A Covariance Based Clustering for Tensor Objects

Rene Gutierrez Marquez, Aaron Wolfe Scheffler, Rajarshi Guhaniyogi,
Abigail Dickinson, Charlotte DiStefano and Shafali Jeste

January 2021

Abstract

Clustering of tensors with limited sample size has become prevalent in a variety of application areas. Existing Bayesian model based clustering of tensors yields less accurate clusters when the tensor dimensions are sufficiently large, sample size is low and clusters of tensors mainly reveal difference in their variability. This article develops a clustering technique for high dimensional tensors with limited sample size when the clusters show difference in their covariances, rather than in their means. The proposed approach constructs several matrices from a tensor, referred to as transformed features, to adequately estimate its variability along different modes and implements a model-based approximate Bayesian clustering algorithm with the matrices thus constructed, in place with the original tensor data. Although some information in the data is discarded, we gain substantial computational efficiency and accuracy in clustering. Simulation study assesses the proposed approach along with its competitors in terms of estimating the number of clusters, identification of the modal cluster membership along with the probability of mis-classification in clustering (a measure of uncertainty in clustering). The proposed methodology provides novel insights into potential clinical subgroups for children with autism spectrum disorder based on resting-state electroencephalography activity.

Keywords: Bayesian statistics; Brain electro-encephalogram signal; Clustering; Chinese restaurant process; Tensor normal distribution.

1 Introduction

In recent times, multidimensional arrays or tensors are encountered in different disciplines including datasets from different brain imaging modalities, multi-omics studies, chemometrics and psychometrics. Statistical analysis of tensor data presents challenges over and above multivariate vector-based methods. First of all, due to the high dimensional nature of tensor data, inference from tensors often require a large parameter space. Also, extra care needs to be exercised to exploit structural information in a tensor object. To address such challenges for tensor data, a plethora of literature has emerged on tensor decomposition (Chi and Kolda, 2012; Dunson and Xing, 2009; Sun and Li, 2019a) and regressions with general and symmetric tensors (Zhou et al., 2013; Guhaniyogi et al., 2017; Lock, 2018; Guhaniyogi and Spencer, 2018; Guha and Guhaniyogi, 2020; Spencer et al., 2020). Most of these approaches employ low-rank and sparse approximations in the tensor structure to reduce the number of parameters considerably, and propose novel estimation tools to draw adequate inference.

This article focuses on clustering of tensors into subgroups when tensors in different subgroups are barely distinguishable in terms of locations (e.g. mean), but exhibit difference in their correlation structures/variability. Examples of such datasets can be found in image analysis, financial, and biological processes. Specifically, we consider a motivating study that collected resting-state electroencephalogram (EEG) data on children with autism spectrum disorder (ASD) to better understand the neural mechanisms underlying observed developmental delays. EEG characterizes cortical brain activity via a high-density electrode array that measures neuronal electrical potentials and their corresponding oscillatory dynamics (i.e. spectral characteristics via frequency decomposition) which results in a two-way tensor composed of an electrode dimension and a time series or frequency dimension. Recent studies in cognitive development using EEG highlight the peak alpha frequency (PAF), defined as the location of a prominent peak in the spectral density within the alpha frequency band (6-14 Hz), as a potential biomarker associated with autism diagnosis (Edgar et al., 2015; Dickinson et al., 2018; Edgar et al., 2019). Thus, we propose to cluster the ASD children into subgroups based on the resting-state EEG data by focusing on alpha band spectral dynamics across electrodes to understand the role of this novel biomarker in segmenting chil-

dren with ASD. In a previous analysis of the motivating alpha spectral density EEG data, Scheffler et al. (2019) found a common alpha spectral mean structure across ASD patients 2-12 years old. However, patients exhibited substantial heterogeneity in patterns of alpha spectral variation across chronological development suggesting that potential subgroups may be identified in second rather than first moment information. Thus, in this application, it is of clinical interest to determine if ASD patients cluster in terms of patterns of alpha spectral variation across the scalp rather than the mean structure. This hypothesis is supported by evidence from prior studies that identified EEG spectral coherence, a specific measure EEG spectral covariation across electrodes, both as a correlate of ASD severity (Duffy and Als, 2012; Duffy et al., 2013) as well as a successful target for unsupervised clustering of ASD patients (Duffy and Als, 2019). See Schwartz et al. (2017) for a thorough review of EEG coherence in patients with ASD. While these previous approaches collapsed EEG coherence across entire power bands prior to analysis, we adopt a more general approach by modeling covariation among electrodes across the entire alpha spectral band via our tensor framework.

