1 RDMDP Clustering

Here, we explain a MCMC algorithm for RDMDP clustering below. In the following discussion, we follow the definitions of the parameters and distributions in the main text.

Algorithm 1: RDMDP Clustering

Data: $D_i = (x_i, y_i), i = 1, 2, ..., n$ for a set of covariates, x_i, and a binary response y_i. T indicates the number of MCMC sampling.

Result: Final cluster assignment $g^* = (g_1^*, g_2^*, ..., g_n^*)$.

1 Initialize $g_{1:n}^{(0)}$ to $K^{(0)}$ initial clusters with unique labels;

2 Initialize $\alpha^{(0)} = 1$ and $\theta^{(0)} = (\beta^{(0)}, \mu^{(0)}, \psi^{(0)})$;

3 **for** $t \leftarrow 1$ **to** T **do**

4   Update $\alpha^{(t)}$ via Metropolis-Hastings with the acceptance rate H in Equation (5);

5   Update Cluster-specific Parameters;

6   **for** $k \in K^{(t-1)}$ **do**

7       $\beta_k^{(t)} \sim p(\beta | Y_k, X_k \propto \pi(\beta_k) \prod_{i=1}^{n_k} \{\Phi(X_{ik}^T \beta_k)\}^{Y_{ik}} \{1 - \Phi(X_{ik}^T \beta_k)\}^{1-Y_{ik}}$;

8       $\mu_{lk}^{(t)} \sim p(\mu | \beta_k^{(t)}, D_i) \propto N \left(\frac{\frac{m_0}{v_0} + \frac{\sum_{j=1}^{n_k} X_{lj}}{\phi_{lk}^{(t-1)}}}{\frac{1}{v_0} + \frac{n_k}{\phi_{lk}^{(t-1)}}}, \sqrt{\frac{1}{\frac{1}{v_0} + \frac{n_k}{\phi_{lk}^{(t-1)}}}} \right), l = 1, ..., p$;

9       $\phi_{lk}^{(t)} \sim p(\phi | \beta_k^{(t)}, D_i) \propto IG \left(g_0 + \frac{n_k}{2}, \frac{1}{2} \sum_{j=1}^{n_k} \left(X_{lj} - \mu_{lk}^{(t)} \right) + b_0 \right), l = 1, ..., p$;

10   **end**

11   Update cluster memberships $g_{1:n}$;

12   **for** $i \leftarrow 1$ **to** n **do**

13   Update cluster membership with a probability $\frac{n_k(g_{-i})}{\alpha + i - 1} P^* \left(D_i | \theta_{i-1}^{(t)} \right) Q(g^{(t)}, A)$ for $k \in K^{(t)}$ or to the newly proposed cluster with a probability $\frac{\alpha}{\alpha + i - 1} P^* \left(D_i | \theta_0^{(t)} \right) Q(g^{(t)}, A)$ where Q is an adjacency matrix;

14   **end**

15 **end**

---

2 Simulation studies

We conducted an additional simulation study to augment our comparative analyses. For every simulation, ground truths were created according to the specifications outlined in Section 3.1 of the main text, under settings 1 and 2. In each scenario, 100 simulated datasets were generated. For setting 1, we have random cluster membership ($c1$, $c2$, and $c3$) generated by 'monte' function in R package 'fungible' (Waller et al., 2024) with η values of (0.9, 0.7, 0.7), a set of random covariates and its probit response (v_1, v_2, v_3, and z), and ground truth coefficients (b_1, b_2, b_3) where b_1 and b_2 are fixed and $b_3 \sim Normal(3, 1)$ for each covariate. The value η close to 1 indicates higher separations between clusters. For generating z with the probit model, we let (b_1, b_2, and b_3) to be (0.8, -1.4, b_3), (2.5, 1.2, b_3) and (-1.3, 0.8, b_3) for $c1$, $c2$ and $c3$. We refer more details to the provided code online. For setting 2, we have fixed cluster membership, a set of random covariates and its probit response, and fixed ground truth coefficients. We gen-

