Journal of Data Science 0 (0), 1–21
??? 2024
DOI: 10.6339/24-JDS1125
Statistical Data Science

Multi-Dimensional Clustering Based on Restricted Distance-Dependent Mixture Dirichlet Process for Diffusion Tensor Imaging

Soyun Park[1], Jihnhee Yu[2,*], and Zohi Sternberg[3]

[1]*Global Biometrics and Data Science, Bristol Myers Squibb, Princeton, New Jersey, USA*
[2]*Department of Biostatistics, University at Buffalo, State University of New York, Buffalo, New York, USA*
[3]*Department of Neurology, University at Buffalo, State University of New York, Buffalo, New York, USA*

Abstract

Brain imaging research poses challenges due to the intricate structure of the brain and the absence of clearly discernible features in the images. In this study, we propose a technique for analyzing brain image data identifying crucial regions relevant to patients' conditions, specifically focusing on Diffusion Tensor Imaging data. Our method utilizes the Bayesian Dirichlet process prior incorporating generalized linear models, that enhances clustering performance while it benefits from the flexibility of accommodating varying numbers of clusters. Our approach improves the performance of identifying potential classes utilizing locational information by considering the proximity between locations as clustering constraints. We apply our technique to a dataset from Transforming Research and Clinical Knowledge in Traumatic Brain Injury study, aiming to identify important regions in the brain's gray matter, white matter, and overall brain tissue that differentiate between young and old age groups. Additionally, we explore a link between our discoveries and the existing outcomes in the field of brain network research.

Keywords *adjacency matrix; Bayesian Dirichlet process prior; brain imaging; clustering; pattern recognition*

1 Introduction

Diffusion tensor imaging (DTI) is a modality of magnetic resonance imaging (MRI) that has become increasingly popular as a diagnostic tool in the field of brain research (Soares et al., 2013; Parekh et al., 2015; ElNakieb et al., 2021). It measures the movement of water molecules in biological tissues, specifically in the white matter tracts of the brain, to gain insights into the brain's microstructures. DTI offers numerous advantages, including its non-invasiveness, high spatial resolution, sensitivity to changes in white matter, and the ability to be combined with other imaging modalities such as fMRI to study the relationship between brain structure and function. DTI data are highly versatile, with applications ranging from research to clinical diagnosis and treatment planning. In neuroscience research, DTI is used to investigate white matter tracts and connectivity in various neurological and psychiatric disorders. Clinically, it can help diagnose and monitor diseases such as multiple sclerosis, traumatic brain injury, and

*Corresponding author. Email: jinheeyu@buffalo.edu.

Received September 22, 2023; Accepted March 12, 2024

brain tumors (Kraus, 2023). Additionally, DTI data can assist clinicians in planning surgeries or radiation treatments for brain tumors by identifying the location of crucial white matter tracts. Overall, DTI is a powerful imaging tool that provides unique insights into the brain's structural connectivity and is widely useful in both research and clinical settings.

However, analyzing DTI or other types of brain imaging data is challenging due to the large volume and complex nature of the data, as well as the statistical demands of analyzing multiple tests. Interpreting this data requires a multidisciplinary approach that incorporates knowledge of both brain anatomy and statistical modeling. Oftentimes, there is no established ground truth for brain function or structure. Due to the challenges in comprehending the intricate connections and functions between different brain regions, the options available for analyzing brain data are limited (Jones and Cercignani, 2010; Jbabdi and Johansen-Berg, 2011; Schilling et al., 2019). To address this, it is of great importance to develop suitable methods and demonstrate the feasibility of data analysis, thereby effectively broadening the spectrum of choices accessible for analyzing DTI data. An ideal approach for DTI data analysis would involve the identification of significant potential structures throughout the entire brain. Ultimately, the intended method should assist clinicians in the diagnosis and treatment of neurological disorders, benefiting patients.

Herein, we propose a novel method called Restricted Distance-dependent Mixtures of Dirichlet Process (RDMDP) within the framework of the Dirichlet process mixture model. RDMDP is proposed to address the need for identifying clustering related features, particularly in domains such as healthcare image and geographic analysis. Our approach is primarily inspired by the idea that brain regions sharing similar traits tend to be located near each other, reflecting the structural organization of functional brain regions (Saad and Mansinghka, 2018). An illustrative example provided in Wehrhahn et al. (2020) demonstrates disease clustering, which indicates an expected emergence of a particular disease with a heightened likelihood of happening in close proximity, both temporally and geographically. This serves as the foundation of our clustering approach and parallels the method proposed by Blei and Frazier (2011), which suggestes a flexible distribution framework for clustering that considers the dependency of elements when assessed based on a distance. Our method incorporates classification information and locational contiguity to provide a form of pseudo-supervised learning. This imposes some level of constraint on cluster modeling rather than allowing for completely unrestrained analysis. The goal is to establish brain mapping that links interrelated regions of the brain, taking into account both neighboring regions and the disease of interest. We aim to visually display this map in a manner that conveys a clinically meaningful summary of the structural and functional relationships within the brain. The proposed method adopts Bayesian nonparametric clustering (Orbanz and Teh, 2010; Wade and Ghahramani, 2018; Seymour, 2020; Masoero et al., 2021; Xian and Wade, 2022; Creswell et al., 2023; Daniel Loyal and Chen, 2023) with a regression component for disease mapping, under the assumption of a restricted Chinese Restaurant Process (CRP) (Fergusom, 1973; Escobar and West, 1995; Neal, 2000; Teh et al., 2006). Cluster assignments are empirically generated by Markov Chain Monte Carlo (MCMC), where a changeable random partition is allowed based on the posterior distribution over partitions, given the observed data and the prior distribution assumed under CRP (Pitman, 1995) based on the Dirichlet prior.

We note that the Dirichlet prior has played an important role in the development of Bayesian nonparametric methods and can be represented in various ways: stick-breaking process, Pólya urn scheme, and CRP. Specifically, the CRP serves as an analogy for a Dirichlet Process (DP) that assumes a customer enters a restaurant sequentially and chooses a table from an infinite number of tables. The CRP is a commonly employed prior in Bayesian nonparametric models and has been extended to include covariate, time, and space-dependent models, as well as hierar-

chical mixture models (MacEachern, 2000; Teh et al., 2006; Griffin and Steel, 2006; Duan et al., 2007). Numerous modifications and applications have been made to the original CRP, such as nested CRP (Blei et al., 2007), distance-dependent CRP (Blei and Frazier, 2011), region-based distance-dependent CRP (Ghosh et al., 2011), similarity-dependent CRP (Socher et al., 2011), temporarily-reweighted CRP (Saad and Mansinghka, 2018), powered CRP (Lu et al., 2018) and restricted CRP (RCRP) (Wehrhahn et al., 2020). Additionally, some imaging data applications have been explored (Baldassano et al., 2015; Ren et al., 2016).

Overall characteristics of these modifications do not maintain the property of exchangeability between subjects (Blei et al., 2007; Blei and Frazier, 2011). However, they offer advantages and are used in many practical applications since they allow for greater control over the clustering structure while imposing restrictions on the assignment of data points to clusters. These modifications can be useful when some prior knowledge about the data is available. While they may not be suitable for situations requiring conditional independence of components, they can provide a flexible and powerful clustering method when used appropriately (Ahmed and Xing, 2008; Ghosh et al., 2011; Blei and Frazier, 2011).

In adapting the CRP for DTI data analysis, we consider two key factors. First, certain areas in the brain may be highly associated with pathogenesis of specific disease or conditions, which can be identified through disease status and observable quantities in DTI. Cluster assignments may need to reflect these associations as they pertain to the structural patterns underlying the pathophysiology of the condition. Second, data points that share locational proximity are likely to form clusters together. Among the many variations of CRP, a modified version known as restricted CRP (RCRP) addresses this issue by using an adjacency matrix. The method enhances the coherence and consistency of the resulting cluster map for disease mapping purposes (Wehrhahn et al., 2020). By employing this strategy, we limit the number of latent clusters considered in the cluster assignment, using proximity information between data points.

In this article, we extend the regression-based CRP (Oganisian et al., 2021) to DTI analysis, aligning with Lan et al. (2021) and incorporating locational considerations. The proposed approach offers flexibility and efficient implementation via MCMC, making it highly effective for clustering and disease mapping applications (Seymour, 2020; Wehrhahn et al., 2020; Creswell et al., 2023).

While RCRP effectively tackles locational challenges, it lacks the capacity to account for relationships between observations and disease status within the model. In contrast, the regression-based CRP overlooks locational information in its classification. By leveraging the strengths of both methods, our method utilizes their respective advantages to enhance classification approach.