We now offer a brief exposition to the current literature for clustering tensors. Loss-based algorithmic approaches for clustering of vectors (Hartigan and Wong, 1979; Banerjee et al., 2004) can be extended to clustering of tensors (Huang et al., 2008), offering a simple approach that is computationally efficient. However, loss-based approaches focuses on the aggregation and separation of a sample into groups depending on similarities in locations of data, and hence is not useful in applications of our interest. Moreover, there is no way to account for clustering uncertainty in these methods. In contrast with algorithmic clustering, model-based clustering (Fraley and Raftery, 2002; Müller et al., 2015) exploits the entire data distribution for clustering. In clustering the tensor observations under the model-based clustering framework, one simple solution would be to vectorize the tensor object followed by unsupervised clustering of these vectors, following the literature on clustering of high dimensional vectors (Medvedovic and Sivaganesan, 2002; Zhong and Ghosh, 2003; Raftery and Dean, 2006; Fröhwirth-Schnatter and Kaufmann, 2008; Pan and Shen, 2007; Wang and Zhu, 2008; Lee et al., 2013; Oh and Raftery, 2007). Vectorization of a K-mode tensor of dimensions $p_1 \times \cdots \times p_K$ results in a long vector of dimension $\prod_{k=1}^{K} p_k$, which often leads to inaccurate clustering (Celeux et al., 2019). Frühwirth-Schnatter (2006) proposes a

specific prior elicitation criterion to overcome this issue for moderate dimensions. However, calibration of hyper-parameters may appear to be difficult for large dimensions.

The model-based clustering typically proceeds by assuming each observation to follow a finite/infinite mixture of distributions, e.g., Gaussian mixture model (GMM). In the context of clustering higher order tensors, an ordinary GMM can be extended to mixture of tensor normal distributions, referred to as tensor normal mixtures (TNM) hereon. The TN distribution expresses the covariance structure of a tensor in terms of covariance structure in every mode of the tensor, i.e., the covariance of a K-mode tensor of dimensions $p_1 \times \cdots \times p_K$ is expressed with covariance matrices of the order $p_1 \times p_1$,..., $p_K \times p_K$ at K modes. Further, the tensor covariance structure can be exploited to simultaneously cluster observations and estimate parameters using either expectation maximization (EM) algorithm, its variants (in the frequentist framework) or Gibbs sampling (in the Bayesian framework) (Viroli, 2011; Anderlucci et al., 2015; Gao et al., 2020; Mai et al., 2021a). However, a standard Gibbs sampling algorithm for the clustering of tensors presents an arduous task of sampling the covariance matrices in each mode of the tensors at every iteration. Besides being computationally inefficient, especially for high-dimensional tensors, this often results in inaccurate estimation of true clusters in presence of limited sample size, as we demonstrate in this article.

This article tackles the problem from a different point of view. In particular, we focus on a set of observations from multiple populations all of which follow tensor normal distributions with the same mean but different covariances. Rather than directly clustering these observations using model-based clustering that presents challenges described earlier, we adopt a two-step approach. As a first step, we construct a set of matrices, referred to as the "transformed features," from each tensor. We prove that when $p_1, ..., p_K$ are large, the transformed features accurately estimate the mode-specific covariance matrices of a TN distribution, thereby turning the curse of dimensionality into a blessing. In the second step, a Bayesian mixture model on transformed features is employed to cluster observations. The proposal makes use of difference between clusters in their covariance structure, and at the same time avoids drawing Markov Chain Monte Carlo (MCMC) samples for high dimensional covariance parameters from tensor normal distributions, resulting in straightforward

computation even with large tensor dimensions. Moreover, we provide clustering uncertainty in terms of mis-classification probabilities.

In the similar spirit as ours, Ieva et al. (2016) developed a novel covariance-based clustering algorithm exploiting the distance between covariances for multi-variate and functional data. Their approach is based on the crucial assumption that the data admits only two groups/clusters, while we do not need to specify the number of clusters. Hallac et al. (2018) proposed a method for multivariate time-series data to segment and cluster, but they posit a restrictive Toeplitz structure for the covariance matrix.

