lambda^{(K^* + 1)(i = 0)}} \exp(S_{i,j})
$$

Given the tree $T$, we can update each scale $\text{pr}(\tilde{\sigma}_i = d_h \mid \cdot) \propto w_h(\prod_{j \geq 1: (i,j) \in T} d_h^{-p/2} \times \exp\{-\sum_{j \geq 1: (i,j) \in T} \|y_i - y_j\|_2^2 / (2\tilde{\sigma}_j d_h)\}$; for the hyper-parameters, $(w_1, \ldots, w_H) \sim \text{Dir}(0.1 + \sum 1(\tilde{\sigma}_i = d_1), \ldots, 0.1 + \sum 1(\tilde{\sigma}_i = d_H))$, and for updating $d_h$, we use a scaled sigmoid reparameterization for $d_h = \text{sigmoid}(\tilde{d}_h)(\max_i, \min_j \tilde{\sigma}_i \|y_i - y_j\|_2 / \sqrt{p})$ and random walk Metropolis to update each $\tilde{d}_h \in \mathbb{R}$.
(a) Starting from two clusters — green and blue, cutting an edge (cyan) detaches a subgraph $G_b$ from the bulk containing the node 0.

(b) Reattaching $G_b$ to one of nodes $i \in G_a$ and $i \neq 0$ corresponds to growing an existing (green) cluster.

(c) Reattaching $G_b$ to the node 0 corresponds to adding a new (red) cluster.

Figure 3: Illustration of the “detach-and-reattach” algorithm to update the clustering based on the forest model. This algorithm gives the discrete probability for sampling an edge from $|V_a| \times |V_b|$ possible candidates.

4 Properties

4.1 Connection to the Spectral Clustering

We now discuss the connection between our method and the spectral clustering algorithm. To provide some background, we first review its technical details (Von Luxburg et al., 2008), using our defined notations.

The spectral clustering algorithm takes a similarity matrix as input, in our case, each conditional density as a similarity $A_{i,j} = \exp(S_{i,j})$ for $i \neq j$, and $A_{i,i} = 0$. Then it calculates the Laplacian matrix $L = D - A$ with $D$ the diagonal matrix with the $i$th diagonal element $D_i = \sum_j A_{i,j}$, and further obtains the normalized Laplacian $N = D^{-1/2}LD^{-1/2}$. The algorithm gets its name because it takes the spectral decomposition of $N = \sum_j \gamma_j \phi_j \phi_j^T$, with $0 = \gamma_1 \leq \gamma_2 \leq \ldots \leq \gamma_n$ the eigenvalues with each $\phi_j$ the corresponding eigenvector, and lastly partitions the rows of the $n \times K$ matrix $(\phi_1 \ldots \phi_K)$ into $K$ clusters, using some heuristic algorithms such as the K-means. To clarify, there are multiple variants of this algorithm, and we focus on the one using the normalized Laplacian $N$ as it leads to consistency in the sense that those $(\phi_1 \ldots \phi_K)$ converge to the eigenvectors of the population version (Von Luxburg et al., 2008).

Now we examine another matrix in our method, and show an asymptotic equivalence in the eigenvectors with a rapid convergence rate, under mild conditions. Suppose we consider a uniform prior for $F$, equivalently $\text{pr}(T = T_1) = (n + 1)^{-(n-1)}$ for all $T_1 \in T$ with $T$ the space of spanning tree for $(n + 1)$ nodes (ignoring the Poisson-style regularization on $K$); further, the leaf density $f(y_i \mid y_j)$ can be any density satisfying the symmetry $f(y_i \mid y_j, \theta) = f(y_j \mid y_i, \theta)$, $r(y_i; \theta)$ can be any density satisfying $r(y_i; \theta) > 0$.

An important quantity in clustering is the co-assignment probability $\text{pr}(c_i = c_j \mid y)$. Note that $(i, j) \in T$ implies $c_i = c_j$; therefore, considering $T$ as a latent variable, the following marginal connecting probability gives an lower bound estimate for the co-assignment $\text{pr}(c_i = c_j \mid y)$,

$$\text{pr}\{(i, j) \in T \mid y\} = \sum_{T_1 \in T} 1\{(i, j) \in T_1\} \text{pr}(T = T_1 \mid y).$$

