A spatial Bayesian semiparametric mixture model for positive definite matrices with applications in diffusion tensor imaging

Zhou LAN, Brian J. REICH, and Dipankar BANDYOPADHYAY

1Center for Outcomes Research and Evaluation, Yale School of Medicine, U.S.A.
2Department of Statistics, North Carolina State University, U.S.A.
3Department of Biostatistics, Virginia Commonwealth University, U.S.A.

Abstract: Studies on diffusion tensor imaging (DTI) quantify the diffusion of water molecules in a brain voxel using an estimated $3 \times 3$ symmetric positive definite (p.d.) diffusion tensor matrix. Due to the challenges associated with modelling matrix-variate responses, the voxel-level DTI data are usually summarized by univariate quantities, such as fractional anisotropy. This approach leads to evident loss of information. Furthermore, DTI analyses often ignore the spatial association among neighbouring voxels, leading to imprecise estimates. Although the spatial modelling literature is rich, modelling spatially dependent p.d. matrices is challenging. To mitigate these issues, we propose a matrix-variate Bayesian semiparametric mixture model, where the p.d. matrices are distributed as a mixture of inverse Wishart distributions, with the spatial dependence captured by a Markov model for the mixture component labels. Related Bayesian computing is facilitated by conjugacy results and use of the double Metropolis–Hastings algorithm. Our simulation study shows that the proposed method is more powerful than competing non-spatial methods. We also apply our method to investigate the effect of cocaine use on brain microstructure. By extending spatial statistics to matrix-variate data, we contribute to providing a novel and computationally tractable inferential tool for DTI analysis.

Résumé: Les études d’imagerie de diffusion par tenseurs (IDT) permettent de quantifier la diffusion des molécules d’eau dans les voxels du cerveau grâce à l’estimé d’un tenseur de diffusion : une matrice symétrique définie positive (d.p.) $3 \times 3$. Compte tenu des défis liés à la modélisation d’une variable réponse matricielle, l’IDT au niveau des voxels est habituellement résumée par des variables univariées telle que l’anisotropie fractionnelle (AF), ce qui mène à une perte d’information. De plus, les analyses d’IDT ignorent souvent l’association spatiale entre les voxels voisins, ce qui conduit à des estimateurs imprécis. Malgré une littérature riche en modélisation spatiale, la modélisation de matrices d.p. constitue un grand défi. Les auteurs proposent un modèle de mélange semi-paramétrique bayésien où des matrices d.p. sont distribuées selon une loi de Wishart, et dont la dépendance spatiale est capturée par un modèle de Markov pour l’appartenance aux composantes du mélange. Les calculs associés à ce modèle bayésien exploitent des résultats liés aux lois conjuguées et un algorithme de Metropolis-Hastings double. Les auteurs présentent une étude de simulation qui montre que la méthode proposée est plus puissante que les approches non spatiales. Ils utilisent également leur méthode afin d’étudier les effets de la cocaïne sur les microstructures.
1. INTRODUCTION

Diffusion tensor imaging (DTI), a popular magnetic resonance imaging technique, characterizes the microstructural changes in the brain by measuring the diffusion process of water molecules (Soares et al., 2013). The diffusion process in the brain reflects interactions with many obstacles, such as fibres, thereby revealing microscopic details about the underlying tissue architecture. Although there are several possible ways to characterize this diffusion process [e.g., fractional anisotropy (FA), diffusion tensor (DT), diffusion weighted signals], we focus on the DT, a $3 \times 3$ symmetric positive definite (p.d.) matrix estimated at each voxel. This DT matrix can be interpreted as the covariance matrix of a three-dimensional (3-D) Gaussian distribution, modelling the local Brownian motion of the water molecules (Schwartzman, Mascarenhas & Taylor, 2008). An important clinical objective of analyzing DTI data is to detect regions of local white matter in the brain that differ between two groups (e.g., normal versus diseased), thereby revealing anatomical structural differences (Lo et al., 2010). For example, the motivating data for this article come from a clinical DTI study (Ma et al., 2017), where the scientific objective is to detect regions that differentiate between cocaine users and non-users.

From the framework of modelling matrix-variate responses, the statistical analysis of voxel-level DTI data is challenging. However, there seems to be continued interest in exploring this direction (Schwartzman, Mascarenhas & Taylor, 2008; Zhu et al., 2009; Yuan et al., 2012; Lin, Kong & Sun, 2019; Kong et al., 2020; Hu, Kong & Shen, 2021). In this vein, a widely adopted procedure involves projecting the DTs into a scalar quantity, such as the FA distributed in $(0, 1)$, characterizing the degrees of anisotropy of a diffusion process and then modelling this voxel-level scalar. This simplification results in the loss of statistical information. For example, different p.d. matrices may produce the same FA value (Ennis and Kindlmann, 2006), creating a roadblock in capturing the matrix-level heterogeneity. Conversely, if matrix-variate methods and related regression tools could be implemented and were scalable, they would be obviously attractive and would provide the key to avoiding information loss. There are relatively few matrix-variate methods available to analyze DTI data, and they can be broadly classified into either the (inverse) Wishart matrix models (Lee & Schwartzman, 2017) or the random ellipsoid (RE) models (Schwartzman, Mascarenhas & Taylor, 2008). However, these voxel-level models involve an important shortcoming. Previous studies have revealed that the effect of the disease at proximally located/neighbouring voxels can be similar (see Figure 1), with recommendations to incorporate this non-negligible spatial association to achieve an efficient and valid inference (Spence et al., 2007; Martín-Fernández, Westín & Alberola-López, 2004; Wu et al., 2013; Xue, Bowman & Kang, 2018). Despite this drawback, the aforementioned models do not quantify information from neighbouring voxels. This deficiency motivates us to develop an improved spatial statistical model that (a) utilizes full matrix information, (b) captures spatial dependence and (c) can be implemented through fast and elegant computing.

The spatial neuroimaging toolbox for univariate modelling is well equipped. Woolrich et al. (2004b) proposed a fully Bayesian model for spatiotemporal data, while Kang et al. (2011) implemented spatial point processes for meta-analysis. To select essential biological features, Musgrove, Hughes & Eberly (2016) explored Bayesian variable selection. Recently, Reich et al. (2018) proposed spectral methods combined with Gaussian processes (GPs) that provide computational benefits. All of these studies demonstrated an improvement in the precision of estimates by properly accounting for spatial dependence. In this vein, the Potts model (Wu, 1982), a generalization of the Ising model in statistical mechanics, has also been successfully applied.

The Canadian Journal of Statistics / La revue canadienne de statistique DOI: 10.1002/cjs.11601
FIGURE 1: Colour-coded neuronal tracts from DTI of a human brain. Positive definite matrices visualized as ellipsoids are produced for each voxel, revealing the anatomical structure of the brain. The ellipsoids are generated using the visualization software BrainSuite (http://brainsuite.org/). The colours red, green and blue represent the $x$ (frontal), $y$ (sagittal) and $z$ (transverse) directions, respectively.

to imaging (Johnson et al., 2013). A desirable property of the Potts model is that it avoids smoothing over abrupt changes in the image intensity (Johnson et al., 2013), thereby making the Potts model more attractive than the available Gaussian kernel methods.

To this end, we assume the p.d. DTs follow a mixture of inverse Wishart distributions, with the mixture component labels modelled via a (spatial) Potts model, representing a discrete Markov random field. This *semiparametric mixture* specification refers to a class of flexible mixture distributions with a finite number of components (Lindsay & Lesperance, 1995). In addition to spatial modelling, another important topic in neuroimaging is the problem of detecting regions of differences between two groups. Previous investigations involved voxel-wise hypothesis testing (Schwartzman, Mascarenhas & Taylor, 2008; Lee & Schwartzman, 2017). An alternative option involves constructing multilevel hierarchical modelling, accounting for both subject-level and group-level variation, and then using the group-level parameters for voxel-wise hypothesis testing (Woolrich et al., 2004a; Liu et al., 2014). In this article, we use the latter approach by extending the latent Potts model into a hierarchical *two-way* framework. This formulation facilitates hypothesis testing via group-level parameters and inter-subject variability simultaneously. Our proposal utilizes the Bayesian approach, accounting for the uncertainty of model parameters at all levels of the hierarchy. However, the Bayesian approach in neuroimaging involves a heavy computational burden (Cohen et al., 2017). Although the associated Markov chain Monte Carlo (MCMC) algorithm is mostly composed of computationally tractable Gibbs steps that can be implemented in parallel, a major drawback of the Potts model is the intractable normalizing constant that creates a bottleneck for hyperparameter updates. In this article, we follow the suggestion of Park & Haran (2018) and resolve this challenge via the double Metropolis-Hastings (DMH) sampler (Liang, 2010).