erate b_1 from $N(1, 2)$, b_2 from $N(5, 2)$, b_3 from $N(0, 1.5)$ for cluster $c1$, b_1 from $N(5, 2)$, b_2 from $N(1, 2)$, b_3 from $N(0, 1.5)$ for cluster $c2$, and b_1 from $N(0, 2)$, b_2 from $N(0, 2)$, b_3 from $N(5, 1.5)$ for cluster $c3$. Illustrations of the generated ground truth patterns and their corresponding clustering results can be found in Figures S1 and S2. It is important to note that these figures offer only a snapshot of the simulation, providing a glimpse of the shapes of clusters and their corresponding assignments. Performance indices are summarized (average) in Table S1. In line with our exploration in Section 3.3, metrics like silhouette, entropy, adjusted rand index (ARI), and Calinski-Harabasz index (CHI) do not consistently indicate the superiority of any particular method. Nevertheless, accuracy (rate of correct predictions of clusters) consistently demonstrates the favorable performance of RDMDP methods overall.
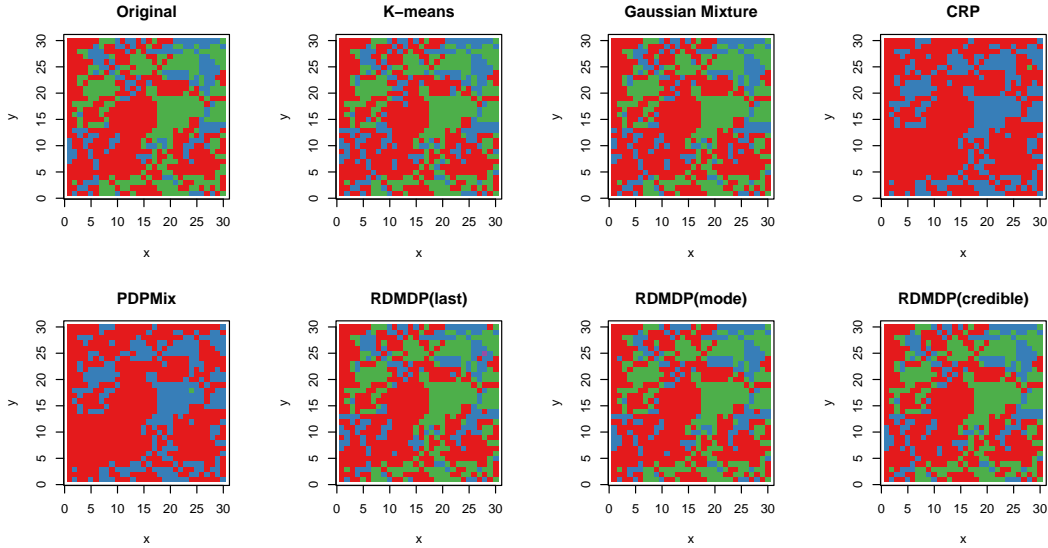

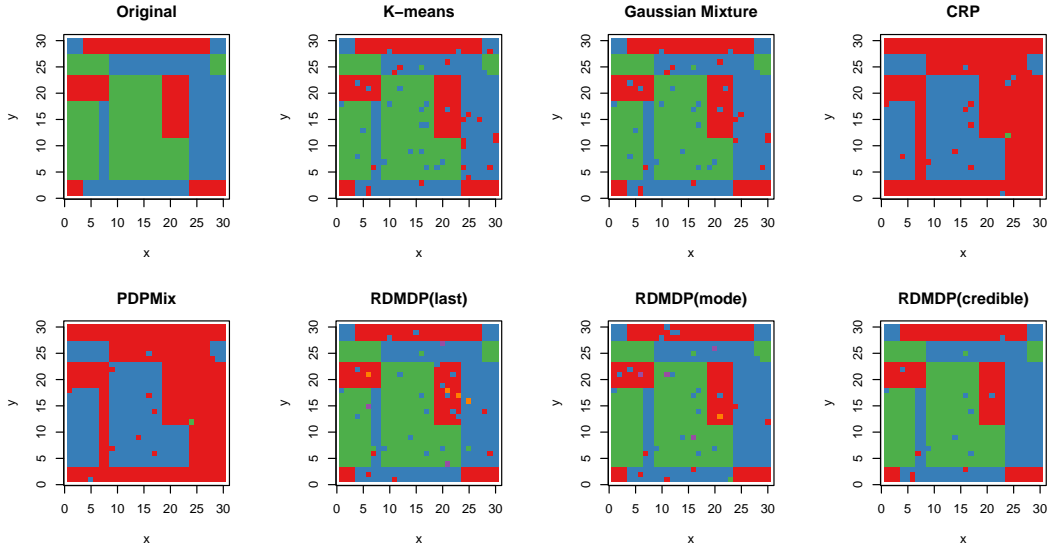Figure S1: A simulation example from Setting 1.