Rather than clustering tensors using the mixture of tensor normal distributions, there is a literature regarding K-means clustering on low-rank approximation of tensors. For example, a class of methods assume tensor decomposition of the mean of the tensor normal distribution, followed by minimization of the total squared Euclidean distance of each observation mean to its cluster centroid (Sun and Li, 2019a). While the low-rank approximation is widely adopted in tensor data analysis, it typically work on identifying clusters through centers of their distributions, and is less suitable for our purpose. Lee et al. (2010), Tan and Witten (2014) develop bi-clustering methods that simultaneously group features and observations into clusters. Extensions of the feature-sample bi-clustering for vector observations are known as the co-clustering or multiway clustering problems (Jegelka et al., 2009; Chi et al., 2020; Wang and Zeng, 2019), where each mode of the tensor is clustered into groups. Our problem is different from these works in that our sole goal is to cluster the observations.

Rest of the article evolves as follows. In section 2 we provide a brief introduction to model based clustering and describe our approach for clustering tensors with covariance estimators. Posterior computation from the model is described in Section 3. Empirical evaluations with simulation studies and the resting-state EEG data analysis are presented in Sections 4 and 5, respectively. Finally, we conclude in Section 6 with an eye towards the future work. Proofs of the theoretical results are presented in the supplementary material.

2 Covariance-based Bayesian Tensor Clustering

2.1 Notations

We begin with a quick review of some tensor notations and operations which will be subsequently used. A more detailed review can be found in Kolda and Bader (2009).

Consider the K-way tensor (also known as K-mode or K-th order tensor) $\mathbf{T} \in \mathbb{R}^{p_1 \times ... \times p_K}$ with its $(i_1, ..., i_K)$-th element denoted by $T_{i_1,...,i_K}$. When $K = 1$, the tensor reduces to a vector and when $K = 2$, the tensor is a matrix. The vec($\mathbf{T}$) operator applied to a tensor $\mathbf{T}$ stacks elements into a column vector of dimension $p = \prod_{k=1}^{K} p_k$ with $T_{i_1,...,i_K}$ mapped to the j-th entry of $vec(\mathbf{T})$, for $j = 1 + \sum_{k=1}^{K}(i_k - 1)\prod_{k'=1}^{k-1} p_{k'}$.

A fiber is the higher order analogue of a matrix row and column, and is defined by fixing every index of the tensor but one. A k-mode fiber is a p_k-dimensional vector obtained by fixing all other modes except the k-th mode. For example, a matrix column is a mode-1 fiber and a row is a mode-2 fiber. There are p/p_k such k-mode fibers for $\mathbf{T}$ each with dimension $p_k \times 1$. The k-mode matricization of a tensor transforms a tensor into a matrix $\mathbf{T}_{(k)} \in \mathbb{R}^{p_k \times \frac{p}{p_k}}$, where $T_{(i_1,...,i_K)}$ mapping to (i_k, j)-th element of the matrix, where $j = \sum_{k' \neq k}(i_{k'} - 1)\prod_{k'' < k', k'' \neq k} p_{k''}$. The k-mode product of a tensor $\mathbf{T} \in \mathbb{R}^{p_1 \times ... \times p_K}$ and a compatible matrix $\mathbf{A} \in \mathbb{R}^{J \times p_k}$, will result in a tensor $\mathbf{T} \times_k \mathbf{A} \in \mathbb{R}^{p_1 \times ... \times p_{k-1} \times J \times p_{k+1} \times ... p_K}$, where each element is the product of mode-k fiber of $\mathbf{T}$ multiplied by $\mathbf{A}$. Notice that this operation reduces to the usual matrix product for a 2-way tensor and to the inner product for a 1-way tensor. Finally, for a list of matrices $\mathbf{A}_1, \ldots, \mathbf{A}_K$ with compatible sizes $A_k \in \mathbb{R}^{J_k \times p_k}$ we define the product $\mathbf{T} \times [\mathbf{A}_1, \ldots, \mathbf{A}_K] = \mathbf{T} \times_1 \mathbf{A}_1 \times_2 \ldots \times_K \mathbf{A}_K \in \mathbb{R}^{J_1 \times ... \times J_K}$. Thus, when $\mathbf{A}_1, \ldots, \mathbf{A}_K$ are square matrices, the resulting tensor is of the same dimension as $\mathbf{T}$. In what follows, we will use $||\cdot||_F$ to denote the Frobenius norm of the tensor $\mathbf{T}$ given by $||\mathbf{T}||_F := \sqrt{\sum_{i_1,...,i_K} T^2_{i_1,...,i_K}}$.