In Section 2 of the article, we first introduce the modelling framework for a single-subject, and then extend it to multiple subjects, and the group hypothesis testing framework. Relevant MCMC computational details are summarized in Section 3. To demonstrate the improvement in performance compared to plausible alternatives, we report the results of two simulation studies in Section 4. In Section 5, we describe the application of our method to the motivating cocaine data. Section 6 consists of some concluding remarks and possible directions for future research.
2. MODEL

In this section, we first introduce the single-subject Bayesian semiparametric mixture model and then extend it to the multi-subject model.

2.1. Single-subject Model

Let \( A_v \) be the \( p \times p \) DT at voxel \( v \in \{1, 2, \ldots, n\} \). To ensure \( A_v \) is symmetric and p.d., it is usually parameterized as a (inverse) Wishart matrix (Lee & Schwartzman, 2017) or a Gaussian symmetric matrix-variate distribution (Schwartzman, Mascarenhas & Taylor, 2008). In this article, we assume that \( A_v \) follows an inverse Wishart distribution as

\[
A_v | M_v, m \overset{\text{iid}}{\sim} \mathcal{IW}_p(M_v, m),
\]

where \( \mathcal{IW}_p(M_v, m) \) is the inverse Wishart distribution parameterized to have mean \( M_v \) (i.e., \( \mathbb{E} A_v = M_v \), see Appendix A) and degrees of freedom \( m > p + 1 \), and the DTs are independently distributed across \( v \) given the mean matrices \( M_v \) and the degrees of freedom \( m \). The mean matrices are modelled as a finite mixture of Wishart distributions, denoted as \( [M_v | g_v = k] := V_k \), where \( g_v \in \{1, 2, \ldots, K\} \) is the latent cluster label. The prior of \( V_k \) is \( V_k \overset{i.i.d.}{\sim} \mathcal{IW}_p(\Sigma, \nu) \), where \( \mathcal{IW}_p(\Sigma, \nu) \) is the Wishart distribution parameterized (see Appendix A for details) to have mean \( \Sigma \) and degrees of freedom \( \nu > p \).

Spatial dependence among the DTs is induced through the dependence of the mean matrices \( M_v \). We assume the latent cluster labels follow a weighted Potts model, specified via the full conditional distribution

\[
\mathbb{P}_k = P(g_v = k | \beta, \eta_k, g_{-v}) \propto \exp \left[ \eta_k + \beta \sum_{u \in N_v} \mathcal{I}(g_u = k) \right],
\]

where \( g_{-v} \) is the full set \( g = \{g_1, g_2, \ldots, g_n\} \) excluding \( g_v ; N_v \) is a set of indices of the neighbouring voxels of \( v \); and \( \mathcal{I}[\cdot] = 1 \) if event \( \cdot \) is true, and 0 otherwise. Given \( g_{-v} \) but marginal over \( g_v \), the distribution of \( A_v \) is the mixture of \( K \) inverse Wishart distributions

\[
\sum_{k=1}^{K} \mathbb{P}_k \mathcal{IW}_p(A_v | V_k, m),
\]

where \( \mathcal{IW}_p(A|V, m) \) is the inverse Wishart density function of \( A \) with mean matrix \( V \) and the degrees of freedom \( m \). This semiparametric mixture model spans a rich class of density functions.

Through the Potts model, an image can be considered a network whose nodes are the voxels. In this network, every voxel is connected to its neighbouring voxels (the nearest six voxels when applicable). The full conditional distribution of \( g_v \) depends only on the voxels in the neighbouring set \( N_v \), and therefore, the process is Markovian. The spatial parameter \( \beta \), the coefficient of the neighbouring term \( \sum_{u \in N_v} \mathcal{I}(g_u = k) \), quantifies the dependence on the neighbouring voxels. Unlike the classic Potts model (Wu, 1982), the terms \( \eta_k \) are added as offset terms that control the overall mass allocated to each cluster. We set \( \eta_k = -k^2 \xi \), such that \( \xi \geq 0 \) is the concentration parameter controlling the homogeneity of the latent cluster labels. It is problematic to pre-specify the number of components \( K \) in a mixture model (McCullagh & Yang, 2008). However, the offset terms allocate more weight to the key components and less weight to the trivial components. We fit the model by setting \( K \) to be an upper bound on the number of active clusters and allow the data to determine the number of active clusters via estimation of \( \xi \). If \( \xi \to 0 \), there are several active clusters, whereas if \( \xi \) is large, there are a few active clusters. As a result, the model is
less sensitive to the number of components $K$ when the offset term $\eta_k$ is included. This claim is verified in the simulation studies (see Section 4) and the illustrative example discussed in Section 5, where we obtain similar results for different values of $K$.

### 2.1.1. Spatial dependence

Quantifying spatial dependence is a vital issue in spatial statistics and neuroimaging. As this model is used to analyze matrix-variate data, we use the expected squared Frobenius norm to measure dependence. The spatial dependence between matrices $A$ and $B$ can be summarized as $E ||A - B||_F^2 = E \text{Tr}[(A - B)^T(A - B)]$. The norm increases as dependence decreases. If $A$ and $B$ are $1 \times 1$, the expected squared Frobenius norm is the classic variogram (Cressie, 1993), the cornerstone of spatial statistics. In this regard, the expected squared Frobenius norm can be treated as the variogram for matrix-variate data. In the rest of the article, we simply refer to the expected squared Frobenius norm as the variogram. For the Potts model described above, the corresponding variogram is

$$V(u, v) = E ||A_u - A_v||_F^2 = \gamma(m, v, \Sigma)P(u, v|\beta, \xi),$$

where $P(u, v|\beta, \xi)$ is the marginal (over all other cluster labels $g$) probability of $g_u \neq g_v$, and $\gamma(m, v, \Sigma)$ is a measure of the variability in $A_v|M_v$ and the variability of $V_k$ across $K$. Therefore, the multivariate spatial dependence structure is separable (Cressie & Wikle, 2015), in that the dependence is the product of a non-spatial term $\gamma(m, v, \Sigma)$ that controls cross dependence and a spatial term $P(u, v|\beta, \xi)$ that controls spatial dependence. The full expression for $\gamma(m, v, \Sigma)$ is available in Appendix B. When $p = 3$ and $\Sigma = I$, the non-spatial term is $\frac{12(m + v - 4)(2m - 7)}{\nu(m - 3)(m - 6)}$, where $m > 6$ and $v > 3$. Therefore, in this special case, the cross dependence decreases if $m$ or $v$ is larger (see Figure 2).

The spatial term $P(u, v|\beta, \xi)$ is intractable, prompting us to use a Monte Carlo approximation to study the function. In Figure 3, the function is computed under the scenario that the image is a one-dimensional (1-D) grid with $K = 100$, and $\xi = 0$. We observe that $P(u, v|\beta, \xi)$ increases with distance, and a larger $\beta$ leads to stronger spatial dependence. In Figure 3, we also display the functional values for varying $K$. Using extreme value theory (Reich & Shaby, 2019), we have

![Figure 2: The density plot of $\gamma(m, v, \Sigma)$ when $p = 3$ and $\Sigma = I$.](image)
\[ P(u,v|\beta,\xi) \]

Figure 3: Monte Carlo approximation of the spatial term \( P(u,v|\beta,\xi) \). The left panel is with \( K=100 \) and the right panel with \( \beta=10 \). The value varies depending on the distance \( |u-v| \), the number of clusters \( K \) and the spatial parameter \( \beta; \xi=0 \).

\[ \lim_{|u-v|\to\infty} P(u,v|\beta,\xi) = 1 - \frac{1}{K}. \]

Thus, increasing \( K \) leads to smaller spatial dependence. Hence, we fix \( K \) to be large to eliminate long-range dependence (i.e., \( P(u,v|\beta,\xi) < 1 \) for large \( |u-v| \)) and estimate \( \beta \) to capture local dependence.