Figure S2: A simulation example from Setting 2.

3 Connection between cluster assignment and the regions of interest

Table S2 presents a depiction of the relationships between the cluster assignments in Figure 3 and the corresponding brain segments. The red cluster in Figure 3 is primarily influenced by the cerebellum (18.15%) but it splits into two clusters colored red and blue, making up 36.39% in slice 2 and 35.27% in slice 3 (Table S2). This suggests that the cerebellum may exhibit different characteristics depending on its location and relationship with adjacent clusters. Additionally, the clusters in the middle section, between slices 5 and 8, involve multiple ROIs in a complex manner. Although this may necessitate cautious interpretation, it is worth noting that the hippocampus (9.47% in the blue cluster at slice 6) and the thalamus (22.77% in the blue cluster at slice 8) constitute a relatively higher percentage, in contrast to just white matter, gray matter, and cerebellum, which are commonly dominant in all slices. In recent findings, it has come to light that both the brain's subcortex and cerebellum may serve as crucial functional brain nodes. This has drawn attention to the importance of the functional network organization that includes the subcortex and cerebellum (Seitzman et al., 2020). Slice 13 is predominantly composed of gray matter (16.44%) and white matter (11.89%). The identified regions could be linked to the dorsal frontoparietal network, consisting of the intraparietal sulcus and frontal eye fields. Age may be related to changes in dorsal frontoparietal cognitive control network showing reduction in an activity related to prospective memory (Lamichhane et al., 2018). Notably, our cluster analysis grouped the cerebellum, which is suggested to be part of this network (Brissenden et al., 2018), into a single cluster (depicted as the red cluster in Figure 3). Additional analyses, such as measurements of gray matter volume or cortical thickness, could provide a more comprehensive and confirmatory understanding of structural changes in the overall brain tissue.

4

Setting 1	Accuracy	ARI	CHI	Silhouette	Entropy
K-means	0.9467	0.8599	1501.0025	0.4178	0.1852
Gaussian Mixture	0.9566	0.8876	1492.0236	0.4193	0.1577
CRP	0.8111	0.6279	1299.0725	0.5101	0.0699
PDPMix	0.8122	0.6289	942.2243	0.8122	0.0646
RDMDP(l)	0.9578	0.8877	Inf	0.3736	0.1421
RDMDP(m)	0.9633	0.9035	975.3994	0.3922	0.1210
RDMDP(c)	0.9744	0.9326	1471.0668	0.4142	0.1043
Setting 2	Accuracy	ARI	CHI	Silhouette	Entropy
K-means	0.9556	0.8754	1239.7903	0.4244	0.1676
Gaussian Mixture	0.9622	0.8927	1237.4805	0.4258	0.1497
CRP	0.7633	0.6350	1267.7842	0.4901	0.0815
PDPMix	0.7622	0.6331	721.0106	0.4720	0.0762
RDMDP(l)	0.9644	0.8937	Inf	0.3473	0.1319
RDMDP(m)	0.9689	0.9111	806.8969	0.3869	0.1126
RDMDP(c)	0.98	0.9429	1218.2551	0.4244	0.0895

Table S1: Clustering performance measures of two simulation settings.

4 R Code

R code is available for the simulation using RDMDP method in online supplementary.

References

Brissenden JA, Tobyne SM, Osher DE, Levin EJ, Halko MA, Somers DC (2018). Topographic cortico-cerebellar networks revealed by visual attention and working memory. *Current Biology*, 28(21): 3364–3372.

Lamichhane B, McDaniel MA, Waldum ER, Braver TS (2018). Age-related changes in neural mechanisms of prospective memory. *Cognitive, Affective, & Behavioral Neuroscience*, 18: 982–999.

Seitzman BA, Gratton C, Marek S, Raut RV, Dosenbach NU, Schlaggar BL, et al. (2020). A set of functionally-defined brain regions with improved representation of the subcortex and cerebellum. *Neuroimage*, 206: 116290.