2.2 Bayesian Model-based Tensor Clustering Approach

Let $\mathbf{T}_i$ be a tensor valued observation in $\mathcal{T}$, $\mathcal{T} \subseteq \mathbb{R}^{p_1 \times \cdots \times p_K}$, for $i = 1, ..., n$. Let $\mathcal{C} = \{\mathcal{C}_1, ..., \mathcal{C}_{n(\mathcal{C})}\}$ be a partition of n observations into $n(\mathcal{C})$ disjoint sets, i.e., $|\mathcal{C}| = n(\mathcal{C})$. Typical

Bayesian models for clustering are based on posterior distributions of the form

$$\pi(\mathcal{C}|\mathbf{T}_1,..,\mathbf{T}_n) \propto \pi(\mathcal{C}) \prod_{h=1}^{n(\mathcal{C})} \left[\int \prod_{i \in \mathcal{C}_h} f(\mathbf{T}_i|\mathbf{\Theta}_h)\pi(\mathbf{\Theta}_h)d\mathbf{\Theta}_h \right] = \pi(\mathcal{C}) \prod_{h=1}^{n(\mathcal{C})} m(\{\mathbf{T}_i : i \in \mathcal{C}_h\}), \quad (1)$$

where $f(\mathbf{T}_i|\mathbf{\Theta}_h)$ denotes the likelihood for a tensor observation belonging to the h-th cluster with the cluster-specific model parameter $\mathbf{\Theta}_h$ and $\pi(\mathbf{\Theta}_h)$ corresponds to the prior distribution on the parameter $\mathbf{\Theta}_h$. The quantity $m(\{\mathbf{T}_i : i \in \mathcal{C}_h\}) = \int \prod_{i \in \mathcal{C}_h} f(\mathbf{T}_i|\mathbf{\Theta}_h)\pi(\mathbf{\Theta}_h)d\mathbf{\Theta}_h$ denotes the marginal distribution of tensors belonging to the h-th cluster which is typically not obtained in a closed form. Alternatively, the partition can be described through cluster labels for n observations given by $\mathbf{c} = (c_1, ..., c_n)'$, so that $c_i = h$, if and only if $i \in \mathcal{C}_h$, for $i = 1, ..., n$. Irrespective of the representation, our interest only lies in the induced partition $\mathcal{C}$ rather than the labels on the indicators $\mathbf{c} = (c_1, ..., c_n)'$.

A natural choice for the likelihood $f(\mathbf{T}_i|\mathbf{\Theta}_h)$ appears to be a tensor normal distribution, denoted as $\text{TN}(\boldsymbol{M}_h, \mathbf{\Sigma}_{1,h}, ..., \mathbf{\Sigma}_{K,h})$, and is given by

$$f(\mathbf{T}_i|\boldsymbol{M}_h, \mathbf{\Sigma}_{1,h}, \ldots, \mathbf{\Sigma}_{K,h}) = (2\pi)^{-\frac{p}{2}} \left\{ \prod_{k=1}^{K} |\mathbf{\Sigma}_{k,h}|^{-\frac{p}{2p_k}} \right\} \exp\left(-\frac{1}{2} \left\| (\mathbf{T}_i - \boldsymbol{M}_h) \times [\mathbf{\Sigma}_{1,h}^{-\frac{1}{2}}, \ldots, \mathbf{\Sigma}_{K,h}^{-\frac{1}{2}}] \right\|_F^2 \right), \quad (2)$$

where $\boldsymbol{M}_h$ is the mean/center of the TN distribution, and $\mathbf{\Sigma}_{k,h}$ is a $p_k \times p_k$ dimensional positive definite matrix, also referred to as the covariance matrix for the k-th mode.

This article focuses on a scenario where the observed tensors in the sample are barely distinguishable in terms of their means and the tensors belonging to different clusters only differ in terms of their variability. Thus, the following crucial assumption is made hereon.

Assumption A: *Different clusters of tensors only vary in terms of their covariance structure and not in their means. Thus, without loss of generality,* $\boldsymbol{M}_h = \mathbf{0}$ *for all* $h = 1, .., n(\mathcal{C})$.