As one reviewer pointed out, various other (non-Euclidean) choices for the dependence metric exist (Arsigny et al., 2007), such as the Log-Euclidean metric. However, under our current framework, the Log-Euclidean distance is also separable, such that \( E||\log A_u - \log A_v||^2_F = \eta(m,v,\Sigma)P(u,v|\beta,\xi) \), where \( \eta(m,v,\Sigma) \) is another non-spatial term and \( P(u,v|\beta,\xi) \) is a spatial term, as we previously observed in Equation (3). For full details of this derivation, see Appendix C. In reality, we posit that the choice between the Euclidean (our Frobenius norm) and Log-Euclidean metric results in a trivial impact, but our choice avoids significant computing challenges and, in particular, working with the second moment of the log-inverse Wishart matrix.

For a more intuitive understanding of this model, we simulated the DTs, which can be visualized as ellipsoids in a \( 40 \times 40 \) grid (Figure 4, right panel). In these simulations, we used \( \xi=0, m=4 \) and \( \nu=30 \). From Figure 4 (left panel), the DTs within the same latent cluster label are similar to each other and indicative of the within-cluster spatial dependence. In Figure 5, a larger \( \beta \) leads to more distinct realization of the spatial dependence. These figures also illustrate that the Potts model allows sharp breaks, which is desirable if neighbouring voxels are present in different tracts.

2.2. Multi-subject Model

We now extend the single-subject model to the multi-subject setting. Through this statistical formulation, we refine our clinical objective of detecting brain regions where the DT distribution across the subjects differs between the cocaine users and non-users.

Let \( A_{iv} \) be the DT and \( g_{iv} \) the cluster label for voxel \( v \in \{1, 2, \ldots, n\} \) for subject \( i \in \{1, 2, \ldots, N\} \). By extending \( g_v \) to \( g_{iv} \), the subject-level cluster labels not only model intra-subject spatial dependence but also allow inter-subject variability. As in the single-subject model, the DTs are conditionally independent given the random matrices \( M_{iv} \) and follow the finite mixture model now extended to multiple subjects:

\[ A_{iv}|M_{iv},m \sim \mathcal{I}(\mathcal{W}_p(M_{iv},m)), \quad M_{iv} := V_{g_{iv}}, \quad V_k \sim \mathcal{W}_p(\Sigma, \nu). \]

Let \( x_i \) denote the binary group indicator of subject \( i \), such that \( x_i = 1 \) indicates a cocaine user and \( x_i = 0 \) identifies a non-user. To model intra-subject spatial dependence within a group, we extend the latent cluster formulation by introducing the group-level cluster labels \( h_{iv} \) for group
FIGURE 4: DT illustration: The left panel represents the latent cluster labels $g_i$, where each colour denotes a distinct latent cluster label. The right panel displays the corresponding simulated DTs $A_i(s)$.

FIGURE 5: From left to right, the panels display the simulated DTs under the models with $\beta = 1, 2$ and $6$; $K = 20; \xi = 0$.

$x \in \{0, 1\}$ and voxel $v \in \{1, 2, \ldots, n\}$. Both $h_{xv}$ and $g_{iv}$ are also spatially dependent, with full conditional distributions

$$P(g_{iv} = k | \alpha, \beta, \xi, g_{(-iv)}, h) \propto \exp \left[ -k^\xi + \beta \sum_{u \in N_v} \mathcal{J}(g_{iu} = k) + \alpha \mathcal{J}(h_{xv} = k) \right]$$

$$P(h_{xv} = k | \alpha, \beta, h_{(-xv)}, g) \propto \exp \left[ \beta \sum_{u \in N_v} \mathcal{J}(h_{xu} = k) + \sum_{j, y \neq x} \alpha \mathcal{J}(g_{jv} = k) \right], \quad (5)$$

where $g_{(-iv)}$ is the set on $g_i = \{g_1, \ldots, g_n\}$ excluding $g_{iv}$, $h_{(-xv)}$ is the set $h_i = \{h_1, \ldots, h_n\}$ excluding $h_{xv}$, $g$ is the set on $\{g_1, \ldots, g_N\}$ and $h$ is the set on $\{h_0, h_1\}$. Contrary to the single-subject model, the group-clustering parameter $\alpha$ is introduced for modelling multiple subjects. If $\alpha = 0$, $A_{iv}$ is independently distributed over subjects; otherwise, the subject-level cluster label $g_{iv}$ depends on the group-level cluster label $h_{xv}$, leading to smaller inter-subject variability of the spatial dependence pattern within one group. The joint probability mass function (PMF) of $\{g_1, g_2, \ldots, g_N\} \cup \{h_0, h_1\}$ is given in Appendix D. As the conditional densities specified in Equation (5) satisfy the conditions of the Hammersley–Clifford theorem (Clifford, 1990), the

DOI: 10.1002/cjs.11601 The Canadian Journal of Statistics / La revue canadienne de statistique
FIGURE 6: The graphical representation of the latent cluster labels is displayed. The group-level cluster labels are \( h_x = \{ h_{x1}, \ldots, h_{xn} \} \), and the subject-level cluster labels are \( g_i = \{ g_{i1}, \ldots, g_{in} \} \). Cluster labels \( h_x \) and \( g_i \) are mutually dependent. The subject-level cluster labels \( g_i \) have inter-subject variability. The group-level cluster labels \( h_x \) summarize the spatial dependence of all subjects.

The existence of the joint distribution of \( \{ g_1, g_2, \ldots, g_N \} \cup \{ h_0, h_1 \} \) is guaranteed; see Appendix D for details. A graphical representation of this latent Potts model is provided in Figure 6, where cluster labels \( g_i \) and \( h_x \) characterize the spatial pattern of subject \( i \) and the general spatial pattern of subjects in group \( x \), respectively.

To further understand the role of \( \alpha \) and \( h_{xv} \), we inspect the density of \( A_{iv} \) conditioned on \( h_{xv} \) and marginal over all other labels (see Appendix E). The conditional density of \( A_{iv} \) given \( x_i = x \) and \( h_{xv} \) is a mixture of inverse Wishart densities, proportional to

\[
\sum_{k=1}^{K} \exp \left[ -k^2 + \alpha \mathcal{I} (h_{xv} = k) \right] \mathcal{I} \mathbb{W} (A_{iv} | V_k, m),
\]

where the term \( \exp \left[ -k^2 + \alpha \mathcal{I} (h_{xv} = k) \right] \) is the weight corresponding to cluster \( k \), and \( \mathcal{I} (h_{xv} = k) \) is the indicator function evaluating the mass on the mixture component \( k \) at voxel \( v \) for all subjects with \( x_i = x \). Assuming \( \alpha \geq 0 \), the density identified in Equation (6) depends on \( x_i \) if and only if \( h_{0v} \neq h_{1v} \).

Now, the clinical objective of finding regional differences between two groups can be formulated as the problem of finding voxels for which the distribution of \( A_{iv} \) is different for \( x_i = 0 \) or \( x_i = 1 \). This reduces to the testing problem

\[
\mathcal{H}_{ov} : h_{0v} = h_{1v} \\
\mathcal{H}_{av} : h_{0v} \neq h_{1v}.
\]

Our Bayesian inference will provide estimates of the posterior probabilities of the hypotheses and is discussed more fully in Section 3.

2.2.1. Spatial dependence

To quantify spatial dependence within and across subjects, we propose the variogram

\[
\mathcal{V}_{ij}(u, v) = E ||A_{iu} - A_{jv}||^2_F = \gamma(m, \nu, \Sigma) P_{ij}(u, v|\alpha, \beta, \xi),
\]

where \( \gamma(m, \nu, \Sigma) \) is the non-spatial term (discussed in Section 2.1), and the spatial term \( P_{ij}(u, v|\alpha, \beta, \xi) \) is the marginal probability of \( g_{iu} \neq g_{jv} \). In Equation (8), \( i = j \) leads to the individual variogram, while \( i \neq j \) corresponds to the inter-subject variogram, and both are
FIGURE 7: Monte Carlo approximation of the spatial term $P_{ij}(u, v|\alpha, \beta, \xi, 0)$ for the individual (left panel) and within-group (middle panel) and between-group (right panel) variograms, plotted as a function of the distance $|u - v|$, the group-clustering parameter $\alpha$ and the spatial parameter $\beta$.

separable (Cressie & Wikle, 2015). For the inter-subject variogram, we can also compare the within-group variogram for subjects with $x_i = x_j$ and the between-group variogram for subjects with $x_i \neq x_j$.