The following sections delineate the structure of the remainder of the paper. In Section 2, we introduce RDMDP, the Restricted Distance-dependent Mixture Dirichlet Process as a tool for identifying areas that reflect both phenotypical differences and the disease of interest. We detail the MCMC algorithm that implements the proposed RDMDP and introduce methods for summarizing partition data generated by MCMC iterations. In Section 3, we conduct a simulation study to demonstrate the performance of RDMDP in identifying interesting patterns across various images and compare it to existing clustering techniques. In Section 4, we apply our method to DTI data from TRACK-TBI, explaining data preparation and providing a detailed account of the data analysis. We also discuss the interpretation of the results, comparing partitions of brain regions based on gray matter, white matter and overall brain tissues. Finally, Section 5 contains our concluding remarks.

2 Method

2.1 Proposed Method: RDMDP

Let $D = (D_1, D_2, \ldots, D_n)$ be a set of data where $D_i = (Y_i, X_i)$ consists of Y_i for the binary response and X_i for covariates. We define a set of parameters, $\theta_i = (\beta_i, \mu_i, \psi_i)$, where β_i is a parameter to describe the relationship between Y_i and X_i, and μ_i and ψ_i are the mean and variance vectors describing the covariate vector $X_i = (X_{i1}, \ldots, X_{iL})^T$, where L is the number of covariates. The full conditional distribution of an individual data point given parameters is expressed as

$$P(D_i|\theta_i) = P(Y_i|X_i, \theta_i)P(X_i|\theta_i).$$

Specifically, we have

$$\begin{aligned} Y_i|X_i, \beta_i &\sim Bernoulli(\phi(X_i^T\beta_i)), \\ X_i|\mu_i, \psi_i &\sim p(x_i|\mu_i, \psi_i), \\ \theta_i|G &\sim G, \\ G|\alpha, G_0 &\sim \mathrm{RCRP}(\alpha, G_0, A), \end{aligned} \tag{1}$$

where ϕ is the inverse of a link function, p is the distribution of the covariate X_i, and G, with its distribution denoted as $\mathrm{RCRP}(\alpha, G_0, A)$, signifies the parameter distribution based on the RCRP prior G_0 and α. Here, α is a non-negative concentration parameter for the RCRP prior. The RCRP prior is restricted by an $n \times n$ adjacency matrix $A = (a_{i,i'})$, a first-order neighborhood matrix as we explain in the subsequent discussion within this section. Consequently, $\theta_i|G$ refers to the random distribution of the i-th cluster parameter θ_i given the distribution G, therefore, it is a distribution of distributions. The matrix's element $a_{i,i'}$ is an indicator function that shows whether regions i and i' share a boundary, i.e., $a_{i,i'} = 1$ if sharing a boundary, or 0 otherwise. The concentration parameter α controls the probability of assigning an observation to a new cluster.

Let K indicate the random number of clusters and $g = (g_1, \ldots, g_n)$ is a cluster assignment for n observations where g_i is the i-th observation's cluster membership. In a typical DP, given cluster memberships of all other observations and α, the i-th observation is assigned to an existing cluster $k \leqslant K$ with probability proportional to the number of observations in existing cluster k, or a new cluster with probability proportional to α

$$\pi(g_i = k|g_{-i}, \alpha) \propto \begin{cases} n_k(g_{-i}) & \text{if } k \leqslant K(g_{-i}) \\ \alpha & \text{if } k = K(g_{-i}) + 1 \end{cases}, \tag{2}$$

where $g_{-i} = (g_1, \ldots, g_{i-1}, g_{i+1}, \ldots, g_n)$ is the cluster assignments for all other observations, $n_k(g_{-i})$ is the number of cluster k in g_{-i}, and $K(g_{-i})$ is the total number of clusters in g_{-i}. The parameter θ_i of the i-th cluster follows a random distribution G from the Dirichlet process $\mathrm{DP}(\alpha, G_0)$, a distribution of distributions, which induces a cluster structure of the parameter space without assuming an exact number of clusters, estimating K directly from data. When α reduces to 0, g_i's are assigned to the same cluster since there is no probability of forming a new cluster, and if α goes to ∞, g_i's become all different clusters. The partition distribution is a distribution of cluster memberships $g = (g_1, g_2, \ldots, g_n)$. With regard to cluster memberships $g_i (i = 1, \ldots, n)$ and the random number of clusters $K(g)$ for all subjects, the

partition distribution is given by the probability $P(g, K(g)|\alpha > 0) = \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)}\alpha^{K(g)}\prod_{j=1}^{K(g)}\Gamma(n_j)$ where $n_j = \sum_{i=1}^{n} 1(g_i = k)$ is the number of elements in cluster k $(k = 1, \ldots, K(g))$ and $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$ is the Gamma function (Antoniak, 1974).

As shown in Equation (1), we adopt the RCRP with the adjacency matrix A that assigns the observation to one of the clusters of observations sharing a boundary, or a new cluster with probability proportional to α. This strategy gives rise to considering adjacent neighboring clusters of regions. Thus, the partition distribution with regard to cluster memberships is now given by

$$P(g|\alpha, A) = \frac{\alpha^{K(g)}}{C(\alpha, A)}\prod_{j=1}^{K(g)} n_j(g)Q(g, A), \tag{3}$$

where $C(\alpha, A)$ is a normalizing constant and $Q(g, A) = 1$ if g is a cluster membership assignment that can be described by the adjacency matrix A (Wehrhahn et al., 2020). Under the consideration of the adjacency, Equation (2) is modified as,

$$\pi(g_i = k|g_{-i}, \alpha, A) \propto \begin{cases} n_k(g_{-i}) & \text{if } k \leqslant K(g_{-i}),\ Q(g, A) = 1 \\ \alpha & \text{if } k = K(g_{-i}) + 1,\ Q(g, A) = 1 \\ 0 & \text{if } Q(g, A) = 0 \end{cases}, \tag{4}$$

In cluster assignments, an increasing trend of K is expected to be controlled by the use of adjacency matrix A, since the cluster g_i is derived by a restricted set of neighboring observations.

Based on the distributions above, first, we update α using a Metropolis-Hastings sampler. The prior distribution for α is an inverse Gamma distribution $IG(1, 1)$ (Roy et al., 2018). Similar to an implementation by Oganisian et al. (2021), the proposed value of α^* is generated from $N(\alpha, 1)$, where α is the value in the previous step. For our parameter vector θ under the DP prior, we accept the proposal with probability of $\min(1, H)$ where H is the Hastings acceptance ratio. Since we assume a symmetric proposal density which is $f(\alpha|\alpha^*) = f(\alpha^*|\alpha)$ for all α and α^*, the ratio of the two proposal distributions $\frac{f(\alpha|\alpha^*)}{f(\alpha^*|\alpha)}$ is 1 and thus the acceptance ratio H is expressed as

$$\begin{aligned} H &= \frac{\pi(\alpha^*)P(g|\alpha^*, A)}{\pi(\alpha)P(g|\alpha, A)} \\ &= \exp\left(\sum_{j=1}^{K} \ln\left(\frac{n_j + \frac{\alpha^*}{K}}{n_j + \frac{\alpha}{K}}\right) - \ln\left(\frac{n + \alpha^*}{n + \alpha}\right) - (K + 1)\ln\left(\frac{\alpha^*}{\alpha}\right) - \left(\frac{1}{\alpha^*} - \frac{1}{\alpha}\right)\right). \end{aligned} \tag{5}$$

After updating α, we update the cluster-wise parameters $\theta_k = (\beta_k, \mu_k, \phi_k)$, $k = 1, \ldots, K$ conditional on cluster membership using a Bayesian binary probit regression model as described in Albert and Chib (1993). First, let the prior distribution of β_k be $\pi(\beta_k)$ over the model parameters. Then, we obtain the posterior distribution of the model parameters as

$$\begin{aligned} p(\beta_k|Y_k, X_k) &= \pi(\beta_k)\prod_{i=1}^{n_k} p(Y_{ik}|\beta_k, X_{ik}) \\ &= \pi(\beta_k)\prod_{i=1}^{n_k}\{\Phi(X_{ik}^T\beta_k)\}^{Y_{ik}}\{1 - \Phi(X_{ik}^T\beta_k)\}^{1-Y_{ik}}, \end{aligned} \tag{6}$$

where Φ is the standard normal cumulative distribution function, and the subscript k in Z_k and X_k indicates the data within the k-th cluster. Assuming that there exists no conjugate prior $\pi(\beta_k)$ for the parameters, a direct calculation of the posterior distribution above is difficult. Instead, we augment the original Bayesian probit model with an additional latent variable Z to compute the posterior distribution by using the conditional distributions equivalent to the conditional distributions under a Bayesian normal linear regression model (Albert and Chib, 1993). For this, we assume $Y_{ik} = I(Z_{ik} > 0)$ where $Z_{ik} = X_{ik}^T\beta_k + \epsilon_{ik}$ with ϵ_{ik} following the standard normal distribution as

$$P(Y_{ik} = 1|X_{ik}, \beta_k) = P(X_{ik}^T\beta_k + \epsilon_{ik} > 0) = P(\epsilon_{ik} < X_{ik}^T\beta_k) = \Phi(X_{ik}^T\beta_k).$$