According to the likelihood specification in (2) and Assumption A, $\mathbf{\Theta}_h$ corresponds to the collection of covariance matrices for all modes, i.e., $\mathbf{\Theta}_h = \{\mathbf{\Sigma}_{1,h}, \ldots, \mathbf{\Sigma}_{K,h}\}$. While we develop our approach based on Assumption A, simulation studies also demonstrate its good performance when Assumption A is violated.

The distributional form of $f(\mathbf{T}_i|\mathbf{\Theta}_h)$, as given in (2), does not yield a closed form integral

for the marginal distribution in (1). The common practice is to begin with the distribution $(\mathbf{T}_i|\mathbf{\Theta}_h, c_i = h) \sim f(\mathbf{T}_i|\mathbf{\Theta}_h)$ and develop a Gibbs sampler to draw posterior samples of $\mathbf{c}$ along with $\mathbf{\Sigma}_{k,h}$'s, for all $k = 1, ..., K$ and $h = 1, ..., n(\mathcal{C})$. However, when $p_1, ..., p_K$ are large and sample size n is moderate, Gibbs sampling of covariance matrices $\mathbf{\Sigma}_{k,h}$'s results in inferential inaccuracy related to clustering, as demonstrated in Section 4. Next section develops an approximate Bayesian clustering algorithm that offers remedies to this challenge.

2.3 A Covariance-Based Bayesian Tensor Clustering Approach

To avoid complications due to model based clustering of high-dimensional tensor observations, we propose a two-step Bayesian clustering approach of tensors. In summary, our approach first extracts important features of tensors to adequately estimate the covariance structure along different modes, followed by model-based clustering of these features. To elaborate on it, let $\mathcal{A}(\mathbf{T}_i)$ be the set of extracted features from tensor $\mathbf{T}_i$ which will be referred to as transformed features (TF) hereon. The transformed features are carefully chosen to estimate variability of the tensor normal distribution in each mode. Section 2.4 details out a specific choice of such transformed features. While the exact distribution of $\mathcal{A}(\mathbf{T}_i)$ is determined by the tensor normal specification given in (2), we focus on a reasonable approximation of the distribution for $\mathcal{A}(\mathbf{T}_i)$ in our goal to cluster these transformed features. Let $\tilde{f}(\mathcal{A}(\mathbf{T}_i)|\tilde{\mathbf{\Theta}}_h, \tilde{\mathbf{\Theta}}_a)$ be the approximated distribution of $\mathcal{A}(\mathbf{T}_i)$ in the h-th cluster, with $\tilde{\mathbf{\Theta}}_h$ as its h-th cluster-specific parameter and $\tilde{\mathbf{\Theta}}_a$ an auxiliary lower dimensional parameter common across all clusters. Let $\tilde{\pi}_h(\tilde{\mathbf{\Theta}}_h)$ and $\tilde{\pi}_a(\tilde{\mathbf{\Theta}}_a)$ denote the prior distribution of $\tilde{\mathbf{\Theta}}_h$ and $\tilde{\mathbf{\Theta}}_a$, respectively, for $h = 1, ..., H$. We choose $\tilde{f}(\cdot)$ and $\tilde{\pi}_h(\cdot)$ to ensure closed form marginal distribution of $\tilde{m}(\{\mathcal{A}(\mathbf{T}_i) : i \in \mathcal{C}_h\}|\tilde{\mathbf{\Theta}}_a) = \int \prod_{i\in\mathcal{C}_h} \tilde{f}(\mathcal{A}(\mathbf{T}_i)|\tilde{\mathbf{\Theta}}_h, \tilde{\mathbf{\Theta}}_a)\tilde{\pi}_h(\tilde{\mathbf{\Theta}}_h)d\tilde{\mathbf{\Theta}}_h$.