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

$$\sum_{\mathcal{T} \in T} \prod_{(i,j) \in \mathcal{T}} \exp(S_{i,j}) = (n + 1)^{-1} \prod_{h=2}^{n+1} \lambda_{h}(L).$$
where $L$ is the Laplacian matrix as described above for the nodes $0, 1, \ldots, n$; $\lambda_{(h)}$ denotes the $h$th smallest eigenvalue. Differentiating its logarithmic transform with respect to $S_{i,j}$,$$
abla S_{i,j} \exp(S_{i,j}) = \frac{\partial}{\partial S_{i,j}} \sum_{\mathcal{T} \in \mathcal{T}} \prod_{(i',j') \in \mathcal{T}} \exp(S_{i',j'}) = \frac{1}{\xi_K - \xi_{K+1}} \max_{i,j} \left\{ (1 + \epsilon)(D_i^{-1/2} - D_j^{-1/2})^2 A_{i,j} \right\},$$we found the convergence to be rather fast, with the two sets of leading eigenvectors becoming almost an approximate posterior mode for $(\Psi_{1:i} \phi_{1:i})$. And we can compare with the $K$ leading eigenvectors of $(-N) \in \mathbb{R}^{n \times n}$, $\phi_1, \ldots, \phi_K$. Using $\Psi_{1:i} \phi_{1:i}$ 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 $\epsilon > 0$, $$\|\Psi_{1:i} - \phi_{1:i} R\|_F \leq \frac{40\sqrt{K(n+1)}}{\xi_K - 2n+1} \max_{i,j} \left\{ (1 + \epsilon)(D_i^{-1/2} - D_j^{-1/2})^2 A_{i,j} \right\},$$with probability at least $1 - \exp(-n)$.

**Remark 4** To make the right hand side go to zero, a sufficient condition is to have all $A_{i,j} / (\sum_{j'=1}^n A_{i,j'}) = O(n^{-\kappa})$ with $\kappa > 1/2$. We provide the detailed definition of the bound constant $\epsilon$ in the supplementary materials.

Therefore, under mild condition, as $n \to \infty$, the two sets of leading eigenvectors coincide. This means that the spectral clustering algorithm that clusters the $K$-dimensional vectors $(\phi_{1:i}, \phi_{2:i}, \ldots, \phi_{K:i})$ for $i = 1, \ldots, n$ (while ignoring $i = 0$) into $K$ groups, gives an approximate solution to the following optimization problem:

$$\max_{(c_1, \ldots, c_n) \in P_K} \sum_{i=2}^n \sum_{j=1}^{n-1} 1(c_i = c_j) \Pr\{(i, j) \in \mathcal{T} | Y\},$$

where $P_K$ denotes the label space corresponding to all $K$-partitions of $(1, \ldots, n)$. Clearly, the solution gives an approximate posterior mode for $(c_1, \ldots, c_n)$ conditioning on having $K$ component trees. Empirically, we found the convergence to be rather fast, with the two sets of leading eigenvectors becoming almost indistinguishable starting around $n \geq 50$.

**Remark 5** Our discovery also means that a simple and often-used heuristic in fact also approximates the spectral clustering algorithm for obtaining $K$ clusters (which has a complexity of $O(n^2)$ due to the use of spectral decomposition) — finding the minimum spanning tree for $n$ data points (that minimizes $\sum_{(i,j) \in \mathcal{T}} \|y_i - y_j\|^2$) and cutting the edges associated with the $(K-1)$ largest distances. This has a computing complexity of $O(p^{-1} n^2 \log n + \log n \log n)$ with $p$ processors, based on the parallel Borůvka’s algorithm (Bader and Cong, 2006).

### 4.2 Consistent Clustering of Separable Sets

We show that clustering consistency is possible, under some separability assumptions. 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 et al., 2019).

To formalize the above statements, we denote the true cluster label used for generating $y_k$ by $c_0^k$ (subject to label permutation among clusters), and we define the enclosing region for all possible $y_k : c_0^k = k$ as $R_k^0$ for $k = 1, \ldots, K_0$ for some true finite $K_0$. And we refer to $R^0 = (R_1^0, \ldots, R_K^0)$ as the “null partition”. By separability, we mean the scenario that $(R_0^1, \ldots, R_K^0)$ 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 $\mathcal{T}$ (and equivalently, $\mathcal{F}$). 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:
Next, we relax the condition by having a

(A3') Let the root density satisfy \( \tilde{\sigma} \)

(A2) (Minimum separation) \( \inf_{x \in R^d_1, y \in R^d_1} \| x - y \|_2 > M_n \), 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 \( y \in X \).

Remark 6 Our assumption (A1) is based on our empirical uniform prior for \( \sigma^{0,n} \) using the minimum distances. Assumption (A2) coupled with (A1) suggests that the minimum separation is allowed to decay with sample size but at a rate slower than \( \sigma^{0,n} \).

Under the null partition, \( \Pi(T | y) \) is maximized at \( T = T_{MST,R_0} \), which contains \( K_0 \) trees with each \( T_k \) being the minimum spanning tree (denoted by subscript “MST”) within region \( R_k \). Similarly, for any alternative \( R \), \( \Pi(T | y) \) is maximized at the \( T = T_{MST,R} \).

Theorem 2 Under \( (A1,A2,A3) \), we have \( \frac{\Pi(T_{MST,R_0} | Y)}{\Pi(T_{MST,R_0} | Y)} \to 0 \) almost surely, unless \( R_0 \subseteq R_{\xi(i)} \) for some permutation map \( \xi(\cdot) \).

Number of cluster 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') Let the root density satisfy \( \tilde{m}_1 e^{-M/2\sigma^{0,n}} \leq r(y) \leq \tilde{m}_2 e^{-M/2\sigma^{0,n}} \) for some \( \tilde{m}_1 < \tilde{m}_2 \).