In Figure 7, we plot $P_{ij}(u, v|\alpha, \beta, \xi)$ as a function of the distance $|u - v|$, obtained via a Monte Carlo approximation, under the scenario that the image is a 1-D grid with $N = 5$ and $K = 100$, for various choices of the group clustering parameter $\alpha$ and the spatial parameter $\beta$. For the individual variogram (left panel), the spatial parameter $\beta$ largely controls the within-subject dependence. For the within-group variogram (middle panel), a larger value of $\alpha$ leads to more dependence in the within-group variogram. Thus, $\alpha$ controls the dependence of subjects within one group. Finally, for the between-group variogram (right panel), as $g_{iu}$ and $g_{jiu}$ are assumed to be independent, the spatial term $P_{ij}(u, v|\alpha, \beta, \xi) = 1 - \frac{1}{K}$ is a constant.

3. COMPUTATION

We use MCMC to fit the model described in Section 2. The source code, written in hybrid R and C++, and example scripts for model implementation can be found at https://github.com/ZhouLanNCSU/Potts_DTI.

The final Potts model is

$$A_{iv} | M_{iv}, m \overset{\text{indep.}}{\sim} \mathcal{F}(\mathcal{W}_p(M_{iv}, m), M_{iv} := V_{g_{iv}}, V_k \overset{i.i.d}{\sim} \mathcal{W}(\Sigma, \nu))$$

$$P(g_{iv} = k|\alpha, \beta, \xi, g_{-iv}, h) \propto \exp \left[ -k\xi + \beta \sum_{u \in N_v} \mathcal{J} (g_{iu} = k) + \alpha \mathcal{J} (h_{xi} = k) \right]$$

$$P(h_{xy} = k|\alpha, \beta, h_{-xy}, g) \propto \exp \left[ \beta \sum_{u \in N_v} \mathcal{J} (h_{xu} = k) + \sum_{j : x_j = x} \alpha \mathcal{J} (g_{jv} = k) \right].$$ (9)

Using the moment method (Robert, 2007, Section 3.2.4), we set $\Sigma$ equal to the sample mean of all observed DTs. The a priori information conveyed by $\Sigma$ has little impact if the number of observations is large. We adopted a uniform prior for the degrees of freedom $m$ and $\nu$ on $[5, 50] \times [4, 50]$. Following Liang (2010), we also chose a uniform prior for $\theta = \{\alpha, \beta, \xi\}$ on $[0, 20] \times [0, 20] \times [0, 1]$, which we denote by $\pi(\theta)$. In what follows, we describe the updating rule for each parameter.

The MCMC algorithm is a combination of Gibbs and Metropolis–Hastings steps. The latent mean matrices and cluster labels are updated via Gibbs steps. Their full conditional distributions are

DOI: 10.1002/cjs.11601
The Canadian Journal of Statistics / La revue canadienne de statistique
DOI: 10.1002/cjs.11601

138

LAN, REICH AND BANDYOPADHYAY
Vol. 49, No. 1

Figure 8: Empirical variograms from the cocaine users’ data; each line corresponds to each subject/pair.
From left to right, the estimates are between-group, within-group and individual variograms.

- $V_k \sim \mathcal{W}_p((\Sigma^{-1})_v + (m - p - 1)\sum_{v' : g_{iv}=k}(A_{iv}^{-1})^{-1}(N_{k}m + \nu), N_{k}m + \nu)$,
- $P(g_{iv} = k|\cdot) \propto \mathcal{I} \mathcal{W}_p(A_{iv}|V_{g_{iv}}, m) \exp \left[ -k^2 + \beta \sum_{u \in N_v} \mathcal{I}(g_{iu} = k) + \alpha \mathcal{I}(h_{iv} = k) \right]$
- $P(h_{iv} = k|\cdot) \propto \exp \left[ \beta \sum_{u \in N_v} \mathcal{I}(h_{iu} = k) + \alpha \sum_{j \neq i} \mathcal{I}(g_{jv} = k) \right]$

where $n_k = \sum_{i,v} \mathcal{I}(g_{iv} = k)$. In addition, $P(g_{iv} = k|\cdot)$ and $P(h_{sv} = k|\cdot)$ can be updated in parallel over $i$ and $x$, respectively. As the uniform prior is not conjugated, we have to sample $[\nu_l]$ and $[\nu_r]$ via Metropolis–Hastings sampling, with log-normal random walk proposals.

To select regions of possible differences via Bayesian hypothesis testing, we reject the null hypothesis identified in Equation (7) if $P(h_{0_v} \neq h_{1_v}|\cdot) > P(h_{0_v} = h_{1_v}|\cdot)$. The fact that the distributions are the same in the two groups if and only if $h_{0_v} = h_{1_v}$ allows us to undertake the hypothesis testing in this manner. Furthermore, the hypothesis $h_{0_v} = h_{1_v}$ averages over the specific labels and thus avoids the label-switching problem (Rodriguez & Walker, 2014). The posterior probabilities of $h_{0_v} = h_{1_v}$ or $h_{0_v} \neq h_{1_v}$ can be estimated through MCMC samples.

Updating the Potts hyperparameters $\alpha$, $\beta$, and $\xi$ is problematic due to the intractability of the normalizing constant in the joint distribution function of the cluster labels; see the PMF identified in Appendix D. A simple approach is to estimate the parameters outside of the MCMC estimation. The plug-in values can be obtained from cross-validation (e.g., Goldsmith, Huang & Crainiceanu, 2014), pseudo-likelihood comparison (e.g., Zhao, Kang & Yu, 2014; Lan et al., 2016) or by comparing empirical and model-based variograms (e.g., Figures 7 and 8). However, these methods fail to account for the uncertainty associated with imputing these parameters. Hence, we resort to using the DMH algorithm (Liang, 2010).

Very recently, Park & Haran (2018) reviewed several Monte Carlo methods for estimating models with intractable normalizing constants and recommended the DMH algorithm proposed by Liang (2010) because of its ease of execution and computational efficiency. The DMH algorithm, combined with usual Bayesian tools, has been found to be effective for the Potts implementation in a variety of applications, such as oral health (Jin, Yuan & Bandyopadhyay, 2016), cancer imaging (Li et al., 2019) and others. The DMH update for $\theta$ is initiated with a candidate $\theta'$ drawn from $q(\theta|\bar{\theta})$, where $\bar{\theta}$ is the current value and $q(\theta|\bar{\theta})$ is a log-normal random walk transitional probability centered at $\bar{\theta}$. Given the candidate $\theta'$, we draw labels $g_{i} = \{g_{i1}', \ldots, g_{im}'\}$ and $h_{x} = \{h_{x1}', \ldots, h_{xm}'\}$ using Gibbs sampling for each $i$ and $x$. The candidate $\theta'$ is accepted with the probability $\min(1, r)$, where $r = \frac{\pi(\theta')\mathcal{P}(g_{i}', h_{x}')}{\pi(\theta)\mathcal{P}(g_{i}, h_{x})}$, with $\mathcal{P}(g, h|\theta)$, the likelihood of $\{g_{1}, g_{2}, \ldots, g_{N}\} \cup \{h_{0}, h_{1}\}$, conditioned on $\theta$, and $g_{i}$ and $h_{x}$ are the current values of $g_{i}$ and
4. SIMULATION STUDIES

In this section, we explore the finite sample performance of our method via two simulation studies, where we generated synthetic data under different scenarios. We compare our method to the RE model of Schwartzman, Mascarenhas & Taylor (2008), which considers a non-spatial Gaussian symmetric matrix modelling for the DTI response to the RE model of Schwartzman, Mascarenhas & Taylor (2008), which considers a non-spatial studies, where we generated synthetic data under different scenarios. We compare our method with the alternative RE model. The accuracy of the ML estimates from the RE model, our fitted model exhibited significantly improved performance in terms of the estimated TPR, FPR and FDR. It is likely that the small number of subjects may have contributed to the low observed TPR for the RE model. The accuracy of the ML estimates from the RE model is dependent on the number of subjects (Schwartzman, Mascarenhas & Taylor, 2008). As the choice of \( K \) does not affect the selection accuracy in the Potts model, the simulation results also support the claim that our model can be less sensitive to the number of clusters if \( K \) is larger than the true cluster size.