With closed form marginals for TFs in each cluster, the posterior distribution of clusters and the auxiliary parameters is given by,

$$\pi(\mathcal{C}, \tilde{\mathbf{\Theta}}_a \mid \mathcal{A}(\mathbf{T}_1), ..., \mathcal{A}(\mathbf{T}_n)) = \pi(\mathcal{C})\tilde{\pi}_a(\tilde{\mathbf{\Theta}}_a) \prod_{h=1}^{n(\mathcal{C})} \tilde{m}(\{\mathcal{A}(\mathbf{T}_i) : i \in \mathcal{C}_h\}|\tilde{\mathbf{\Theta}}_a), \tag{3}$$

where $\pi(\mathcal{C})$ denotes the prior on partitions. In the interests of computational convenience, we propose to adopt prior models on partitions for which posterior simulation methods are fully

developed (Ferguson, 1973; Antoniak, 1974; Gopalan and Berry, 1998). More specifically, with the posterior distribution of partitions given in (3), the computation proceeds through a Chinese restaurant sampler described below (Lau and Green, 2007).

1. *Initialize:* Choose an initial partition $\mathcal{C}^{(0)}$. Common options are either to set singleton clusters or to put all observations in the same cluster.

2. *Obtain s-th iterate of $\mathcal{C}$:* To obtain s-th iterate of the partition $\mathcal{C}^{(s)}$ do:

   (a) *Initialize the Partition:* Set $\mathcal{C} = \mathcal{C}^{(s-1)}$, and let $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{n(\mathcal{C})}\}$.

   (b) *Loop through every observation:*

      i. *Remove observation $\mathcal{A}(\mathbf{T}_i)$ from the partition:* Remove i-the observation from the partition $\mathcal{C}$ to obtain a new partition $\mathcal{C}_{-i} = \{\mathcal{C}_{1,-i}, ..., \mathcal{C}_{n(\mathcal{C}_{-i}),-i}\}$.

      ii. *Assign observation i:* Either assign the i-th observation to a new cluster, that is update $\mathcal{C}$ to $\mathcal{C} = \{\mathcal{C}_{1,-i}, ..., \mathcal{C}_{n(\mathcal{C}_{-i}),-i}, \{i\}\}$ with probability proportional to:

$$\tilde{m}(\mathcal{A}(\mathbf{T}_i)|\tilde{\mathbf{\Theta}}_a) \times \frac{\pi(\{\mathcal{C}_{1,-i}, ..., \mathcal{C}_{n(\mathcal{C}_{-i}),-i}, \{i\}\})}{\pi(\{\mathcal{C}_{1,-i}, ..., \mathcal{C}_{n(\mathcal{C}_{-i}),-i}\})}, \tag{4}$$

      or, assign the i-th observation to the existing j-th cluster $\mathcal{C}_{j,-i}$, that is update $\mathcal{C}$ to $\mathcal{C} = \{\mathcal{C}_{1,-i}, ..., \mathcal{C}_{j,-i} \cup \{i\}, \ldots, \mathcal{C}_{n(\mathcal{C}_{-i}),-i}\}$ with probability proportional to:

$$\frac{\tilde{m}(\{\mathcal{A}(\mathbf{T}_s) : s \in \{\{i\} \cup \mathcal{C}_{j,-i}\}\}|\tilde{\mathbf{\Theta}}_a)}{\tilde{m}(\{\mathcal{A}(\mathbf{T}_s) : s \in \mathcal{C}_{j,-i}\}|\tilde{\mathbf{\Theta}}_a)} \times \frac{\pi(\{\mathcal{C}_{1,-i}, ..., \mathcal{C}_{j,-i} \cup \{i\}, \ldots, \mathcal{C}_{n(\mathcal{C}_{-i}),-i}\})}{\pi(\{\mathcal{C}_{1,-i}, ..., \mathcal{C}_{n(\mathcal{C}_{-i}),-i}\})} \tag{5}$$

   (c) *Set the partition $\mathcal{C}^{(s)}$:* After updating $\mathcal{C}$, going through every observation, set $\mathcal{C}^{(s)} = \mathcal{C}$.

3. *Sample the s-th iterate of $\tilde{\mathbf{\Theta}}_a$:* Draw s-th iterate of $\tilde{\mathbf{\Theta}}_a$ from its full conditional distribution derived from (3).

Notably, steps (a)-(c) involve approximate marginal distribution of TFs which are available in closed form by our assumption. In fact, the algorithm bypasses updating high dimensional

parameters at any step, which leads to rapid mixing of the Markov Chain. Since the algorithm uses transformed features $\mathcal{A}(\mathbf{T}_i)$ of the tensor $\mathbf{T}_i$, the clustering accuracy is naturally dependent on the choice of these features. Next section describes a specific choice of TFs which leads to desirable clustering performance for tensors.