In this assumption, we essentially assume the root distribution to be increasingly “flatter” with \( n \). Then we have the following results.

Theorem 3 1) If \( K < K_0 \), under the assumptions \( (A1,A2,A3) \), we have \( \frac{\Pi(T_{MST,R_0} | Y)}{\Pi(T_{MST,R_0} | Y)} \to 0 \) almost surely.

2) If \( K > K_0 \), under the assumptions \( (A1,A2,A3') \), we have \( \frac{\Pi(T_{MST,R_0} | Y)}{\Pi(T_{MST,R_0} | Y)} \to 0 \) almost surely.

5 Numerical Experiments

5.1 Convergence of Eigenvectors

We now use simulations to illustrate the closeness between the \( K \) leading eigenvectors of the marginal connecting probability matrix \( M \) and the ones of the normalized Laplacian \( N \). It is important to clarify that such closeness does not depend on how the data are generated. Therefore, 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 \( I_2 \), then we fit our forest model, and estimate \( \sigma_{i,j} \)'s using posterior mean. Based on the posterior mean of \( \sigma_{i,j} \), we compute \( M \) and \( N \), and then compute distances between their leading eigenvectors \( \| \Psi_{1,K} - \Phi_{1,K} R \| \); for \( R \), we use a diagonal matrix of either 1 or \(-1\), corresponding to a simple sign change in each column.

We conduct such experiments under different sample sizes \( n \) ranging from 10 to 200; for each \( n \), we repeat experiments for 30 times. As our theory requires a spectral gap \( \xi_K - \xi_{K+1} \) not too close to zero, we choose to compare the top \( K = 5 \) eigenvectors.

As shown in the boxplot of Figure 4(a), the distance between two sets of eigenvectors quickly drops to near zero, for \( n \geq 50 \). To show more details, we plot leading 3 eigenvectors from one simulation at \( n = 60 \). As shown in Figure 4(b), two sets of eigenvectors are almost indistinguishable.
(a) The difference between eigenvectors converges to zero rapidly as \( n \) increases.

(b) Comparing the eigenvectors of a marginal connecting probability matrix \( M \) and the ones of normalized Laplacian \( N \).

Figure 4: Numeric experiments showing that the top few eigenvectors of a marginal connecting probability matrix \( M \) and the ones of normalized Laplacian become almost indistinguishable starting from \( n = 50 \).

5.2 Clustering Manifold Data

To illustrate the capability of uncertainty quantification, we carry out clustering tasks on those manifold data commonly used for benchmarking clustering algorithms. For conciseness, we present one example of three-ring manifolds in the main text, and present the other examples in the supplementary materials.

Figure 5: Illustration of uncertainty quantification on the clustering of near-manifold data, three circles. The posterior mode of the forest model recovers the ground truth, while the posterior distribution characterizes the uncertainty, such as the co-assignment probability \( \text{pr}(c_i = c_j | y) \).

In this simulation, we start with 300 points drawn 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 cases, one with noises from \( \text{No}(0, 0.05^2 I_2) \), and one with noises \( \text{No}(0, 0.1^2 I_2) \).

As shown in Figure 5, when these data are well separated (panel a, showing posterior mode), there is very little uncertainty on the clustering (panel b), with the posterior co-assignment \( \text{pr}(c_i = c_j | y) \) close to zero for any two data points near different rings. As noises increase, these data become more difficult to separate (although the posterior mode still recovers the ground truth, shown in panel c). There is a considerable amount of uncertainty for those red and blue points: these two sets of points are sometimes assigned into one cluster, leading to high co-assignment \( \text{pr}(c_i = c_j | y) \) between red and blue points (panel d).
6 Application: Hierarchical Modeling of Multi-subject Functional Magnetic Resonance Imaging Data

Besides uncertainty quantification, one advantage of having a generative model is that we can now develop model-based extensions for spectral clustering algorithms, for example, including the forest model in a hierarchical model.

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-stage functional magnetic resonance imaging scan data, collected from \( S = 144 \) subjects at different diseased stages of Alzheimer’s disease. Each subject has scans over \( n = 116 \) regions of interest and over \( T = 119 \) time points. We denote an observation for the \( s \)th subject in the \( i \)th region by \( y_i^{(s)} \in \mathbb{R}^T \).

Our goal is to find and distinguish between distinct clustering patterns for these subjects, while allowing some subjects to share same clustering of brain regions. To achieve this purpose, we assign each subject with an individual forest \( T^{(s)} \), while imposing a discrete prior for \( T^{(s)} \) over \( s = 1, \ldots, S \).

\[
T^{(s)} \sim \sum_{k=1}^{\infty} v_k \delta_{T_0^{(k)}}(\cdot), \quad \{y_1^{(s)}, \ldots, y_n^{(s)}\} \sim \text{Forest Model}(T^{(s)}).
\]

In the above, since we do not know the number of clustering patterns, we assume it to be unbounded from above and use a stick-breaking parameterization, \( v_k \sim \text{Beta}(1, 0.1) \), \( v_1 = \nu_1 \), and \( v_k = \nu_k \prod_{l<k}(1 - \nu_l) \) for \( k \geq 2 \); and we model each \( T_0^{(k)} \) as drawn from a base measure uniformly in the space of spanning trees with \( (n + 1) \) nodes.