Next, to understand the true effect of the spatial term (as pointed out by a reviewer), we fitted additional models, forcing \( \beta \) to small values, such as 1, 3 and 5, with \( K \) fixed to 10. Under this scenario, the estimated FPRs were 0.35, 0.21 and 0, while the estimated TPR values all exceeded 0.99. While the estimated TPR for the RE model was noticeably smaller (0.29; see Table 1, column 5), the FPRs from these new fits (fixing \( \beta \)) were much larger than the corresponding values derived from the RE model. This implies that, if the ground truth data are generated under spatial association (as in the mixture model), our spatial Potts model (with various degrees of spatial association) is more adequate in recovering that association, via a large estimated TPR, compared to the alternative non-spatial RE model. However, as expected, fixing \( \beta \) (and the induced spatial association) instead of estimating it in our fitted Potts model led to a spike in the estimated FPR, with the rate becoming smaller with increasing \( \beta \) (and increased spatial
### Table 1: Simulation results under data generation from mixture models. Estimated values of the true positive rate, false positive rate, false discovery rate and computation time for the Potts and random ellipsoid models are summarized.

| Characteristic | $K = 10$ | $K = 50$ | $K = 100$ | Random ellipsoid |
|----------------|----------|----------|-----------|-----------------|
| TPR            | 0.99     | 0.97     | 0.98      | 0.29            |
| FPR            | 0.013    | 0.010    | 0.009     | 0.00            |
| FDR            | 0.025    | 0.025    | 0.024     | 0.002           |
| Time (h)       | 0.5      | 0.8      | 1.0       | < 0.01          |

### Table 2: Simulation results exploring robustness when the data were generated using a spatial Cholesky process. Estimated values of the true positive rate, false positive rate, false discovery rate and computation time for the Potts and random ellipsoid models are summarized.

| Characteristic | $K = 10$ | $K = 50$ | $K = 100$ | Random ellipsoid |
|----------------|----------|----------|-----------|-----------------|
| TPR            | 0.79     | 0.80     | 0.77      | 0.50            |
| FPR            | 0.01     | 0.01     | 0.01      | 0.00            |
| FDR            | 0.03     | 0.03     | 0.02      | 0.03            |
| Time (h)       | 1.0      | 1.5      | 1.8       | < 0.01          |

We concluded that the spatial term plays a major role in controlling the FPR within our model framework.

To determine robustness in the face of model misspecification, we also simulated data from the spatial Cholesky process as we now demonstrate. The DT matrix for subject $i$ at voxel $v$ is determined by six independent spatial GPs $U_{ivk}$ ($k \in \{1, 2, \ldots, 6\}$). These spatial GPs constitute the lower triangular matrix $L_{iv} = \begin{bmatrix} e^{U_{iv1}} & 0 & 0 \\ U_{iv4} & e^{U_{iv2}} & 0 \\ U_{iv5} & U_{iv6} & e^{U_{iv3}} \end{bmatrix}$. The responses $A_{iv}$ were then constructed as $A_{iv} = L_{iv} L_{iv}^T$, thereby introducing spatial dependence and guaranteeing positive definiteness. We again used a $40 \times 40$ grid with spacing 1 between adjacent grid points as an image. We also chose to use 10 subjects in each of the treatment and control groups. The six spatial GPs were simulated with variance $\tau^2 = 0.1$ and exponential correlation function with range parameter $\rho = 2$. The mean values of the six GPs were all 0, except for a $10 \times 10$ region in the centre of the image for the treatment group subjects, where $U_{ivk}$ had mean 0.5 for $k \leq 3$ and 0.25 for $k > 3$. This feature represented the brain with a small region of difference between the two groups. Once again, we compared the results obtained using our model with $K = 10, 50, and 100$ to the corresponding results derived from the RE model fitted to the same data. The observed results, averaged over 50 simulations, are summarized in Table 2. These estimated values demonstrate that our Potts model is robust to this form of misspecification. In addition, under the spatial dependence assumption, the spatial models exhibited better performance overall than the non-spatial model.

A major drawback of the use of Bayesian methods in neuroimaging is the associated computational burden. In both simulation studies, we were able to implement the Potts model...
within a few hours. On the other hand, the RE model avoids the computational burden of MCMC. The time taken (in hours) to fit each model under the two simulation schemes is reported in Tables 1 and 2. However, in light of the conservative performance exhibited by the RE model, we believe our Potts model represents a reasonable compromise.

5. APPLICATION: COCAINE DEPENDENCE STUDY

We now use our Potts model to analyze the dataset concerning cocaine users (Ma et al., 2017), which we previously described in Section 2. The data were provided by the Institute for Drug and Alcohol Studies, Virginia Commonwealth University. The study recruited 11 cocaine users and 11 controls to explore possible differences in the microstructural changes in the brain via DTI. We focus on the corpus callosum (CC), a large white matter fibre network that connects the two cerebral hemispheres in the brain and plays important roles such as transferring motor, sensory and cognitive information between the left and right hemispheres (Ma et al., 2009). The CC region contains 15,273 voxels. DTI, a powerful non-invasive technique, quantifies the degree of directionality of the diffusion of water molecules in the white matter tissues within the CC via popular parameters (Ma et al., 2009), such as FA, with preference for a single direction. Previous DTI studies of cocaine dependence (Moeller et al., 2005; Lim et al., 2008; Lane et al., 2010) revealed reduced FA (i.e., the diffusion need not be restricted to one direction due to damage of the nerve fibres) and white matter integrity (altered microstructure) consistently in the CC among cocaine users compared to controls. These white matter alterations in cocaine dependence studies are presumed to be related to impulsivity (Moeller et al., 2005), decision-making (Lane et al., 2010), etc. We would now like to re-evaluate these findings in light of DT matrices, using the tools associated with our Potts model.

Prior to any model fitting, we examined the fit of the proposed model to the cocaine users’ data via empirical estimates of variograms. We let \( \hat{\gamma}_{ij}(d) = \frac{1}{N_d} \sum |u - v| = d ||A_{iu} - A_{ju}||_F^2 \) denote the empirical variogram value of subjects \( i \) and \( j \) at distance \( d \), where \( N_d \) is the number of pairs with \( |u - v| = d \). We plotted these empirical variograms in Figure 8. The DTs have a strong within-subject spatial dependence. The empirical within-group variogram also increases with distance, indicating inter-subject dependence within a group; however, the between-group empirical variogram was almost flat, which suggests that the subjects are independent if they belong in different groups. As these empirical variograms perfectly matched the theoretical variograms displayed in Figure 7, the hierarchical Potts model assumptions concerning spatial dependence appear reasonable for these data.

For Bayesian inference, we used 11,000 MCMC samples with 3,000 discarded as burn-in. The computing time typically took 5 h using a CPU with 3.4 GHz Intel Core i5. To study the sensitivity to \( K \), we fit the model with \( K \) equal to 100, 200, 300, 400 and 500 and used the Hubert–Arabie-adjusted Rand index (Rand, 1971; Steinley, 2004) to measure the similarities of regions of detected differences, with varying \( K \). The adjusted Rand index measures clustering similarity: If the two clusters are almost identical, the index is close 1; otherwise, the index is close to 0. As summarized in Table 3, the Rand indices for any two values of \( K \) are close to 1. Hence, the selection is not sensitive to \( K \). For the rest of this section, we use the result corresponding to \( K = 100 \).

We first used the \texttt{R} package \texttt{brainR} (Muschelli, Sweeney & Crainiceanu, 2014) for 3-D visualization of the regions of possible differences. To investigate if the performance of our Potts model is improved by introducing spatial dependence, we also compared our results to estimates derived using the RE model with a confidence level of 0.9. The selected regions of differences are displayed in Figure 9. The adjusted Rand index value of 0.86 implies “good” recovery of clustering similarity. While the Potts model recovered some additional regions of differences (compared to the RE model), the regions detected using the fitted Potts model are spatially contiguous. We observe in Figure 9 that a region of difference was detected in the splenium, a
Table 3: The adjusted Rand index for measuring clustering similarities. The off-diagonals of the table are the estimated Rand indices for the corresponding two values of $K$.

| $K$ | 100 | 200 | 300 | 400 | 500 |
|-----|-----|-----|-----|-----|-----|
| 100 | .   | 0.92| 0.91| 0.92| 0.92|
| 200 | .   | .   | 0.94| 0.95| 0.94|
| 300 | .   | .   | .   | 0.96| 0.95|
| 400 | .   | .   | .   | .   | 0.96|
| 500 | .   | .   | .   | .   | .   |

Figure 9: A 3-D representation of the cocaine dependence data. The regions of possible differences between the cocaine users and non-users are shown in red. The left and right panels indicate that the image was derived from the Potts model and the random ellipsoid model, respectively.