Taking into account the status of observed Y_{ik}, the full conditional distribution of Z_k given β_k, Y_k, X_k has the form

$$Z_{ik}|\beta_k, Y_{ik}, X_{ik} \sim \begin{cases} \psi(\bar{\delta}_k, 1, -\infty, 0) & \text{if } Y_{ik} = 0, \\ \psi(\bar{\delta}_k, 1, 0, \infty) & \text{if } Y_{ik} = 1, \end{cases}$$

where ψ is a truncated normal distribution with mean $\bar{\delta}_k = X_{ik}^T\beta_k$ and variance 1 within an interval $(-\infty, 0)$ or $(0, \infty)$ for $Y_{ik} = 0$ and 1, respectively. The posterior distribution of $p(\beta_k|Z_{ik}, X_{ik})$ in Equation (6) is expressed as

$$p(\beta_k|Z_k, X_k) = \pi(\beta_k)\prod_{i=1}^{n_k} p(Z_{ik}|\beta_k, X_{ik}). \tag{7}$$

With a constant prior for β_k, i.e., $\pi(\beta_k) \propto 1$, the conditional distribution (7) is given as $\beta_k|Z_k, X_k \sim N((X_k^TX_k)^{-1}X_k^TZ_k, (X_k^TX_k)^{-1})$. Assume the proper conjugate prior for β_k as $\pi(\beta_k) \sim N(\beta_0, P_0)$. Then, we have $\beta_k|Z_k, X_k \sim N((P_0^{-1} + X_k^TX_k)^{-1}(P_0^{-1}\beta_{k,old} + X_k^TZ_k), (P_0^{-1} + X_k^TX_k)^{-1})$ using the previous draw $\beta_{k,old}$.

Next, we update the mean and variance μ_{lk} and ϕ_{lk} of the covariates for the l-th covariate $(l = 1, \ldots, p)$ in cluster k by Gibbs samplers. The priors used are $\mu_{lk} \sim N(m_0, v_0)$ and $\phi_{lk} \sim IG(g_0, b_0)$. The covariates are sampled from the posterior distributions

$$\mu_{lk}^{(t+1)} \sim N\left(\frac{\frac{m_0}{v_0} + \frac{\sum_{j=1}^{n_k} X_{lj}}{\phi_{lk}^{(t)}}}{\frac{1}{v_0} + \frac{n_k}{\phi_{lk}^{(t)}}}, \sqrt{\frac{1}{\frac{1}{v_0} + \frac{n_k}{\phi_{lk}^{(t)}}}}\right),$$

$$\phi_{lk}^{(t+1)} \sim IG\left(g_0 + \frac{n_k}{2}, \frac{1}{2}\sum_{j=1}^{n_k}\left(X_{lj} - \mu_{lk}^{(t+1)}\right) + b_0\right),$$

where $\sum_{j=1}^{n_k} X_{lj}$ indicates the sum of X_{lj} within cluster k. The hyperparameters m_0 and v_0 are determined by the mean and variance matrix of covariates β_k, but for v_0, the variance matrix is multiplied by an identity matrix with a small constant, i.e., 0.001 to keep noise under control. Although there is no theoretical guidance, the hyperparameter g_0 is given as a constant, i.e., $g_0 = 2$, and b_0 is the variance of X_{jk}, as suggested to bring the center of the prior distribution closer to the empirically estimated values based on observed data (Oganisian and Roy, 2021).

Now, we sample $\theta_i|D$ as

$$P(\theta_i|D) \propto \frac{1}{\alpha + i - 1}\left\{\alpha P(D_i|\theta_i)G_0(\theta_i) + \sum_{j<i} P(D_i|\theta_j)\delta_{\theta_j}(\theta_i)\right\}Q(g, A),$$

where $\delta_{\theta_j}(\theta_i)$ is a Dirac measure of θ_j whose value is 1 if the i-th and j-th observations are in the same cluster, or 0 otherwise, and $G_0(\theta_i)$ is a prior. $\sum_{j<i} P(D_i|\theta_j)$ expresses the posterior distribution of the data based on the prior distribution in Equation (4). After updating both the concentration parameter α and the set of parameters θ, we update the cluster membership of the i-th observation via MCMC samplers using the categorical distribution

$$g_i^{(t+1)}|g_{-i}^{(t)} \sim Cat\Bigg(\frac{1}{\alpha+i-1}P^*\left(D_i|\theta_1^{(t+1)}\right)Q(g^{(t)},A),\ldots,\frac{1}{\alpha+i-1}P^*\left(D_i|\theta_{i-1}^{(t+1)}\right)Q(g^{(t)},A),\\ \frac{\alpha}{\alpha+i-1}P^*\left(D_i|\theta_0^{(t+1)}\right)Q(g^{(t)},A)\Bigg),$$

where $P^*(D_i|\theta_k^{(t+1)})$ is the scaled posterior distribution $P(D_i|\theta_k^{(t+1)})$ to be a probability given by the parameters at $(t+1)$-th iteration, $Q(g_i, A)$ is an indicator determined by the neighbors of the i-th observation and the adjacency matrix A.

Oganisian et al. (2021) offers an R function called 'PDPMix' for posterior sampling of DP mixture of logistic regression in the R package 'ChiRP' (Oganisian, 2019), extending the traditional CRP method by incorporating a regression model to capture the relationship between the covariates and the cluster assignments. For the implementation of our proposed method, we modified PDPMix and summarized partition results from MCMC iterations as shown in the following section.

2.2 Summary of MCMC Iterations

Mode clustering

Inference in Bayesian nonparametric clustering models usually relies on MCMC techniques, which produce a large number of cluster assignments approximating samples from the posterior distribution. However, displaying all unique clustering results is neither feasible nor applicable in real data analysis. To address this, various efforts have been made to summarize the iterations and construct a final cluster result from the entire set of iterations. These methods include multi-clustering fusion techniques such as 'boost-clustering' (Frossyniotis et al., 2004), posterior mode (Heller and Ghahramani, 2005; Dahl, 2009; Raykov et al., 2016), and posterior similarity matrices (Medvedovic and Sivaganesan, 2002; Medvedovic et al., 2004; Rasmussen et al., 2008). The posterior mode clustering is defined as the partition that maximizes the posterior distribution $p(\pi|D)$ of the parameter of interest over the entire partition space. This is also known as the maximum a posteriori (MAP) and is often obtained through the optimal Bayes estimate of the clustering under a specific loss function, such as 0-1 loss. In our example, we use mode clustering in the application of the posterior similarity matrix. We consider label switching since cluster information is a nominal variable and not fixed; that is, cluster 1 at iteration i could correspond to cluster 2 at iteration $i+1$. The issue is resolved through deterministic relabeling of clusters based on posterior sampling (Rodriguez and Walker, 2014). The optimal cluster is determined by finding the posterior mode matrix that minimizes the $L2$ loss. Each element of the matrix represents the average number of times that subjects i and j are assigned to the same cluster across all iterations.

Credible ball

Bayesian clustering outputs typically consist of the posterior mode, obtained from multiple sets of cluster structures generated through MCMC iterations. Wade and Ghahramani (2018)

offers an approach for point estimation of the posterior distribution by defining credible balls, which represent the posterior distribution over the entire clustering space. These credible balls, equivalent to credible intervals, reflect the uncertainty in the clustering structure given the data. To compare cluster assignments from two iterations (c_1, c_2), a modified loss function known as the variation of information (VI) is used as follows (Meilă, 2007):

$$\mathrm{VI}(c_1, c_2) = H(c_1) + H(c_2) - 2I(c_1, c_2),$$

where $H(c_i)$ is the entropy of the cluster assignment c_i measuring the uncertainty of a particular clustering, and $I(c_1, c_2)$ is the mutual information between two clustering assignments c_1, c_2, indicating the decrease in uncertainty of data points' cluster assignment in c_1 given its clustering assignment in c_2 (Wade and Ghahramani, 2018). Under the application of VI, the optimal partition c^* is found by identifying the cluster assignment with the minimum expected VI given the data D as follows:

$$\begin{aligned} c^* &= \mathrm{argmin}_{c_2} E(\mathrm{VI}(c_1, c_2|D)) \\ &= \mathrm{argmin}_{c_2} \left[\sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{n} 1(c_{2j} = c_{2i}) \right\} - 2 \sum_{i=1}^{n} E \left\{ \log \left(\sum_{j=1}^{n} 1(c_{1j} = c_{1i}, c_{2j} = c_{2i})|D \right) \right\} \right], \end{aligned}$$

where c_{1i} and c_{2i} indicate the i-th subject's cluster given c_1 and c_2, respectively, and the expectation in the second term is approximated by the posterior similarity matrix using the entire MCMC cluster assignments (Wade and Ghahramani, 2018).