We ran the Markov chain Monte Carlo for 5,000 iterations and discarded the first 2,500 as burn-in. Among the posterior samples, most correspond to having three distinct values in \( \{T^{(s)}\}_{s=1}^S \). In Figure 6 we plot three forests at the empirical posterior mode. By grouping subjects \( s \) and \( s' \) with \( T^{(s)} = T^{(s')} \) together, we obtain three-group assignments for these subjects as well. By comparison, the first one (Panel a) has nodes more connected in 6 clusters (along with a few isolated data points), whereas the other two (Panels b and c) have more small clusters that make the brain regions more fragmented. Using a validating external variable about the disease status of each subject, we find our result to be clinically interpretable: the first clustering pattern mostly corresponds to healthy subjects, while the second and third do to subjects that are in stages of Alzheimer’s disease (Panel d). Combining the last two estimated groups, we obtain a match rate of 78% when compared against the healthy/diseased labels. To provide a baseline, a simple K-means on the raw data with \( K = 2 \) yields a match rate of only 56%.

We find that the across hemisphere connections from Panel a are mostly missing in Panels b and c, supporting the fact that inter-hemispheric connectivity decreases among the diseased population (Qiu et al., 2016). Additionally, connections associated with the temporal lobe, occipital gyrus, angular gyrus, fusiform, cingulate anterior, and frontal cortex from Panel a are also found to be absent in the other two. These regions are primarily responsible for different cognitive tasks, memory management, and are known to be affected by Alzheimer’s disease (Galton et al., 2001; Yun et al., 2015; Cai et al., 2015).
Figure 6: Joint modeling of the brain regions for multiple subjects: each subject is assigned to a latent group label, with each group corresponding to one clustering pattern parameterized by a forest graph (shown in Panels a–c, colors representing different clusters, and white representing singletons). The estimated group labels at the posterior mode are very similar to the observed labels of diseased or healthy subjects, as a validation variable (Panel d).

Acknowledgement

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. 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.

Proof of Theorems

Proof of Theorem 1

Proof 1 For ease of notation, we use $p = (n + 1)$.

1. Obtain closed-form of the marginal connecting probability.
Since $L$ correspond to a connected graph with weight $A_{i,j} = \exp(W_{i,j})$ for $i \neq j$ and $A_{i,i} = 0$, hence only has one eigenvalue equal to 0 and with eigenvector $\bar{1}/\sqrt{p}$, therefore we have:

$$p^{-1} \prod_{i=2}^{n+1} \lambda_{i}(L) = |L + J/p^2|,$$

where $J = \bar{1}\bar{1}^T$. Let $\tilde{L} = L + J/p^2$, differentiating $\log |\tilde{L}|$ with respect to $W_{i,j}$ yields:

$$M_{i,j} = (\Omega_{i,i} + \Omega_{j,j} - 2\Omega_{i,j})A_{i,j},$$

where $\Omega$ is due to $\Omega_{i,i} + \Omega_{j,j} - 2\Omega_{i,j} = 0$ if $1(j \neq i) = 0$, hence $1(k \neq i)$ can be omitted; $\bar{b}_{i,j}$ is a binary vector with the $i$th element 1 and the $k$th element $-1$, and all other elements 0.

Let $\alpha_{i,j} := D_i^{1/2}\bar{b}_{i,j}(L + J/p^2)^{-1}\bar{b}_{i,j}D_j^{1/2}$ for $i \neq j$. Since $N_{i,i} = 1$, and $x(I - N)$ has the same eigenvectors as $N$ for any scalar $x > 0$, we see that $M = -\alpha \circ (I - N)$ is an element-wise perturbation of $x(I - N)$. Therefore, our next task is to show $\alpha$ is close to a simple $xJ$ for some $x > 0$.

**2. Obtain $M$ as a perturbation form.**

Let $L = \sum_{i=1}^{p} \lambda_i \psi_i \psi_i^T$ be the eigendecomposition of $L$, and $N = D^{-1/2}LD^{-1/2}$. Note that,

$$M_{i,j} = (\Omega_{i,i} + \Omega_{j,j} - 2\Omega_{i,j})A_{i,j}$$

$$= (\Omega_{i,i} + \Omega_{j,j} - 2\Omega_{i,j})(-L_{i,j}1(j \neq i))$$

$$= (b_{i,j}^T(L + J/p^2)^{-1}b_{i,j})(-L_{i,j})$$

$$= D_i^{1/2}\bar{b}_{i,j}^T(L + J/p^2)^{-1}\bar{b}_{i,j}D_j^{1/2}(-N_{i,j})$$

where $(a)$ is due to $\Omega_{i,i} + \Omega_{j,j} - 2\Omega_{i,j} = 0$ if $1(j \neq i) = 0$, hence $1(k \neq i)$ can be omitted; $\bar{b}_{i,j}$ is a binary vector with the $i$th element 1 and the $k$th element $-1$, and all other elements 0.