Waller N, Kracht J, Jones J, Giordano C, Nguyen HV (2024). *Package 'fungible'*. R package version 3.5.
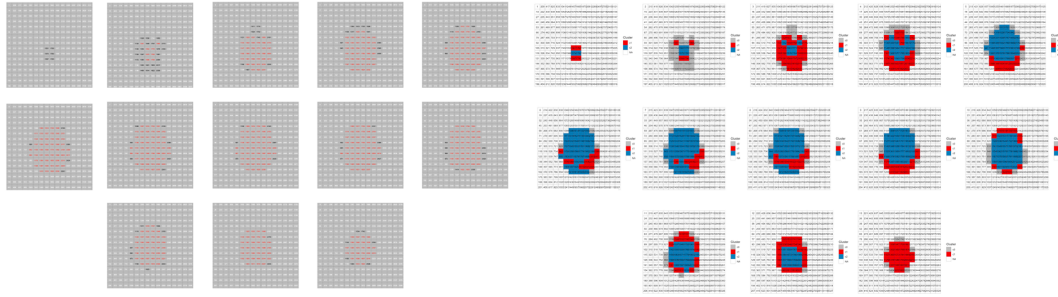
Figure S3: White matter map (left) and RDMDP Clustering on white matter cubes (right). Clusters w1 and w2, according to Table 3, are colored red and blue. Cubes in gray are not used for clustering because they contain fewer than 10 subjects. Each slice, arranged from left to right and bottom to top, depicts a series of two-dimensional representations of cubes spanning from the bottom to the top of the brain.
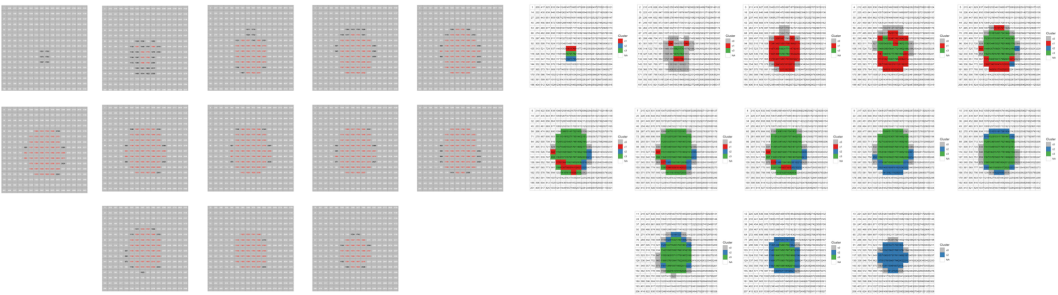

Figure S4: White matter map (left) and K-means Clustering on white matter cubes (right).

	c1 = Red	c2 = Blue	c3 = Green
Slice 1	Cerebellum 18.15% White matter 9.16%	White matter 6.3% Cerebellum 3.72% Gray matter 0.29%	
Slice 2	Cerebellum 29.55% White matter 2.77%	Cerebellum 36.39% White matter 17.52% Gray matter 9.24%	Cerebellum 8.73% White matter 7.63% Gray matter 2.2%
Slice 3	Cerebellum 47.54% White matter 18.83% Gray matter 4.61%	Cerebellum 35.27% Gray matter 8.67% White matter 8.14%	Gray matter 15.89% White matter 9.33% Cerebellum 8.14% Amygdala 0.68%
Slice 4	Cerebellum 34.06% Gray matter 13.21% White matter 4.00%	Cerebellum 51.51% Gray matter 16.14% White matter 14.32%	Gray matter 20.83% Cerebellum 20.69% White matter 15.04% Hippocampus 3.03%
Slice 5	Gray matter 20.71% White matter 12.10% Cerebellum 6.07%	Cerebellum 33.68% Gray matter 24.12% White matter 21.57% Hippocampus 3.50%	Gray matter 2.23% Whiter matter 20.98% Hippocampus 6.87% Amygdala 5.19% Globus Pallidus 5.18%
Slice 6	Gray matter 28.10% White matter 26.05%	White matter 39.10% Gray matter 23.46% Cerebellum 15.27% Hippocampus 9.47% Thalamus 5.32%	White matter 26.13% Gray matter 23.52% Cerebellum 13.96% Putamen 5.97% Amygdala 5.85%
Slice 8	White matter 28.06% Gray matter 26.44% Thalamus 1.13% Hippocampus 0.95% Putamen 0.76%	White matter 45.20% Gray matter 25.76% Thalamus 22.77% Putamen 8.06% Amygdala 6.08%	White matter 35.56% Gray matter 23.76% Caudate 13.10% Putamen 8.66% Amygdala 6.47%
Slice 13	Gray matter 16.44% White matter 11.89%		

Table S2: Top 5 ROIs. Three clusters (c1, c2, c3) from RDMDP in Figure 4 are used in this brain segmentation analysis using 23 ROIs by 'BrainSeg' in DSIStudio.