Let $d(c_1, c_2)$ indicate a distance metric such as VI between cluster assignment c_1 and c_2, and C indicate the space of partitions. Then, a ball around c_1 of size ϵ is constructed as $B_\epsilon(c_1) = \{c_2 \in C : d(c_1, c_2) \leqslant \epsilon\}$, reflecting an intuition of the closest set of clusters to c_1. The credible ball is defined to characterize the uncertainty in the point estimate c^* with a given credible level $1 - \alpha, \alpha \in [0, 1]$, as $B_{\epsilon^*}(c^*) = \{c_1 : d(c^*, c_1) \leqslant \epsilon^*\}$ where ϵ^* is the smallest $\epsilon \geqslant 0$ such that $P(B_\epsilon(c^*)|D) \geqslant 1 - \alpha$. An R package 'mcclust.ext' finds a cluster assignment minimizing the posterior expected VI (Wade and Wade, 2015).

3 Simulation Study

3.1 Image Simulation

We conduct two simulation studies to demonstrate the performance of our proposed RDMDP model in identifying different patterns in images. We first generate image patterns on a 30×30 grid, consisting of 900 locations. An adjacency matrix defines the spatial relationship between locations based on a binary first-order neighborhood structure (Figure 1). In the simulation, we first generate observations within each cluster. Each observation comprises three numerical data points (v_1,v_2,v_3) drawn from a multivariate normal distribution, as well as a random binary outcome (z) generated from a Probit model based on the covariates of v_1,v_2 and v_3. For each location on the 30×30 grid, starting from the bottom-left corner, we randomly assign a cluster label in accordance with the binary first-order neighborhood structure. The pattern in Figure 1 is designed to resemble the spatial connections in a disease map, as presented in Wehrhahn et al. (2020), where adjacent locations are likely to share the same cluster memberships. To simulate this pattern, we determine the ground truth class assignment for each grid location based on a basic rule: it assigns a class with a probability of 0.9 using the class information of neighboring

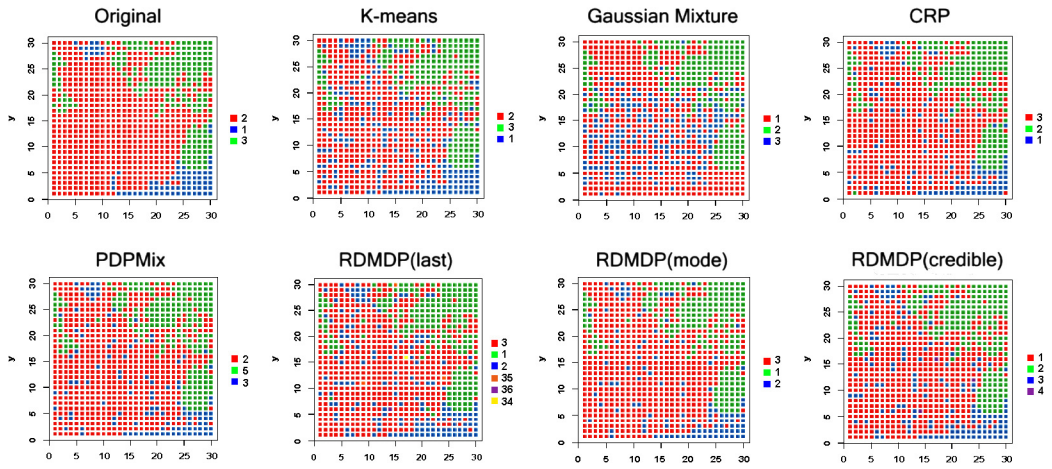

Figure 1: Simulation results of Pattern 1.

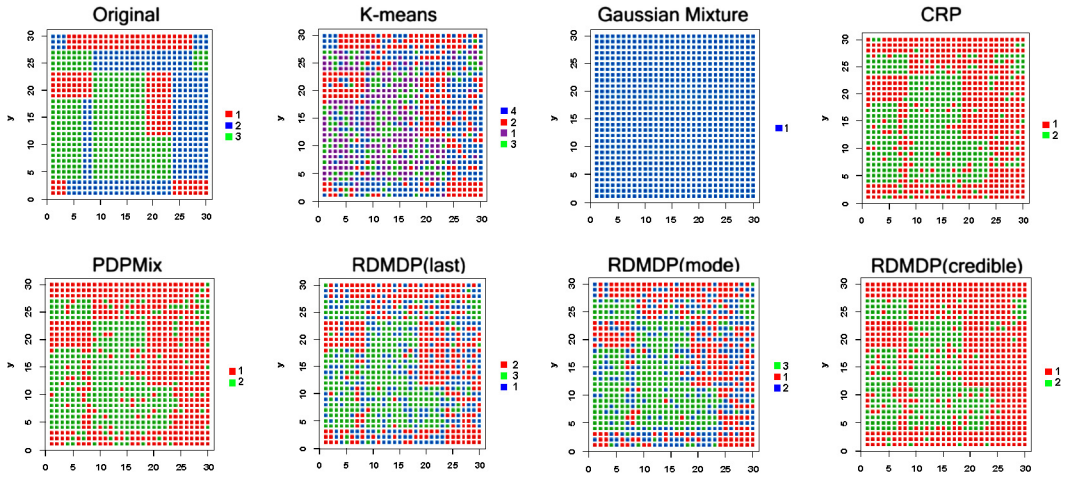

Figure 2: Simulation results of Pattern 2.

grids. The number of neighboring grid clusters in this example ranges from three to eight. At each location, the associated data are drawn from the pool of generated data according to the assigned clusters. The different clusters are color-coded in red, green, and blue, respectively.

For an alternative cluster assignment strategy, we also generate patterns where the images consist of streaks of clusters (Figure 2). The pattern in Figure 2 features a combination of elongated rectangles within the 30×30 grid, varying in width and length. This design is inspired by de Stijl art (Wilkins et al., 2009). According to this pattern, we assign three ground truth classes to each location following the predetermined pattern. Associated numeric and binary data are similarly assigned using the method mentioned earlier. The first images (top-left image) in Figure 1 and Figure 2 represent the ground truth patterns. We evaluate the performance of pattern identification by both the RDMDP and other existing techniques in comparison with these ground truth patterns.

3.2 Comparison Methods

We compare the performance of our proposed RDMDP model with existing techniques such as K-means clustering, hierarchical mixture model, and regression-based and non-regression-based CRP models. Although DP-based clustering can be considered a nonparametric version of K-means, as suggested by Raykov et al. (2016), there are notable distinctions in both methodology and performance that warrant further investigation.

K-means clustering (K-means) is a widely used algorithm in unsupervised learning due to its simplicity (Hartigan and Wong, 1979). It aims to group data points into K clusters based on centroids, resulting in the minimum total within-cluster variance. The optimal value of K is typically determined by examining the scree plot.

Gaussian mixture modeling for clustering (Gaussian mixture) employs finite Gaussian mixture models for model-based clustering using the finite Gaussian mixture models (Fraley et al., 2012). The algorithm is initiated with hierarchical model-based agglomerative clustering and selects up to nine clusters using the BIC criteria. The Expectation-Maximization (EM) steps estimate the conditional probability of the i-th data point being assigned to cluster k, given the current parameter estimates, and compute the maximum likelihood estimates of the parameters based on these conditional probabilities. This method has the advantage of handling complex data structures by providing detailed relationships between data points. However, its performance suffers when the components are not well-separated. Unlike predetermined K from K-means and optimized K from EM-based optimization from Gaussian mixture, CRP models determine the number of clusters empirically in a Bayesian approach.

The regression-based CRP model, implemented using the 'PDPMix' function in the R package 'ChiRP' (PDPMix), employs a probabilistic approach within the realm of Bayesian nonparametric methods to cluster data based on a set of features or covariates. Unlike traditional CRP models, the regression-based CRP variants such as PDPMix include a regression component that allows the probability of a data point joining a specific cluster to depend on its features or covariates. This regression component assesses the likelihood of a data point being explained by the regression model, given the coefficients of its covariates. The regression-based CRP is particularly useful when there is a strong correlation between the covariates and the response variable, as it enables more accurate clustering based on their relationships. Various types of regression models, including linear regression, logistic regression, and Gaussian processes, can be used in implementing the regression-based CRP. In practice, the regression model is trained using Bayesian methods, allowing for the uncertainty in model parameters to be incorporated into the clustering process.

For comparison purposes, we also modify the regression-based CRP by omitting the regression component. In this version, the relationship between the outcome data and the covariates is not considered. We refer to this method simply as CRP. Additional simulation results based on replicated ground truth figures are also provided in Supplementary Material 2.

3.3 Comparisons Between Clustering Methods

The ground truth patterns and their corresponding clustering results are displayed in Figures 1 and 2. Cluster assignments obtained from MCMC are presented in three ways: the last iteration, mode clustering, and credible ball. In Figure 1, for the K-means clustering method, the number K is selected to be 3 based on the elbow observed in the scree plot. While most methods effectively distinguish the green area in the pattern, significant confusion appears between the red and blue areas in cluster assignments. The bottom-right blue area in the ground truth is not

Table 1: Clustering performance measures of two patterns. K-means, Gaussian mixture, CRP and PDPMix do not consider pixel location, while RDMDP utilizes an adjacency matrix for locational variation caused by covariates and location. RDMDP results are represented by three outcomes (l: last iteration, m: mode clustering, c: credible ball).