Let $\alpha_{i,j} := D_i^{1/2}\bar{b}_{i,j}(L + J/p^2)^{-1}\bar{b}_{i,j}D_j^{1/2}$ for $i \neq j$. Since $N_{i,i} = 1$, and $x(I - N)$ has the same eigenvectors as $N$ for any scalar $x > 0$, we see that $M = -\alpha \circ (I - N)$ is an element-wise perturbation of $x(I - N)$. Therefore, our next task is to show $\alpha$ is close to a simple $xJ$ for some $x > 0$.

**3. Bound the difference between the $K$ leading eigenvectors.**

Slightly changing the above,

$$\alpha_{i,j} = D_i^{1/2}\bar{b}_{i,j}(L + J/p^2)^{-1}\bar{b}_{i,j}D_j^{1/2}$$

$$= D_i^{1/2}\bar{b}_{i,j}^T D^{-1/2}(N + D^{-1/2}JD^{-1/2}/p^2)^{-1}D^{-1/2}\bar{b}_{i,j}D_j^{1/2}.$$

Using $N = I - D^{-1/2}AD^{-1/2}$, we have

$$(N + D^{-1/2}JD^{-1/2}/p^2)^{-1} = \{I - D^{-1/2}(A - J/p^2)D^{-1/2}\}^{-1}$$

$$= I + \sum_{k=1}^{\infty} \{D^{-1/2}(A - J/p^2)D^{-1/2}\}^k$$

$$= I + E.$$

where $(a)$ uses the Neumann expansion, as $E$ is not divergent:

$$E = D^{1/2}(L + J/p^2)^{-1}D^{1/2} - I = D^{1/2}(\sum_{i=2}^{p} \lambda_i^{-1}\psi_i\psi_i^T + J)D^{1/2} - I,$$

as $\lambda_2 > 0$, $E$ is bounded element-wise for any $D$ with finite-value elements. Further, note that $D_i^{-1/2}(A_{i,j} - 1/p^2)D_j^{-1/2} \to 0$ and monotonically decreasing for fixed $A_{i,j}$ and increasing $D_i$ or $D_j$; hence $E$ is always bounded elementwise even as all $D_i \to \infty$. We denote the bound constant by $\max_{i,j} |E_{i,j}| \leq \epsilon$.

Combining the above,

$$\alpha_{i,j} = D_i^{1/2}D_j^{1/2}\bar{b}_{i,j}\left\{D^{-1} + D^{-1/2}ED^{-1/2}\right\}\bar{b}_{i,j}$$

$$= D_i^{1/2}D_j^{1/2}\left\{(D_i^{-1} + D_j^{-1}) + (D_i^{-1}E_{i,i} + D_j^{-1}E_{j,j} - 2D_i^{-1/2}D_j^{-1/2}E_{i,j})\right\}.$$
Now we can bound the difference between $M$ and $x(I - N)$ minimized over $x$:

$$\min_{x} \max_{i,j} |(x - xJ) \circ (I - N)\{i,j\}|$$

$$= \min_{x} \max_{i,j} |(x_i - x)(D_i^{-1/2}D_j^{-1/2}A_{i,j})|$$

$$= \min_{x} \max_{i,j} \left\{|(D_i^{-1} + D_j^{-1}) - xD_i^{-1/2}D_j^{-1/2} + (D_i^{-1}E_{i,i} + D_j^{-1}E_{j,j} - 2D_i^{-1/2}D_j^{-1/2}E_{i,j})\} A_{i,j}\right\}$$

$$\leq \min_{x} \max_{i,j} \left\{|(D_i^{-1} + D_j^{-1}) - xD_i^{-1/2}D_j^{-1/2} + (D_i^{-1}D_j^{-1} + 2D_i^{-1/2}D_j^{-1/2})\} A_{i,j}\right\}$$

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

where ($a$) chooses $x = 2(1 + \epsilon) + 2\epsilon$.

Using Theorem 2 from Yu et al. (2015), there exists an orthonormal matrix $R \in \mathbb{R}^{K \times K}$, such that,

$$\|\Psi_{1:K} - \phi_{1:K}R\|_F \leq 2^{3/2} \min\{\sqrt{K}\|\alpha - xJ\circ (I - N)\|_{op}, \|\alpha - xJ\circ (I - N)\|_F\}$$

$$\xi_K - \xi_{K+1}$$

for any $x > 0$.

Since $|\{(\alpha - xJ) \circ (I - N)\}_{i,j}|$ is upper-bounded hence is sub-Gaussian with bound parameter $\sigma_e = \max_{i,j}\{(1 + \epsilon)(D_i^{-1/2} - D_j^{-1/2})A_{i,j}\}$.