A thick collection of axonal fibres located on the posterior end of the CC; the splenium is thought to play an essential role in cognition. This finding is consistent with previous clinical studies on cocaine use (Lane et al., 2010). As several previous studies have revealed that the disease status at proximally located/neighbouring voxels can be similar (see Wu et al., 2013; Xue, Bowman & Kang, 2018), our method is expected to be more clinically meaningful than other studies given that the detected regions are spatially contiguous. Furthermore, our test is able to detect additional regions of differences compared to clinical cocaine dependence studies that rely on FA summaries, such as Ma et al. (2017), which did not find these regions despite using the same dataset.

The MCMC trace plots and posterior densities of the Potts spatial dependence parameters $\theta$ are displayed in Figure 10. The concentration parameter $\xi$ has a 95% credible region equal to [0.884, 0.888], indicating there are a few active clusters. The group-clustering parameter $\alpha$ and spatial parameter $\beta$ control the within- and between-subject spatial dependence and have corresponding 95% credible regions that equal [0.323, 0.327] and [18.698, 18.703], respectively. The dependence information revealed by the two credible regions is identical to the information obtained from the empirical variograms (see Figure 8). Thus, similar to our use of the classic variogram, the generalized empirical variograms may also serve as an alternative tool for obtaining the plug-in hyperparameter values (Reich & Shaby, 2019).
Figure 10: The MCMC summaries of the group-clustering parameter $\alpha$, spatial parameter $\beta$ and the concentration parameter $\xi$ as obtained after fitting the Potts model to the cocaine dependence data. In each row, the left panel displays a histogram of the posterior samples with the 95% credible regions shaded, while the corresponding right panel shows the associated MCMC trace plot.

6. DISCUSSION

Although the spatial statistics literature concerning models and tools for matrix-variate data is sparse, the use of p.d. matrix-variate responses is extensive and includes multiple-input multiple-output systems (Smith & Garth, 2007) and computer vision (Cherian, Morellas & Papanikolopoulos, 2015) to name a few. Our contribution to these efforts has been to develop a spatial formulation of matrix-variate data and employ a Bayesian semiparametric mixture model for related estimation and inference. With a focus on DTI, our formulation retains the original data structure and avoids modelling popular data summaries, such as the FA. The spatial dependence is accounted for by a computationally elegant model. In simulation studies, our Potts model exhibited significantly improved performance compared to the non-spatial alternative RE model. Using our Potts model to fit the motivating DTI cocaine dependence data demonstrated the novelty of our approach to the problem of detecting clinically meaningful regions of differences.

DTI studies mostly consider modelling a single primary fibre direction per voxel. As pointed out by a reviewer, the fibre crossing (FC) problem, i.e., when there are two or more differently oriented fibre bundles in the same imaging voxel (Alexander & Seunarine, 2010), remains a fundamental limitation of diffusion MRI studies of the brain (Schilling et al., 2017). Note that the target area of the brain in our application, the CC, is believed to contain a single homogeneous fibre population (Alexander et al., 2007), with fewer white matter crossings. However, recent structure tensor analysis of the CC revealed (Schilling et al., 2017) dispersion, or heterogeneity of orientations, with the prevalence of FC increasing with the spatial resolution. Our proposed Potts model cannot handle this FC problem. However, we believe this limitation does not necessarily compromise the potential of our current modelling. This is because, even if there are crossing fibres, the analysis using DTs can still characterize the diffusion processes (Le Bihan
et al., 2001), although at the cost of more precise scientific interpretation. We hope to explore this challenge in future research.

Our current work primarily focuses on finding between-region differences in the brain at a single time point (baseline). Temporally dependent matrix-variate data have already been studied (Smith & Garth, 2007) in the literature, and corresponding spatiotemporal extensions of our model are possible, although they are non-trivial. Extensions to incorporate covariates (e.g., socio-demographics, such as age, gender, etc.) are possible by incorporating a regression term in the full conditional distribution of the cluster labels. These enhancements of our Potts model represent important avenues for future research and will be pursued elsewhere.

Appendix: A: Density Functions

1. $X$ follows the Wishart distribution, written as $X \sim \mathcal{W}_p(V, n)$, with the pdf

$$f(X|V, n) = \frac{1}{2^{np/2}|V/n|^{n/2} \Gamma_p \left( \frac{n}{2} \right)} |X^{(n-p-1)/2} e^{-(1/2)tr(V/n^{-1}X)}|.$$ 

2. $X$ follows the inverse-Wishart distribution, written as $X \sim \mathcal{IW}_p(\Psi, \nu)$, with the pdf

$$f(X|\Psi, \nu) = \frac{|(v-p-1)\Psi|^{v/2} |X|^{-(v+p+1)/2} e^{-\frac{1}{2}tr((v-p-1)\Psi^{-1}X)}|}{2^{vp/2} \Gamma_p \left( \frac{v}{2} \right)}.$$ 

Appendix: B: Variograms

We start with the variogram for the single-subject model, which is

$$\mathbb{E}[|A_u - A_v|^2] = Tr(\mathbb{E}[(A_u - A_v)(A_u - A_v)]) = Tr(\mathbb{E}[A_u A_u]) + Tr(\mathbb{E}[A_v A_v]) - 2Tr(\mathbb{E}[A_u A_v]) = [Tr(\mathbb{E}[A_u A_u]|g_u = g_v)] + [Tr(\mathbb{E}[A_v A_v]|g_u = g_v)] - 2[Tr(\mathbb{E}[A_u A_v]|g_u = g_v)] \times P(g_u = g_v) \beta + [Tr(\mathbb{E}[A_u A_v]|g_u \neq g_v)] + [Tr(\mathbb{E}[A_v A_v]|g_u \neq g_v)] - 2[Tr(\mathbb{E}[A_u A_v]|g_u \neq g_v)] \times P(g_u \neq g_v) \beta.

We now present three propositions.

**Proposition 1.** $\mathbb{E}[A_u A_u] = \mathbb{E}[A_v A_u|g_u = g_v] = \mathbb{E}[A_u A_v|g_u \neq g_v].$

**Proof.** The marginal density function of $A_u$ is independent of either $g_u = g_v$ or $g_u \neq g_v$ as the labels $g$ have already been marginalized. ■

**Proposition 2.** $\mathbb{E}[A_u A_v|g_u = g_v] = \mathbb{E}[A_u A_v].$

**Proof.** When $g_u = g_v$, we have $A_u | M_{g_u} \sim \mathcal{W}(M_{g_u}, m), A_v | M_{g_v} \sim \mathcal{W}(M_{g_v}, m)$ and $M_{g_u} = M_{g_v}$. Let $\mathcal{D}$ represent “equal in distribution.” Then, we have $[A_u A_v|g_u = g_v, M_{g_u}, M_{g_v}] \mathcal{D} = [A_u A_v]$. 

The Canadian Journal of Statistics / La revue canadienne de statistique DOI: 10.1002/cjs.11601
The result is trivial to prove because the marginal expectations are same for any $u$ or $v$. Given these three propositions, we now have

$$\mathbb{E}[|A_u - A_v|^2] = (\text{Tr}(\mathbb{E}[A_u A_u^T | g_u \neq g_v])) + \text{Tr}(\mathbb{E}[A_v A_v^T | g_u \neq g_v]) - 2\text{Tr}(\mathbb{E}[A_u A_v^T | g_u \neq g_v])$$

The variogram for multiple subjects, $\mathbb{E}[|A_{iu} - A_{iv}|^2]$, can be derived via standard book-keeping along similar lines; the details are omitted for brevity. Next, we present the explicit expressions of $\gamma(m, v, \Sigma)$. We first have

$$\text{Tr}(\mathbb{E}[A_u A_v^T | g_u \neq g_v]) = \mathbb{E}[\text{Tr}(\mathbb{E}[A_u A_v^T | g_u \neq g_v, M_{g_u}, M_{g_v}])]
= \mathbb{E}[\text{Tr}(\mathbb{E}[A_v | M_{g_v}]\mathbb{E}[A_u | M_{g_u}])]
= \sum_{i=1}^{p} \sigma_i(\Sigma)^2 = \lambda,$$

where $\sigma_i(.)$ returns the $i$th eigenvalue of the input function. The term $\text{Tr}(\mathbb{E}[A_u A_u^T])$ is complex. Gupta & Nagar (1999)[Section 3.3.6] provide trace moments of the Wishart and inverse Wishart distribution.