Pattern 1	Silhouette	Entropy	ARI	CHI	Median rank
K-means	0.3209	0.3637	0.6145	89.4481	6
Gaussian Mixture	0.2796	0.9034	0.4370	83.9682	7
CRP	0.2821	0.2707	0.7628	603.888	4.5
PDPMix	0.3102	0.2567	0.6854	597.9104	4
RDMDP(l)	0.2647	0.2256	1.0000	202.9998	3
RDMDP(m)	0.3153	0.2285	0.7047	593.1295	3
RDMDP(c)	0.3179	0.2443	0.7976	360.59	2.5
Pattern 2	Silhouette	Entropy	ARI	CHI	Median rank
K-means	0.2019	0.6167	0.3858	26.3301	4.5
Gaussian Mixture	0.1820	0.9703	0.2012	10.2430	7
CRP	0.1478	0.4811	0.6395	633.4640	3
PDPMix	0.1411	0.4662	0.6809	631.8258	2.5
RDMDP(l)	0.1459	0.6655	1.0000	339.2168	5.5
RDMDP(m)	0.1796	0.6207	0.3903	396.4520	4
RDMDP(c)	0.3476	0.4225	0.3691	635.2373	1

well captured by the Gaussian mixture modeling for clustering method. Although both K-means and RDMDP accurately identify the true blue pattern, K-means shows more blue assignments in the central red area than RDMDP. Among the summarization methods for RDMDP, we note that the mode clustering method exhibits the best performance. The credible ball method tends to have more misclassifications in the red area compared to mode clustering. We also include the cluster assignments from the last MCMC iteration for RDMDP. As expected, the last iteration reveals miscellaneous and small clusters, although the overall patterns are similar to summarized patterns from MCMC. Clustering performance indices do not consistently demonstrate the superiority of any single method (Table 1). Therefore, we calculate the median rank using ranks based on silhouette, entropy, adjusted rand index (ARI), and Calinski-Harabasz index (CHI). Comparisons of clustering performance indices using mode clustering and credible ball methods show either the best performances or performance comparable to other methods. These results underscore the superiority of RDMDP algorithms over alternative methods.

In Figure 2, the Gaussian mixture modeling for clustering method fails to identify any meaningful pattern, assigning all locations to a single cluster. In both regression-based CRP and original CRP, the red and blue patterns are combined and assigned to one cluster. K-means clustering, guided by the scree plot, selects four clusters but fails to distinguish the blue and green patterns. It also introduces a new cluster, represented by the color purple, that extends across both the green and blue areas. On the other hand, RDMDP with mode clustering better identifies the streak patterns compared to other methods. A comparison of clustering performance indices based on their median rank reveals that RDMDP with credible ball ranks the best among all methods, underscoring the effectiveness of RDMDP. However, this method fails to recognize the

blue cluster in the actual pattern. Although regression-based CRP and original CRP perform better in terms of clustering indices, RDMDP with mode clustering still achieves values that are comparable. The performance of RDMDP with mode clustering, particularly its ability to determine the optimal number of clusters and effectively differentiate between the red and blue clusters, strongly suggests that it may be the preferred method for pattern identification compared to other clustering techniques.

4 Real Data Analysis

By employing clustering techniques on brain regions, we conduct an exploratory analysis for feature selection and partitioning of brain regions. This allows us to gain initial insights into the overall structure of the brain. Using the proposed method RDMDP, our goal is to identify clusters of regions that exhibit a strong correlation with specific phenotypes, such as age groups.

4.1 Data Source

We use data from Transforming Research and Clinical Knowledge in Traumatic Brain Injury (TRACK-TBI) study, conducted by the International Traumatic Brain Injury Research Initiative (https://tracktbi.ucsf.edu/transforming-research-and-clinical-knowledge-tbi). The study aims to advance our understanding of TBI pathophysiology and improve patient selection and stratification in future clinical trials through the validation of imaging biomarkers in DTI. Data were collected from participating trauma patients at multiple centers in the US from 2013 to 2018 and are shared through the Federal Interagency Traumatic Brain Injury Research (FITBIR) informatics system (https://fitbir.nih.gov). The study includes 2,539 adult patients with TBI, with age ranging from 0 to over 90. The patients consist of 69% male. The downloadable data sets contain demographic data such as age and sex, as well as imaging biomarkers such as DTI scans.

A total of 47 subjects in the younger age group (age < 52 years) have fully available DTI scans. To create a balanced comparison group, we randomly select 47 subjects from the older age group (age $\geqslant 52$ years). Age 52 is selected as the age cut-off point based on previous clinical studies, which have indicated that declines in brain structure and cognitive functions can be observed when comparing groups divided by a certain age threshold. This threshold is based on accelerated changes that occur in the brain between the middle ages of 40 and 59 (e.g., Elliott et al. (2021); Park and Festini (2016)). For these 94 subjects, we obtain baseline DTI scans in the Neuroimaging Informatic Technology Initiative (NIfTI) format, along with demographic information.

DTI scans are standardized into the MNI152 template. Subsequently, the DTI tensor for each voxel is estimated and can be visualized using RGB color mapping to correspond to transverse, anteroposterior, and superior-inferior directions. After this estimation step, we export DTI metrics based on diffusion in compressed NIfTI format. These processes are carried out using DSIStudio (http://dsi-studio.labsolver.org/). Once the DTI metrics are exported, the data can be imported into R using packages such as 'oro.nifti' (Whitcher et al., 2011). DTI has dimensions of $128 \times 128 \times 59$ (973,312 voxels) or $256 \times 256 \times 59$ (3,875,328 voxels), depending on the resolution of the images.

4.2 Data Preparation

To reduce the computational burden, we combine neighboring voxels and calculate the average values of DTI metrics. To preserve the integrity of major structural components of the brain, such as its distinct lobes, it is crucial to ensure that the aggregated volumes are appropriately small. We define a constant size for unit cubes in the 3D DTI. For example, one cube contains $16 \times 16 \times 5$ voxels on x, y, and z axes where the x-axis represents the left-right direction, the y-axis represents the anterior-posterior direction, and the z-axis represents the superior-inferior direction. Cubes on the boundary may contain fewer voxels. This approach results in a total of 3,328 cubes, each with dimensions of $16 \times 16 \times 13$, effectively covering the whole brain while considerably reducing the dimensionality. For each cube, we calculate average DTI metrics (FA, MD, AD, RD) and strengths of diffusion in three directions (Dir1, Dir2, Dir3) for 94 subjects. The size of the unit cube may vary depending on the level of detail desired by the researcher, and the optimal sizes can be determined empirically during data analysis and result interpretation.

We then narrow down the selection to 651 cubes from the initial pool of 3,328 ($16 \times 16 \times 13$) cubes. These selected cubes contain data from more than 10 subjects and are used for cube-based analyses such as logistic regression or K-means clustering. In logistic regression, we model the age group using covariates that consist of subject-level summaries (i.e., averages and standard deviations) of DTI metrics and strengths in each direction. K-means clustering includes the age group as a covariate, along with DTI metrics and strengths of directions. For each cube, we estimate the predictive ability of the binary age group as either the area under the receiver operating characteristic curve (AUC) using logistic regression, or entropy using K-means clustering. AUC is a metric used for evaluating the performance of binary classification, while entropy measures uncertainty in clustering. We then create a classification index (CI) using either binary AUC or binary entropy. Using AUC, CI is 1 if the AUC exceeds a threshold (0.7 in our analysis) or 0 otherwise. Using entropy, CI is 1 if entropy is less than a threshold (0.5 in our analysis), or 0 otherwise. The final dataset has dimensions of 651×471, where 471 includes two DTI metrics (FA, MD) and the three strengths of directions for 94 subjects, as well as the CI for each cube. DTI metrics such as AD and RD are excluded from the final dataset since most classification patterns in AD and RD overlap with FA. We analyze the prepared data using RDMDP and, for comparison, K-means clustering.

Since the cubes encompass all brain tissues, we create separate datasets for white matter and gray matter, maintaining the data structures described above. In DSIStudio, we export the selected regions of white matter and gray matter to separate NIfTI files by filtering them out from the whole-brain scan. These files contain a binary variable indicating whether a given location in the x, y, z axes has white or gray matter in the voxel.

4.3 Results

Clustering results from RDMDP and K-means clustering, using the whole-brain tissue data, are compared in Table 2. This comparison focuses on the average values of DTI metrics assigned to each cluster. The scree plot helps to identify the optimal number of clusters for K-means clustering. For RDMDP, the mode clustering method involves 5,000 MCMC iterations following a 1,000-iteration burn-in period and determines the number of clusters. RDMDP identifies one distinctive cluster (c1) with a high mean binary AUC, as shown in Table 2. The mean FA values are lower in the red cluster (c1), at 0.3446, compared to those in the blue (c2) and green (c3) clusters, which have values of 0.9907 and 0.9019, respectively. Lower FA values may suggest that the tissue in these clusters is less structurally intact (Basser and Jones, 2002).