Using Theorem 1 of Duan, Michaelidis and Ding (2020 (arXiv preprint:1910.02471)), with probability $1 - \delta_t$,

$$\|\Psi_{1:K} - \phi_{1:K}R\|_F \leq \frac{2^{3/2}(\sqrt{Kp}\sigma_e)}{\xi_K - \xi_{K+1}} t,$$

where $\delta_t = \exp[-(t^2/64 - \log(5\sqrt{2}))^2]$. Taking $t = 14$, we have $(t^2/64 - \log(5\sqrt{2})) > 1$. Therefore, we have with probability at least 1 $- \exp(-p)$ $\{\text{which is greater than } 1 - \exp(-n)\},$

$$\|\Psi_{1:K} - \phi_{1:K}R\|_F \leq \frac{40\sqrt{Kp}\sigma_e}{\xi_K - \xi_{K+1}}.$$

**Proof of Theorem 2 and 3**

Let the conditional probability associated with Gaussian leaf density $f$ be $\text{pr}\{B(y_1, M_n/2) \mid y_1\} = m_n$, where $B(y_1, M_n/2)$ stands for an open ball of radius $M_n/2$ around $y_1$. If the true number of clusters is $K = K_0$, then $m_n^{n-K-1}$ is the probability that the distances $\{d_{\ell,n}\}$ in the minimum spanning tree under null are all below $M$. Specifically, let $E_n = \{d_{\ell,n} \leq M_n/2 : 1 \leq \ell \leq n - K - 1\}$. Then $\text{pr}(E_n) = m_n^{n-K-1}$.

With $x_i \sim \text{No}(0, \sigma^{0,n})$, we have $m_n = \text{pr}(\sum_{i=1}^{n} x_i^2 < \frac{M_n^2}{2\sigma^{0,n}}) = 1 - \frac{\Gamma(p/2, \frac{M_n^2}{2\sigma^{0,n}})}{\Gamma(p/2)}$, where $\Gamma(\cdot, \cdot)$ stand for the upper incomplete gamma function using cumulative distribution function of $\chi^2$ distribution. Since $p$ belongs to the set of natural numbers, we have $\Gamma(p/2, x) < C_1 x^{p/2} e^{-x}$ except for $p = 1.5$ and any $x > 0$ with some constant $C_1$ which depends on $p$ (Pinelis 2020, arXiv preprint 2005.06384). However, for large $x$, we have $\Gamma(p/2, x) < C_1 x^{p/2} e^{-x}$ even for $p = 1.5$. We then have $m_n > 1 - \frac{C_2}{\Gamma(p/2)} (\log n)^{p/2-1} e^{-\tilde{m}_0 \log n}$ where $C_2 = C_1(\tilde{m}_0)^{p/2-1}$. Since $\tilde{m}_0 > (p/2 + 2)$, we have $m_n^{n-K-1} > 1 - (n - K - 1) \frac{C_2}{\Gamma(p/2)} (\log n)^{p/2-1} \to 1$ as $n \to \infty$ as $\log n$ goes to 0. Hence $\text{pr}(E_n) \to 1$.

We further have, $\sum_{n \geq K} [1 - (1 - \frac{C_2}{\Gamma(p/2)} (\log n)^{p/2-1})^{n-K-1}] < \sum_{n} n \frac{C_2}{\Gamma(p/2)} (\log n)^{p/2-1}$. Thus for $\tilde{m}_0 > 2 + p/2$, we have $\sum_{n} [1 - (1 - \frac{C_2}{\Gamma(p/2)} (\log n)^{p/2-1})^{n-K-1}] < \infty$. Then, by the Borel-Cantelli Lemma, we also have almost sure convergence of this event.

We now show for $y \in E_n$, the ratio of the maximum posterior probability assigned to a “non-true” clustering arrangement to the posterior probability assigned to the “true” clustering arrangement converges to zero.

**Proof of Theorem 2**
Note that for any \( \sigma \) and a given \( R \), the posterior \( \Pi(\mathcal{T}_{\text{MST},R}, \sigma | y) \) is maximized at the \( T \), which is a combination of minimum spanning trees constructed within the regions \( R_k \)’s. Thus,

\[
\frac{\Pi(\mathcal{T}_{\text{MST},R}, \sigma^{0,n} | y)}{\Pi(\mathcal{T}_{\text{MST},R^0}, \sigma^{0,n} | y)} \leq \left( \frac{c_2}{\epsilon_1} \right)^K \exp \left( -\sum_{\ell=1}^{n-K_0-1} d_{\ell,n}^2/(2\sigma^{0,n}) + \sum_{\ell=1}^{n-K-1} (d_{\ell,n}^0)^2/(2\sigma^{0,n}) \right),
\]

where \( \sum_{\ell=1}^{n-K-1} d_{\ell,n}^2 \) is the total squared norm distance on the minimum spanning tree under the partition regions \( R \) excluding the edges with the root node and \( \sum_{\ell=1}^{n-K_0-1} (d_{\ell,n}^0)^2 \) is the same under \( \mathcal{T}_{\text{MST},R^0} \). The above is because that based on the Prim’s algorithm (Prim, 1957), the minimum spanning tree is equal to the result of sequential growing a tree starting from one node, each time by adding an edge (along with a node) with the shortest distance between one node in the existing tree and one of remaining nodes not yet in the tree. Clearly, at each step, the edge choice is unaffected when changing distance from \( d \) to \( d^2 \); therefore, the minimum spanning trees based on sum of \( d_{\ell,n}^2 \) and sum of \( d_{\ell,n} \) are the same.