**Proposition 4.** Let $S \sim \mathcal{W}_p(\Sigma, m)$ or $W \sim \mathcal{W}_p(M, m)$, where $p$ is the matrix dimension. Also, $S^{-1} = W$ and $M = \Sigma^{-1}$. Then

1. $\mathbb{E}WW = (c_1 + c_2)mM(m-p-1)^2 + c_2\text{Tr}(M)M(m-p-1)^2$,
2. $\mathbb{E}SS = \frac{m^2}{m} \sum_{i=1}^{p} \sigma_i(\Sigma)^2 + \frac{1}{m} \text{Tr}(\Sigma)\Sigma$,
3. $\text{Tr}(S)S = 2\sum_{i=1}^{m} \sigma_i(\Sigma)^2 + \text{Tr}(\Sigma)\Sigma$, where $c_1 = (m-p-2)c_2$ and $c_2 = \frac{1}{(m-p)(m-p-1)(m-p-2)}$.

Thus

$$\mathbb{E}[A_{iu} A_{iv}^T | M_{g_u} = M_{g_v} = M] = (c_1 + c_2)mM(m-p-1)^2 + c_2\text{Tr}(M)M(m-p-1)^2 = \lambda^*$$

and

$$\text{Tr}(\mathbb{E}[A_{iu} A_{iv}]) = \text{Tr}(\mathbb{E}[\lambda^*]) = \{(c_1 + c_2)(v+1)v^{-1}\sum_{i=1}^{p} \sigma_i(\Sigma)^2 + (c_1 + c_2)v^{-1}\sum_{i=1}^{p} \sigma_i(\Sigma)\sum_{j=1}^{p} \sigma_j(\Sigma)
+ c_2(2v^{-1}\sum_{i=1}^{p} \sigma_i(\Sigma)^2 + \sum_{i=1}^{p} \sigma_i(\Sigma)\sum_{j=1}^{p} \sigma_j(\Sigma)))\}(m-p-1)^2
= \sigma.$$
Appendix: C: Euclidean vs Log-Euclidean metric

The Log-Euclidean metric is defined as

$$E || \log A_u - \log A_v ||^F_2,$$

where the log of a matrix is defined as

$$\log(B) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{(B-I)^k}{k} = (B-I) - \frac{(B-I)^2}{2} + \frac{(B-I)^3}{3} \ldots,$$

with the operator returning a matrix. Let \(\log A_u = B_u\). Then, \(E ||B_u - B_v||^F_2\) can be expressed as

$$E ||B_u - B_v||^F_2 = P(g_u \neq g_v \mid \beta) \left[ Tr(E[B_u B_u]) + Tr(E[B_v B_v]) - 2Tr(E[B_u B_v | g_u \neq g_v]) \right].$$

We observe that the expression for the log-Euclidean metric has the same architecture as the Euclidean metric: a product of a spatial term, \(P(g_u \neq g_v \mid \beta)\), and a non-spatial term. Hence, spatial separability also exists here.

Appendix: D: The joint PMF of \(\{g_1, g_2, \ldots, g_N\} \cup \{h_0, h_1\}\)

The joint PMF is

$$P(h, g) \propto \exp \left[ \sum_{i=1}^{N} \sum_{v=1}^{n} \alpha \mathcal{I}(g_{iv} = h_{x_{iv}}) + \sum_{x=0}^{1} \sum_{u \sim v} \beta \mathcal{I}(h_{xu} = h_{x_{iv}}) + \sum_{i=1}^{N} \sum_{u \sim v} (\beta \mathcal{I}(g_{iu} = g_{iv}) - g_{iu}^2) \right].$$

where \(u \sim v\) means \(u\) and \(v\) are connected. It is obvious that the probability is positive and satisfies the pairwise Markov property stated in the Hammersley and Clifford theorem (Clifford, 1990). The normalizing constant \(Z(\alpha, \beta, \xi) = \sum_{g,h} \exp(U(g,h,\theta))\) is intractable. As the summation is over finite and discrete indices, we find that \(0 < Z(\alpha, \beta, \xi) < \infty\), revealing that \(P(h,g)\) is proper.

Appendix: E: The statistical role of \(h_{xv}\)

Step 1: Marginalizing \(g_{iv}\)

$$[A_{iv} : u \in N_v, h_{x_{iv}}, \{V_k : k\}, m] = \sum_{k=1}^{K} P(g_{iv} = k \mid \cdot) \mathcal{I}_p(V_k, m) \cdot C_1 \exp \left[ -k^2 + \beta \sum_{u \in N_v} \mathcal{I}(g_{iu} = k) + \alpha \mathcal{I}(h_{x_{iv}} = k) \right] \mathcal{I}_p(V_k, m),$$

where \(C_1\) represents the normalizing constant.
Step 2: Marginalizing \( \{ g_{iu} : u \in \mathcal{N}_v \} \)

\[
Q = \sum_{g_{iu} = 1, K} \exp \left[ \beta \sum_{u \in \mathcal{N}_v} \mathcal{I} \left( g_{iu} = k \right) \right] \mathcal{P}_{g_{iu}, u \in \mathcal{N}_v} \]

\[
\left[ A_{lv} | h_{xlv}, \{ V_k : k \}, m \right] = \sum_{k=1}^{K} C_i Q \exp \left[ -k^\delta + \alpha \mathcal{I} \left( h_{xlv} = k \right) \right] \mathcal{J} \mathcal{W}_p (V_k, m) \]

ACKNOWLEDGEMENTS

We thank the Institute for Drug and Alcohol Studies of the Virginia Commonwealth University (VCU) for providing us with access to the cocaine users’ dataset. We also acknowledge financial support via two grants (U54DA038999 and R01DE024984) awarded by the United States National Institutes of Health.

BIBLIOGRAPHY

Alexander, A. L., Lee, J. E., Lazar, M., & Field, A. S. (2007). Diffusion tensor imaging of the brain. *Neurotherapeutics*, 4, 316–329.

Alexander, D. & Seunarine, K. (2010). Mathematics of crossing fibers. In *Diffusion MRI: Theory, Methods, and Application*, D. Jones, editor. Oxford University Press, New York, NY, pp. 451–464.

Arsigny, V., Fillard, P., Pennec, X., & Ayache, N. (2007). Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29, 328–347.

Cherian, A., Morellas, V., & Papanikolopoulos, N. (2015). Bayesian nonparametric clustering for positive definite matrices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38, 862–874.

Clifford, P. (1990). Markov random fields in statistics. In *Disorder in Physical Systems: A Volume in Honour of John M. Hammersley*, G. R. Grimmett & D. J. A. Welsh, editors. Oxford University Press, Oxford, UK, pp. 19–32.

Cohen, J. D., Daw, N., Engelhardt, B., Hasson, U., Li, K., Niv, Y., Norman, K. A., Pillow, J., Ramadge, P. J., Turk-Browne, N. B., et al. (2017). Computational approaches to fMRI analysis. *Nature Neuroscience*, 20, 304–313.

Cressie, N. & Wikle, C. K. (2015). *Statistics for Spatio-Temporal Data*. John Wiley & Sons, Hoboken, NJ.

Cressie, N. A. (1993). *Statistics for Spatial Data: Revised Edition*. John Wiley & Sons, New York.

Ennis, D. B. & Kindlmann, G. (2006). Orthogonal tensor invariants and the analysis of diffusion tensor magnetic resonance images. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 55, 136–146.

Goldsmith, J., Huang, L., & Crainiceanu, C. M. (2014). Smooth scalar-on-image regression via spatial Bayesian variable selection. *Journal of Computational and Graphical Statistics*, 23, 46–64.

Gupta, A. K. & Nagar, D. K. (1999). *Matrix Variate Distributions*, Vol. 104. CRC Press, Boca Raton, FL.

Heidelberger, P. & Welch, P. D. (1981). A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM*, 24, 233–245.

Hu, W., Kong, D., & Shen, W. (2021). *Nonparametric Matrix Response Regression with Application to Brain Imaging Data Analysis*. Biometrics. Early View.

Jin, I. H., Yuan, Y., & Bandyopadhyay, D. (2016). A Bayesian hierarchical spatial model for dental caries assessment using non-Gaussian Markov random fields. *The Annals of Applied Statistics*, 10, 884–905.

Johnson, T. D., Liu, Z., Bartsch, A. J., & Nichols, T. E. (2013). A Bayesian non-parametric Potts model with application to pre-surgical fMRI data. *Statistical Methods in Medical Research*, 22, 364–381.