Table 2: Clustering results of 651 cubes (whole brain scans). c1, c2, and c3 refer to three clusters determined by our proposed method and K-means. $\overline{FA}$: the average of the FA in all cubes by each cluster, same for other metrics; $\overline{AUC07}$, $\overline{Ent}$: the averages of binary values (0 or 1) of classification indices.

RDMDP	$\overline{FA}$	$\overline{MD}$	$\overline{Dir1}$	$\overline{Dir2}$	$\overline{Dir3}$	$\overline{AUC07}$	$\overline{Ent}$
c1	0.3446	-0.0058	-0.0082	0.0065	0.0518	0.7816	0.8506
c2	0.9907	0.0135	-0.0035	0.0020	0.2391	0.0190	0.9924
c3	0.9019	-0.0210	-0.0070	0.0067	0.1192	0.2383	0.8738
K-means	$\overline{FA}$	$\overline{MD}$	$\overline{Dir1}$	$\overline{Dir2}$	$\overline{Dir3}$	$\overline{AUC07}$	$\overline{Ent}$
c1	1.0092	0.0012	-0.0043	0.0041	0.2025	0.0558	0.9563
c2	0.2806	-0.0120	-0.0077	0.0045	0.0560	0.8303	0.8667
c3	0.6950	-0.0062	-0.0109	0.0090	0.0640	0.4324	0.7973

Table 3: Clustering results of 504 cubes (white matter). RDMDP identifies clusters w1 and w2, and K-means clustering identify clusters w1, w2 and w3. $\overline{FA}$ refers to the average of FA in all cubes by each cluster, and same for other DTI metrics (FA, MD, Dir1, Dir2, Dir3). $\overline{AUC07}$ and $\overline{Ent}$ are the averages of binary values (0 or 1) of classification indices.

RDMDP	$\overline{FA}$	$\overline{MD}$	$\overline{Dir1}$	$\overline{Dir2}$	$\overline{Dir3}$	$\overline{AUC07}$	$\overline{Ent}$
w1	0.8382	0.0045	-0.0065	0.0046	0.2672	0.0387	0.9903
w2	0.3777	-0.0010	-0.0035	0.0020	0.0705	0.5825	0.9278
K-means	$\overline{FA}$	$\overline{MD}$	$\overline{Dir1}$	$\overline{Dir2}$	$\overline{Dir3}$	$\overline{AUC07}$	$\overline{Ent}$
w1	0.8430	0.0063	-0.0060	0.0028	0.2651	0.0478	0.9873
w2	0.4050	-0.0055	-0.0049	5E-4	0.0837	0.5543	0.9674
w3	0.3180	-0.0027	-0.0036	0.0090	0.0567	0.6020	0.8980

Figures 3 and 4 illustrate the spatial distribution of clusters from RDMDP and K-means clustering, respectively, across 13 x-y slices (each comprising 16 x 16 cubes) in the z direction. These slices cover the entire span from the bottom (1st slice) to the top (13th slice) of the brain. The cluster c1 (colored red) identified by RDMDP is mainly located at the bottom and top boundaries of the brain (1st, 2nd, 11th, and 12th slices by z-axis) and notably occupies the entire top of the brain (13th slice by z-axis). It is evident that the cluster c1 does not align well with areas of white matter, as shown in the white matter map in Supplementary Figure S3. On the other hand, clustering results based solely on white matter using RDMDP identify two clusters (colored red and blue in Supplementary Figure S3). These are comparable to the blue (c2) and green (c3) clusters in Figure 3. The average values of DTI metrics for clusters c2 and c3 in Tables 2 and for clusters w1 and w2 in Table 3 are similar in terms of both relative magnitude and direction. Based on these observations, it can be inferred that clusters c2 and c3 in Figure 3 are primarily influenced by the white matter.

The interpretation offered by RDMDP appears clearer (Figure 3) than that provided by K-means clustering, which lacks a similar level of clarity (Figure 4). In K-means clustering, two identified clusters (c2 and c3) display relatively high mean binary AUC values, as opposed to a distinctive high value achieved by cluster c1 in RDMDP in clustering (Table 2). RDMDP (Fig-

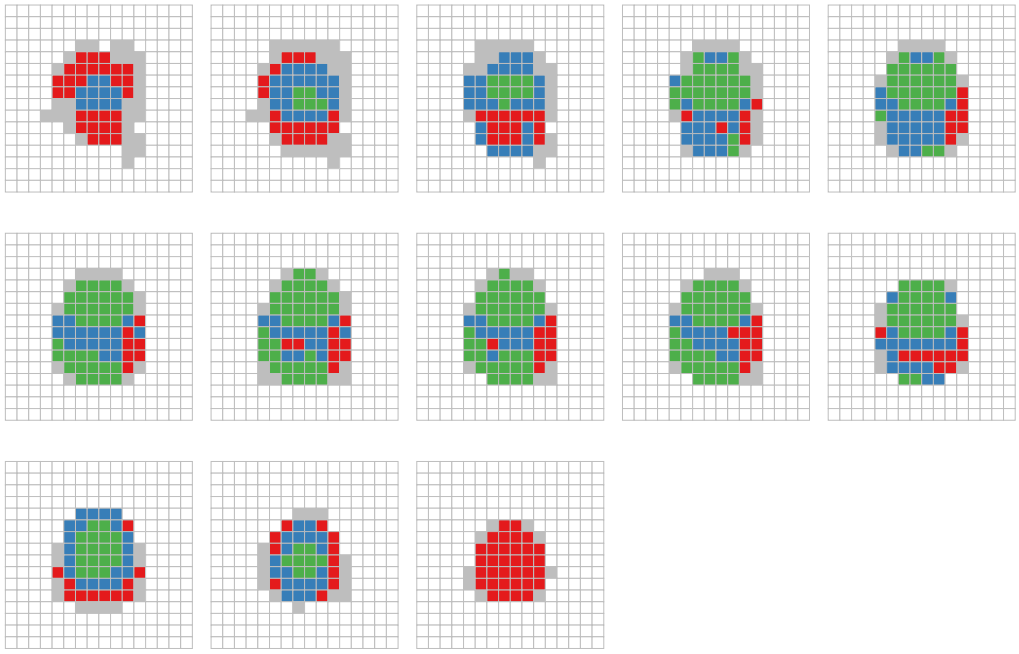

Figure 3: RDMDP clustering on 651 cubes. The clustering for all 13 layers were done all together and displayed by each layer in this figure. Clusters c1, c2, and c3, according to Table 2, are colored red, blue and green, respectively. Cubes in gray are not used for clustering because they contain fewer than 10 subjects. Each slice, arranged from left to right and bottom to top, depicts a series of two-dimensional representations of cubes spanning from the bottom to the top of the brain. Grey = Brain part not clustered, Red = c1, Blue = c2, Green = c3.

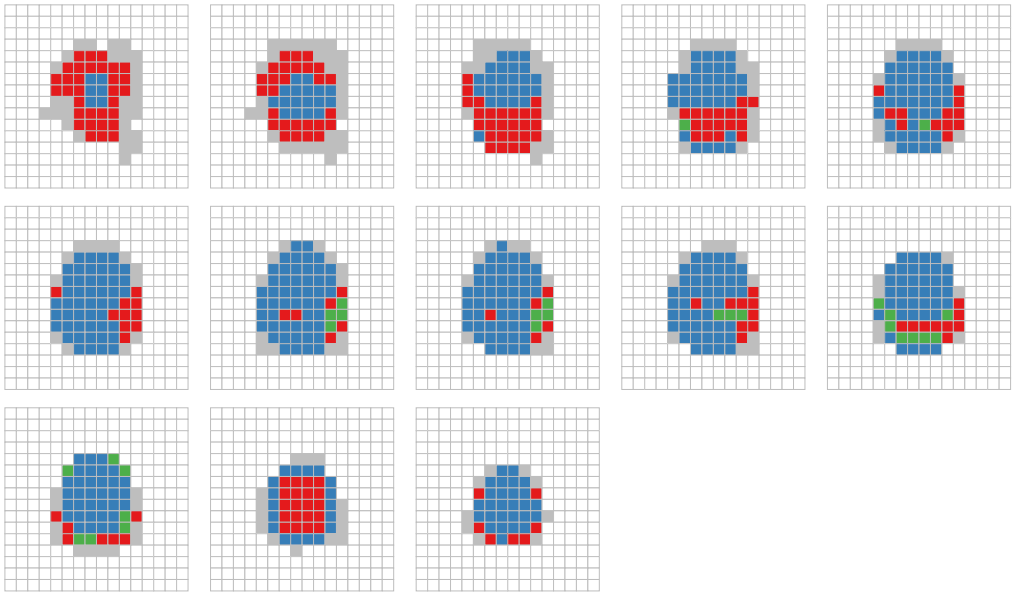

Figure 4: K-means clustering on 651 cubes where at least 10 subjects have DTI on each cube. Grey = brain part not clustered, red = c1, blue = c2, green = c3.

ure 3) reveals a stripe of the blue cluster that expands from the left to the right side of the brain, a pattern that is not distinguishable in K-means clustering (Figure 4). Notably, clustering results obtained from K-means clustering based solely on white matter (Supplementary Figure S4) exhibit patterns remarkably similar to those achieved through RDMDP (Supplementary Figure S3). This suggests that clustering performance is more influenced by the DTI metrics themselves than by age classification. Furthermore, K-means clustering does not effectively isolate the region associated with age when using data based on whole-brain tissues (AUC07 in Table 2).