Since, \( \inf_{x \in R^0, y \in R^0} \| x - y \|_2 > M_n \), for all \( i \neq j \), for at least one \( \ell \), we must have \( d_{\ell,n} > M \). Due to the above result, with probability at least \( m_n^0 \), we have \( \sum_{\ell=1}^{n-K-1} (d_{\ell,n}^0)^2 > \sum_{\ell=1}^{n-K_0-1} (d_{\ell,n}^0)^2 + M_n^2/4 \), which implies \( \frac{\Pi(\mathcal{T}_{\text{MST},R}, \sigma^{0,n} | y)}{\Pi(\mathcal{T}_{\text{MST},R^0}, \sigma^{0,n} | y)} < n^{-\tilde{m}_n} \). And we further have that \( m_n^0 \to 1 \) as \( n \to 1 \).

**Proof of Theorem 3**

First, we consider that the alternative partitioning has a lower number of clusters than the null. Let that be \( K \), which is less than \( K_0 \). Then we have

\[
\frac{\Pi(\mathcal{T}_{\text{MST},R}, \sigma^{0,n} | y)}{\Pi(\mathcal{T}_{\text{MST},R^0}, \sigma^{0,n} | y)} \leq \lambda^{K-K_0} \frac{K_0!}{K!} \left( \frac{c_2}{\epsilon_1} \right)^K \frac{\sigma^{0,n} (K_0-K)/2}{\epsilon_1^{K_0-K}} \exp \left( -\sum_{\ell=1}^{n-K_0-1} d_{\ell,n}^2/(2\sigma^{0,n}) + \sum_{\ell=1}^{n-K_0-1} (d_{\ell,n}^0)^2/(2\sigma^{0,n}) \right),
\]

We again must have \( \sum_{\ell=1}^{n-K_0-1} (d_{\ell,n}^0)^2 > \sum_{\ell=1}^{n-K_0-1} (d_{\ell,n}^0)^2 + M_n^2/4 \) with probability \( m_n^0 \to 1 \) as the alternative partitioning will have edges with length greater than \( M_n \).

Next we show the above when the alternative partitioning has larger number of clusters than null. Specifically say \( K > K_0 \). In this case, we replace A3 and vary the conditions on \( r(y) \) with \( n \).

Then we have

\[
\frac{\Pi(\mathcal{T}_{\text{MST},R}, \sigma^{0,n} | y)}{\Pi(\mathcal{T}_{\text{MST},R^0}, \sigma^{0,n} | y)} \leq \lambda^{K-K_0} \frac{K_0!}{K_2!} \left( \frac{c_2}{\epsilon_1} \right)^K c_2^{K-K_0} (\sigma^{0,n})^{(K_0-K)/2} \times \exp \left( -(K-K_0)M_n^2/(2\sigma^{0,n}) - \sum_{\ell=1}^{n-K_0-1} d_{\ell,n}^2/(2\sigma^{0,n}) + \sum_{\ell=1}^{n-K_0-1} (d_{\ell,n}^0)^2/(2\sigma^{0,n}) \right),
\]

Again, for any \( K > K_0 \), the above ratio goes to zero as \( n \to \infty \) we have \( 1/(\sigma^{0,n}) \to 0 \) with probability at least \( m_n^0 \to 1 \).

**References**

Bader, D. A. and G. Cong (2006). Fast Shared-Memory Algorithms For Computing the Minimum Spanning Forest of Sparse Graphs. *Journal of Parallel and Distributed Computing* 66(11), 1366–1378.

Belkin, M. and P. Niyogi (2002). Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. In T. Dietterich, S. Becker, and Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems*, Volume 14. MIT Press.

15
Belkin, M. and P. Niyogi (2004). Semi-Supervised Learning on Riemannian Manifolds. *Machine learning 56*(1), 209–239.

Belkin, M. and P. Niyogi (2008). Towards a Theoretical Foundation for Laplacian-Based Manifold Methods. *Journal of Computer and System Sciences 74*(8), 1289–1308.

Blackwell, D. and J. B. MacQueen (1973). Ferguson Distributions via Polya Urn Schemes. *The Annals of Statistics 1*(2), 353–355.

Blei, D. M. and P. I. Frazier (2011). Distance Dependent Chinese Restaurant Processes. *Journal of Machine Learning Research 12*(8).

Cai, S., T. Chong, Y. Zhang, J. Li, K. M. von Deneen, J. Ren, M. Dong, L. Huang, and A. D. N. Initiative (2015). Altered Functional Connectivity of Fusiform Gyrus in Subjects With Amnestic Mild Cognitive Impairment: A Resting-State fMRI Study. *Frontiers in Human Neuroscience 9*, 471.