Kang, J., Johnson, T. D., Nichols, T. E., & Wager, T. D. (2011). Meta analysis of functional neuroimaging data via Bayesian spatial point processes. *Journal of the American Statistical Association*, 106, 124–134.
Kong, D., An, B., Zhang, J., & Zhu, H. (2020). L2RM: Low-rank linear regression models for high-dimensional matrix responses. *Journal of the American Statistical Association*, 115, 403–424.

Lan, Z., Zhao, Y., Kang, J., & Yu, T. (2016). Bayesian network feature finder (BANFF): An R package for gene network feature selection. *Bioinformatics*, 32, 3685–3687.

Lane, S. D., Steinberg, J. L., Ma, L., Hasan, K. M., Kramer, L. A., Zuniga, E. A., Narayana, P. A., & Moeller, F. G. (2010). Diffusion tensor imaging and decision making in cocaine dependence. *PLoS One*, 5, e11591.

Le Bihan, D., Mangin, J.-F., Poupon, C., Clark, C. A., Pappata, S., Molko, N., & Chabriat, H. (2001). Diffusion tensor imaging: Concepts and applications. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 13, 534–546.

Lee, H. N. & Schwartzman, A. (2017). Inference for eigenvalues and eigenvectors in exponential families of random symmetric matrices. *Journal of Multivariate Analysis*, 162, 152–171.

Li, Q., Wang, X., Liang, F., Yi, F., Xie, Y., Gazdar, A., & Xiao, G. (2019). A Bayesian hidden Potts mixture model for analyzing lung cancer pathology images. *Biostatistics*, 20, 565–581.

Liang, F. (2010). A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants. *Journal of Statistical Computation and Simulation*, 80, 1007–1022.

Lim, K. O., Wozniak, J. R., Mueller, B. A., Franc, D. T., Specker, S. M., Rodriguez, C. P., Silverman, A. B., & Rotrosen, J. P. (2008). Brain macrostructural and microstructural abnormalities in cocaine dependence. *Drug and Alcohol Dependence*, 92, 164–172.

Lin, Z., D. Kong, and Q. Sun (2019). *Modeling Symmetric Positive Definite Matrices with An Application to Functional Brain Connectivity*. (under review), https://arxiv.org/abs/1907.03385.

Lindsay, B. G. & Lesperance, M. L. (1995). A review of semiparametric mixture models. *Journal of Statistical Planning and Inference*, 47, 29–39.

Liu, W., Awate, S. P., Anderson, J. S., & Fletcher, P. T. (2014). A functional network estimation method of resting-state fMRI using a hierarchical Markov random field. *NeuroImage*, 100, 520–534.

Lo, C.-Y., Wang, P.-N., Chou, K.-H., Wang, J., He, Y., & Lin, C.-P. (2010). Diffusion tensor tractography reveals abnormal topological organization in structural cortical networks in Alzheimer’s disease. *Journal of Neuroscience*, 30, 16876–16885.

Ma, L., Hasan, K. M., Steinberg, J. L., Narayana, P. A., Lane, S. D., Zuniga, E. A., Kramer, L. A., & Moeller, F. G. (2009). Diffusion tensor imaging in cocaine dependence: Regional effects of cocaine on corpus callosum and effect of cocaine administration route. *Drug and Alcohol Dependence*, 104, 262–267.

Ma, L., Steinberg, J. L., Wang, Q., Schmitz, J. M., Boone, E. L., Narayana, P. A., & Moeller, F. G. (2017). A preliminary longitudinal study of white matter alteration in cocaine use disorder subjects. *Drug and Alcohol Dependence*, 173, 39–46.

Martín-Fernández, M., Westin, C.-F., & Alberola-López, C. (2004). 3D Bayesian regularization of diffusion tensor MRI using multivariate Gaussian Markov random fields. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 351–359. https://link.springer.com/chapter/10.1007/978-3-540-30135-6_43.

McCullagh, P. & Yang, J. (2008). How many clusters? *Bayesian Analysis*, 3, 101–120.

Moeller, F. G., Hasan, K. M., Steinberg, J. L., Kramer, L. A., Dougherty, D. M., Santos, R. M., Valdes, I., Swann, A. C., Barratt, E. S., & Narayana, P. A. (2005). Reduced anterior corpus callosum white matter integrity is related to increased impulsivity and reduced discriminability in cocaine-dependent subjects: Diffusion tensor imaging. *Neuropsychopharmacology*, 30, 610–617.

Muschelli, J., Sweeney, E., & Crainiceanu, C. (2014). brainR: Interactive 3 and 4D images of high resolution neuroimage data. *The R Journal*, 6, 41–48.

Musgrove, D. R., Hughes, J., & Eberly, L. E. (2016). Fast, fully Bayesian spatiotemporal inference for fMRI data. *Biostatistics*, 17, 291–303.

Park, J. & Haran, M. (2018). Bayesian inference in the presence of intractable normalizing functions. *Journal of the American Statistical Association*, 113, 1372–1390.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66, 846–850.

Reich, B. J., Guinness, J., Vandekar, S. N., Shinohara, R. T., & Staicu, A.-M. (2018). Fully Bayesian spectral methods for imaging data. *Biometrics*, 74, 645–652.
Reich, B. J. & Shaby, B. A. (2019). A spatial Markov model for climate extremes. *Journal of Computational and Graphical Statistics*, 28, 117–126.

Robert, C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer-Verlag, New-York.

Rodriguez, C. E. & Walker, S. G. (2014). Label switching in Bayesian mixture models: Deterministic relabeling strategies. *Journal of Computational and Graphical Statistics*, 23, 25–45.

Schilling, K., Gao, Y., Janve, V., Stepniewska, I., Landman, B. A., & Anderson, A. W. (2017). Can increased spatial resolution solve the crossing fiber problem for diffusion MRI? *NMR in Biomedicine*, 30, e3787.

Schwartzman, A., Mascarenhas, W. F., & Taylor, J. E. (2008). Inference for eigenvalues and eigenvectors of Gaussian symmetric matrices. *The Annals of Statistics*, 36, 2886–2919.

Smith, P. J. & Garth, L. M. (2007). Distribution and characteristic functions for correlated complex Wishart matrices. *Journal of Multivariate Analysis*, 98, 661–677.

Soares, J. M., Marques, P., Alves, V., & Sousa, N. (2013). A Hitchhiker’s guide to diffusion tensor imaging. *Frontiers in Neuroscience*, 7, 31.

Spence, J. S., Carmack, P. S., Gunst, R. F., Schucany, W. R., Woodward, W. A., & Haley, R. W. (2007). Accounting for spatial dependence in the analysis of SPECT brain imaging data. *Journal of the American Statistical Association*, 102, 464–473.

Steinley, D. (2004). Properties of the Hubert–Arabie adjusted Rand index. *Psychological Methods*, 9, 386–396.

Woolrich, M. W., Behrens, T. E., Beckmann, C. F., Jenkinson, M., & Smith, S. M. (2004a). Multilevel linear modelling for fMRI group analysis using Bayesian inference. *NeuroImage*, 21, 1732–1747.

Woolrich, M. W., Jenkinson, M., Brady, J. M., & Smith, S. M. (2004b). Fully Bayesian spatio-temporal modeling of fMRI data. *IEEE Transactions on Medical Imaging*, 23, 213–231.

Wu, F.-Y. (1982). The Potts model. *Reviews of Modern Physics*, 54, 235.

Wu, G.-R., Stramaglia, S., Chen, H., Liao, W., & Marinazzo, D. (2013). Mapping the voxel-wise effective connectome in resting state fMRI. *PloS ONE*, 8, e73670.

Xue, W., Bowman, F. D., & Kang, J. (2018). A Bayesian spatial model to predict disease status using imaging data from various modalities. *Frontiers in Neuroscience*, 12, 184.

Yuan, Y., Zhu, H., Lin, W., & Marron, J. (2012). Local polynomial regression for symmetric positive definite matrices. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74, 697–719.

Zhao, Y., Kang, J., & Yu, T. (2014). A Bayesian nonparametric mixture model for selecting genes and gene subnetworks. *The Annals of Applied Statistics*, 8, 999.

Zhu, H., Chen, Y., Ibrahim, J. G., Li, Y., Hall, C., & Lin, W. (2009). Intrinsic regression models for positive-definite matrices with applications to diffusion tensor imaging. *Journal of the American Statistical Association*, 104, 1203–1212.

Received 11 December 2019
Accepted 28 August 2020