The spatial distribution of the red cluster identified by RDMDP (Figure 3) suggests that the associated region is not primarily composed of white matter. Instead, this region may contain other components, such as gray matter or other non-white matter structures.

Numerous studies have shown age-related decreases in the volume of various brain regions. The regions identified through RDMDP could be especially susceptible to age-related degeneration. For example, Bernard and Seidler (2014) highlighted age-related volume changes in the cerebellum. However, evidence suggests that age-related differences in diffusion properties, as measured by DTI, may be more sensitive indicators of age-related changes than volumetric measures (Leritz et al., 2014). To delve deeper into this, we utilized 23 regions of interest (ROIs) from BrainSeg in DSIStudio. Our findings suggest that the bottom part of the brain, specifically in slices 1 to 3, may be related to the cerebellum (Supplementary Table S2).This observation aligns with the presence of the red cluster identified in our RDMDP analysis (Figure 3), supporting the hypothesis of age-related changes in cerebellum DTI metrics. More discussions relevant to the connection between identified clusters and ROIs are found in the Supplementary Material 3.

5 Discussion

In this article, we introduced the RDMDP method, which capitalized on the advantages of Bayesian nonparametric approaches for pattern identification. Our simulation demonstrated that RDMDP was either comparable or superior to existing clustering methods such as K-means, Gaussian mixture model for clustering, and the original and regression-based CRP methods. We detailed the steps involved in data preparation to make it suitable for analysis and conducted a thorough examination of the results. Through both simulation studies and the analysis of DTI data from the TRACK-TBI study, we showcased the benefit of RDMDP in yielding a nuanced understanding of brain structure.

The proposed method based on the Dirichlet mixture model shows considerable promise for several compelling reasons. First, a key strength of this approach lies in its ability to identify clusters without requiring a priori specification of the number of clusters. This is particularly valuable in situations where ground truth is not available and the data structures are complex. Second, our method integrates the relationships between patients' conditions and DTI metrics at individual locations, potentially identifying clinically significant patterns and negating the need for manual annotation of ROIs. Third, our method assumes that clusters with similar characteristics are more likely to be in close proximity to one another, an assumption that aligns with the anatomical organization of functional brain areas (Saatman et al., 2008; Blei and Frazier, 2011; Wehrhahn et al., 2020).

However, there are avenues for further investigation. Specifically, the use of a single cube size to encompass multiple tissue types may introduce excessive variability into our measurements, especially when ground truth is not available. On a technical level, some voxels within a cluster may be uninformative and offer no insight (Lazar, 2008). Separating these uninformative voxels

from informative ones could enhance both the performance and interpretability of our method. Additionally, translating the results of clustering into clinically interpretable information remains a challenge.

For future directions, we aim to delve deeper into the study of specific brain regions, such as the cerebellum, in relation to both white and gray matter. This will provide a more detailed understanding of these regions and contribute to our overarching goal of improved brain mapping. It is known that the aging brain exhibits redundancy, meaning it possesses multiple regions or networks capable of performing similar functions (Sadiq et al., 2021). Additionally, the brain shows plasticity by continually reorganizing itself through the formation of new neural connections over the course of one's life (Park and Bischof, 2022). Furthermore, the aging brain has adaptability by employing compensation mechanisms to address age-related changes or declines in specific functions (Stern et al., 2019). In the analysis of age-related brain imaging data, we may need to take into account not only the patterns and clusters in the data but also consider how the brain's functional redundancy, plasticity, and compensation mechanisms might influence or shape those patterns. This understanding may draw comprehensive conclusions from the analysis in the future.

In conclusion, RDMDP offers a promising approach to addressing both technical and clinical challenges in neuroscience, particularly in the evolving field of DTI analysis methods. Its capacity to identify clinically relevant clusters without manual annotation and its demonstrated ability to provide a coherent interpretation of brain structure make it a valuable tool for future research.

Data and Code Availability Statement

The dataset is available on FITBIR website with administrative approval. The R code for RDMDP is available online as the Supplementary Material.

Supplementary Material

Supplementary Materials include a MCMC algorithm for RDMDP method, a simulation study for 100 replications, explanation of the connection between identified brain clusters and the region of interest, and the statement regarding R code for the RDMDP method.

Acknowledgement

The data used in this manuscript were obtained and analyzed from a controlled access dataset distributed by Federal Interagency Traumatic Brain Injury Research (FITBIR) Informatics Systems, supported by both the DOD and NIH. The TRACK-TBI prospective study (Study DOI: 10.23718/FITBIR/1518881, ORC ID: 0000-0002-0926-3128, Grant ID: 1U01NS086090-01) is funded by NINDS and is a multicenter initiative aimed at Transforming Research and Clinical Knowledge in Traumatic Brain Injury. The study is led by multiple Principal Investigators including Geoffrey Manley (MD, PhD), Ramon Diaz-Arrastia (MD, PhD), Joe Giacino (PhD), Pratik Mukherjee (MD, PhD), David Okonkwo (MD, PhD), Claudia Robertson (MD), and Nancy Temkin (PhD).

References

Ahmed A, Xing E (2008). Dynamic non-parametric mixture models and the recurrent Chinese restaurant process: With applications to evolutionary clustering. In: *Proceedings of the 2008 Siam International Conference on Data Mining*, 219–230. SIAM.

Albert JH, Chib S (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422): 669–679. https://doi.org/10.1080/01621459.1993.10476321

Antoniak CE (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6): 1152–1174.

Baldassano C, Beck DM, Fei-Fei L (2015). Parcellating connectivity in spatial maps. *PeerJ*, 3: e784. https://doi.org/10.7717/peerj.784

Basser PJ, Jones DK (2002). Diffusion-tensor mri: Theory, experimental design and data analysis–a technical review. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo*, 15(7–8): 456–467.

Bernard JA, Seidler RD (2014). Moving forward: Age effects on the cerebellum underlie cognitive and motor declines. *Neuroscience and Biobehavioral Reviews*, 42: 193–207. https://doi.org/10.1016/j.neubiorev.2014.02.011

Blei DM, Frazier PI (2011). Distance dependent Chinese restaurant processes. *Journal of Machine Learning Research*, 12(8): 2461–2488.

Blei DM, Griffiths TL, Jordan MI (2007). The nested Chinese restaurant process and hierarchical topic models. *Journal of the ACM*, 57(2): 1–30. https://doi.org/10.1145/1667053.1667056

Creswell R, Robinson M, Gavaghan D, Parag KV, Lei CL, Lambert B (2023). A Bayesian nonparametric method for detecting rapid changes in disease transmission. *Journal of Theoretical Biology*, 558: 111351. https://doi.org/10.1016/j.jtbi.2022.111351

Dahl DB (2009). Modal clustering in a class of product partition models. *Bayesian Analysis*, 4(2): 243–264. https://doi.org/10.1214/09-BA409

Daniel Loyal J, Chen Y (2023). A Bayesian nonparametric latent space approach to modeling evolving communities in dynamic networks. *Bayesian Analysis*, 18(1): 49–77.

Duan JA, Guindani M, Gelfand AE (2007). Generalized spatial Dirichlet process models. *Biometrika*, 94(4): 809–825. https://doi.org/10.1093/biomet/asm071

Elliott ML, Belsky DW, Knodt AR, Ireland D, Melzer TR, Poulton R, et al. (2021). Brain-age in midlife is associated with accelerated biological aging and cognitive decline in a longitudinal birth cohort. *Molecular Psychiatry*, 26(8): 3829–3838. https://doi.org/10.1038/s41380-019-0626-7

ElNakieb Y, Ali MT, Elnakib A, Shalaby A, Soliman A, Mahmoud A, et al. (2021). The role of diffusion tensor MR imaging (DTI) of the brain in diagnosing autism spectrum disorder: Promising results. *Sensors*, 21(24): 8171. https://doi.org/10.3390/s21248171

Escobar MD, West M (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430): 577–588. https://doi.org/10.1080/01621459.1995.10476550

Fergusom T (1973). A Bayesian analysis of some nonparametric hierarchical models. *The Annals of Statistics*, 1: 209–230.