Cao, X., K. Khare, and M. Ghosh (2019). Posterior Graph Selection and Estimation Consistency for High-Dimensional Bayesian DAG Models. *The Annals of Statistics 47*(1), 319–348.

Chaiken, S. and D. J. Kleitman (1978). Matrix Tree Theorems. *Journal of Combinatorial Theory, Series A 24*(3), 377–381.

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

Ester, M., H.-P. Kriegel, J. Sander, and X. Xu (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, KDD’96*, pp. 226–231. AAAI Press.

Fraley, C. and A. E. Raftery (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association 97*(458), 611–631.

Frey, B. J. and D. Dueck (2007). Clustering by Passing Messages Between Data Points. *Science 315*(5814), 972–976.

Galton, C., B. Gomez-Anson, N. Antoun, P. Scheltens, K. Patterson, M. Graves, B. Sahakian, and J. Hodges (2001). Temporal Lobe Rating Scale: Application to Alzheimer’s Disease and Frontotemporal Dementia. *Journal of Neurology, Neurosurgery & Psychiatry 70*(2), 165–173.

Geng, J., A. Bhattacharya, and D. Pati (2019). Probabilistic Community Detection With Unknown Number of Communities. *Journal of the American Statistical Association 114*(526), 893–905.

Hartigan, J. A. (1981). Consistency of Single Linkage for High-Density Clusters. *Journal of the American Statistical Association 76*, 388–394.

Kosmidis, I. and D. Karlis (2016). Model-Based Clustering Using Copulas With Applications. *Statistics and Computing 26*(5), 1079–1099.

Lei, J. and A. Rinaldo (2015). Consistency of Spectral Clustering in Stochastic Block Models. *The Annals of Statistics 43*(1), 215–237.

Maharaj, E. A., P. D’Urso, and J. Caiado (2019). *Time Series Clustering and Classification*. Chapman and Hall/CRC.

McLachlan, G. J., D. Peel, and R. Bean (2003). Modelling High-Dimensional Data by Mixtures of Factor Analyzers. *Computational Statistics & Data Analysis 41*(3-4), 379–388.

Miller, J. W. (2019). An Elementary Derivation of the Chinese Restaurant Process From Sethuraman’s Stick-Breaking Process. *Statistics & Probability Letters 146*, 112–117.
Miller, J. W. and D. B. Dunson (2018). Robust Bayesian Inference via Coarsening. *Journal of the American Statistical Association* **114**(527), 1113–1125.

Müller, P. and F. Quintana (2010). Random Partition Models With Regression on Covariates. *Journal of Statistical Planning and Inference* **140**(10), 2801–2808.

Ng, S.-K., G. J. McLachlan, K. Wang, L. Ben-Tovim Jones, and S.-W. Ng (2006). A Mixture Model With Random-Effects Components for Clustering Correlated Gene-Expression Profiles. *Bioinformatics* **22**(14), 1745–1752.

Prim, R. C. (1957). Shortest Connection Networks and Some Generalizations. *The Bell System Technical Journal* **36**(6), 1389–1401.

Qiu, Y., S. Liu, S. Hilal, Y. M. Loke, M. K. Ikram, X. Xu, B. Y. Tan, N. Venketasubramanian, C. L.-H. Chen, and J. Zhou (2016). Inter-Hemispheric Functional Dysconnectivity Mediates the Association of Corpus Callosum Degeneration With Memory Impairment in AD and Amnestic MCI. *Scientific Reports* **6**(1), 1–12.

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

Rohe, K., S. Chatterjee, and B. Yu (2011). Spectral Clustering and the High-Dimensional Stochastic Blockmodel. *The Annals of Statistics* **39**(4), 1878–1915.

Socher, R., A. Maas, and C. Manning (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.

Steinwart, I. (2011). Adaptive Density Level Set Clustering. In S. M. Kakade and U. von Luxburg (Eds.), *Proceedings of the 24th Annual Conference on Learning Theory*, Volume 19 of *Proceedings of Machine Learning Research*, Budapest, Hungary, pp. 703–738. PMLR.

Vidal, R. (2011). Subspace Clustering. *IEEE Signal Processing Magazine* **28**(2), 52–68.

Von Luxburg, U. (2007). A Tutorial on Spectral Clustering. *Statistics and Computing* **17**(4), 395–416.

Von Luxburg, U., M. Belkin, and O. Bousquet (2008). Consistency of Spectral Clustering. *The Annals of Statistics*, 555–586.

Yu, Y., T. Wang, and R. J. Samworth (2015). A Useful Variant of the Davis–Kahan Theorem for Statisticians. *Biometrika* **102**(2), 315–323.

Yun, H. J., K. Kwak, J.-M. Lee, and A. D. N. Initiative (2015). Multimodal Discrimination of Alzheimer’s Disease Based on Regional Cortical Atrophy and Hypometabolism. *PloS one* **10**(6), e0129250.