Fraley C, Raftery AE, Murphy TB, Scrucca L (2012). mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. 597: 1.

Frossyniotis D, Likas A, Stafylopatis A (2004). A clustering method based on boosting. *Pattern Recognition Letters*, 25(6): 641–654. https://doi.org/10.1016/j.patrec.2003.12.018

Ghosh S, Ungureanu A, Sudderth E, Blei D (2011). Spatial distance dependent chinese restaurant processes for image segmentation. *Advances in Neural Information Processing Systems*, 24.

Griffin JE, Steel MJ (2006). Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, 101(473): 179–194. https://doi.org/10.1198/016214505000000727

Hartigan JA, Wong MA (1979). Algorithm as 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, 28(1): 100–108.

Heller KA, Ghahramani Z (2005). Bayesian hierarchical clustering. In: *Proceedings of the 22nd International Conference on Machine Learning*, 297–304.

Jbabdi S, Johansen-Berg H (2011). Tractography: Where do we go from here? *Brain Connectivity*, 1(3): 169–183. https://doi.org/10.1089/brain.2011.0033

Jones DK, Cercignani M (2010). Twenty-five pitfalls in the analysis of diffusion MRI data. *NMR in Biomedicine*, 23(7): 803–820. https://doi.org/10.1002/nbm.1543

Kraus G (2023). *Traumatic Brain Injury: A Neurosurgeon's Perspective*. CRC Press.

Lan Z, Reich BJ, Bandyopadhyay D (2021). A spatial Bayesian semiparametric mixture model for positive definite matrices with applications in diffusion tensor imaging. *Canadian Journal of Statistics*, 49(1): 129–149. https://doi.org/10.1002/cjs.11601

Lazar NA (2008). *The Statistical Analysis of Functional MRI Data*, volume 7. Springer.

Leritz EC, Shepel J, Williams VJ, Lipsitz LA, McGlinchey RE, Milberg WP, et al. (2014). Associations between T1 white matter lesion volume and regional white matter microstructure in aging. *Human Brain Mapping*, 35(3): 1085–1100. https://doi.org/10.1002/hbm.22236

Lu J, Li M, Dunson DB (2018). Reducing over-clustering via the powered chinese restaurant process. ArXiv preprint: https://arxiv.org/abs/1802.05392

MacEachern SN (2000). Dependent dirichlet processes. *Unpublished manuscript, Department of Statistics, The Ohio State University*, 5.

Masoero L, Schraiber J, Broderick T (2021). Bayesian nonparametric strategies for power maximization in rare variants association studies. ArXiv preprint: https://arxiv.org/abs/2112.02032

Medvedovic M, Sivaganesan S (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18(9): 1194–1206. https://doi.org/10.1093/bioinformatics/18.9.1194

Medvedovic M, Yeung KY, Bumgarner RE (2004). Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, 20(8): 1222–1232. https://doi.org/10.1093/bioinformatics/bth068

Meilă M (2007). Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5): 873–895. https://doi.org/10.1016/j.jmva.2006.11.013

Neal RM (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2): 249–265. https://doi.org/10.1080/10618600.2000.10474879

Oganisian A (2019). Chirp: Chinese restaurant process mixtures for regression and clustering. *Journal of Open Source Software*, 4(35): 1287. https://doi.org/10.21105/joss.01287

Oganisian A, Mitra N, Roy JA (2021). A Bayesian nonparametric model for zero-inflated outcomes: Prediction, clustering, and causal estimation. *Biometrics*, 77(1): 125–135. https://doi.org/10.1111/biom.13244

Oganisian A, Roy JA (2021). A practical introduction to Bayesian estimation of causal effects: Parametric and nonparametric approaches. *Statistics in Medicine*, 40(2): 518–551. https://doi.org/10.1002/sim.8761

Orbanz P, Teh YW (2010). Bayesian nonparametric models. In: *Encyclopedia of Machine Learning*. Springer.

Parekh MB, Gurjarpadhye AA, Manoukian MA, Dubnika A, Rajadas J, Inayathullah M (2015). Recent developments in diffusion tensor imaging of brain. *Radiology Open Journal*, 1(1): 1. https://doi.org/10.17140/ROJ-1-101

Park DC, Bischof GN (2022). The aging mind: Neuroplasticity in response to cognitive training. *Dialogues in Clinical Neuroscience*, 15(1): 109–119. https://doi.org/10.31887/DCNS.2013.15.1/dpark

Park DC, Festini SB (2016). The middle-aged brain. In: *Cognitive Neuroscience of Aging*, 363–388. Oxford University Press.

Pitman J (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2): 145–158. https://doi.org/10.1007/BF01213386

Rasmussen C, De la Cruz BJ, Ghahramani Z, Wild DL (2008). Modeling and visualizing uncertainty in gene expression clusters using Dirichlet process mixtures. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(4): 615–628. https://doi.org/10.1109/TCBB.2007.70269

Raykov YP, Boukouvalas A, Little MA (2016). Simple approximate map inference for Dirichlet processes mixtures. *Electronic Journal of Statistics*, 10(2): 3548–3578. https://doi.org/10.1214/16-EJS1196

Ren Q, Wang Q, Zhang J, Chen S (2016). Unordered images selection for dense 3D reconstruction based on distance dependent Chinese restaurant process. In: *2016 12th World Congress on Intelligent Control and Automation (WCICA)*, 2969–2973. IEEE.

Rodriguez CE, Walker SG (2014). Label switching in Bayesian mixture models: Deterministic relabeling strategies. *Journal of Computational and Graphical Statistics*, 23(1): 25–45. https://doi.org/10.1080/10618600.2012.735624

Roy J, Lum KJ, Zeldow B, Dworkin JD, Re III VL, Daniels MJ (2018). Bayesian nonparametric generative models for causal inference with missing at random covariates. *Biometrics*, 74(4): 1193–1202. https://doi.org/10.1111/biom.12875

Saad F, Mansinghka V (2018). Temporally-reweighted Chinese restaurant process mixtures for clustering, imputing, and forecasting multivariate time series. In: *International Conference on Artificial Intelligence and Statistics*, 755–764. PMLR.

Saatman KE, Duhaime AC, Bullock R, Maas AI, Valadka A, Manley GT (2008). Classification of traumatic brain injury for targeted therapies. *Journal of Neurotrauma*, 25(7): 719–738. https://doi.org/10.1089/neu.2008.0586

Sadiq MU, Langella S, Giovanello KS, Mucha PJ, Dayan E (2021). Accrual of functional redundancy along the lifespan and its effects on cognition. *NeuroImage*, 229: 117737. https://doi.org/10.1016/j.neuroimage.2021.117737

Schilling KG, Daducci A, Maier-Hein K, Poupon C, Houde JC, Nath V, et al. (2019). Challenges in diffusion MRI tractography – lessons learned from international benchmark competitions. *Magnetic Resonance Imaging*, 57: 194–209. https://doi.org/10.1016/j.mri.2018.11.014

Seymour RG (2020). Bayesian nonparametric methods for individual-level stochastic epidemic models, Ph.D. thesis, University of Nottingham.

Soares JM, Marques P, Alves V, Sousa N (2013). A hitchhiker's guide to diffusion tensor imaging. *Frontiers in Neuroscience*, 7: 31.

Socher R, Maas A, Manning C (2011). Spectral Chinese restaurant processes: Nonparametric clustering based on similarities. In: *Proceedings of the Fourteenth International Conference on*

Artificial Intelligence and Statistics, 698–706. JMLR Workshop and Conference Proceedings.

Stern Y, Barnes CA, Grady C, Jones RN, Raz N (2019). Brain reserve, cognitive reserve, compensation, and maintenance: Operationalization, validity, and mechanisms of cognitive resilience. *Neurobiology of Aging*, 83: 124–129. https://doi.org/10.1016/j.neurobiolaging.2019.03.022

Teh J, Teh YW, Jordan MI, Beal MJ, Blei DM (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476): 1566–1581. https://doi.org/10.1198/016214506000000302

Wade S, Ghahramani Z (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis*, 13(2): 559–626. https://doi.org/10.1214/17-BA1073

Wade S, Wade MS (2015). Package 'mcclust. ext'. *Journal of Computational and Graphical Statistics*, 16: 526–558.

Wehrhahn C, Leonard S, Rodriguez A, Xifara T (2020). A Bayesian approach to disease clustering using restricted Chinese restaurant processes. *Electronic Journal of Statistics*, 14(1): 1449–1478.

Whitcher B, Schmid VJ, Thornton A (2011). Working with the DICOM and NIfTI data standards in R. *Journal of Statistical Software*, 44(6): 1–28. https://doi.org/10.18637/jss.v044.i06

Wilkins DG, Schultz B, Linduff KM (2009). *Art Past, Art Present*. Prentice Hall.

Xian MTS, Wade S (2022). Bayesian nonparametric scalar-on-image regression via potts-gibbs random partition models. ArXiv preprint: https://arxiv.org/abs/2206.11051