2.4 Transformed Features and Their Distributions

This section discusses the specific choice of transformed features $\mathcal{A}(\mathbf{T})$ and the approximate distribution $\tilde{f}(\mathcal{A}(\mathbf{T})|\tilde{\mathbf{\Theta}}_h, \tilde{\mathbf{\Theta}}_a)$ of the transformed features used in this article. Specifically, we propose to work with the collection of transformed features given by $\mathcal{A}(\mathbf{T}_i) = \{\frac{p_k}{p}\mathbf{T}_{i,(k)}\mathbf{T}'_{i,(k)} : k = 1, ..., K\}$, where $\mathbf{T}_{i,(k)}$ is the k-th mode matrix of the tensor $\mathbf{T}_i$. Therefore, given a K-way tensor observation $\mathbf{T}_i$ of dimension $p_1 \times \cdots \times p_K$ (where $p = \prod_{i=1}^{K} p_i$), we extract a collection of K matrices of sizes $p_1 \times p_1, \ldots, p_K \times p_K$, which will suitably capture the covariance structure of the observed tensor, as described by the lemma below.

Lemma 2.1 *Let* $\mathbf{T}_i \sim TN(\mathbf{0}, \mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_K)$ *and* $\mathcal{A}(\mathbf{T}_i)^{(k)} = \frac{p_k}{p}\mathbf{T}_{i,(k)}\mathbf{T}'_{i,(k)}$. *Assume that for all* $k = 1, ..., K$, *as* $p \to \infty$, *we have (i)* $\frac{p_k}{p} \to 0$ *(ii)* $\frac{p_k}{p} tr(\otimes_{k' \neq k}\mathbf{\Sigma}_{k'}) \to w_k$ *and (iii)* $\frac{p_k^2}{p^2}\sum_{l,r}\{\otimes_{k' \neq k}\mathbf{\Sigma}_{k'}\}_{l,r} \to 0$, *for all* $l, r = 1, ..., p/p_k$, *where* $\{\otimes_{k' \neq k}\mathbf{\Sigma}_{k'}\}_{l,r}$ *denotes the* (l, r)*th entry of the matrix* $\otimes_{k' \neq k}\mathbf{\Sigma}_{k'}$. *(i)-(iii) together imply that* $\{\mathcal{A}(\mathbf{T}_i)^{(k)}\}_{l,r} \to \{\mathbf{\Sigma}_k\}_{l,r} w_k$, *where* w_k *is a constant.*

The result implies that under regularity conditions, as the tensor dimensions grow, the transformed features converge to mode-specific covariance matrices upto a scale factor, recovering their shapes and orientations. The proof of Lemma 2.1 is provided in the supplementary material. Before discussing the implication of Lemma 2.1, some discussions on assumptions (i)-(iii) is warranted. Assumption (i) is a mild one only guaranteeing growth of tensor along every dimension. Broadly, the conditions (ii) and (iii) restrict the number of nonzero elements in the mode-specific covariance matrices generating the data, which turn out to be crucial in dictating the clustering performance of the approach, as observed in Section 4. In particular, when $\mathbf{\Sigma}_k$ is an identity matrix of dimension $p_k \times p_k$, (ii) and (iii) are trivially satisfied with $w_k = 1$ for all $k = 1, ..., K$.

The lemma reveals an interesting aspect of the transformed features. Note that the major inferential and computational challenges of clustering high-dimensional tensors using

mixture models stems from estimating high dimensional covariance matrices for different modes. In contrast, when sparsity of the true covariance matrices are restricted following assumptions (ii) and (iii), higher tensor dimensions will guarantee accurate estimation of mode-specific covariance matrices by the transformed features, which is conducive to our approximate tensor clustering approach, as revealed in Lemma 2.1.

2.4.1 The Approximate TF Distribution and Prior On Parameters

Following the mixture of tensor normal specification for the tensors $\mathbf{T}_i$ as in (2), the transformed features (TFs) $\mathcal{A}(\mathbf{T}_i)$ follow a mixture distribution which does not allow straightforward posterior computation. We propose an approximation wherein we employ cluster-specific normal means model on the upper triangular entries of $\mathcal{A}(\mathbf{T}_i)^{(k)}$ in all clusters and for all modes $k = 1, ..., K$. More specifically, the (l, r)-th entry of $\mathcal{A}(\mathbf{T}_i)^{(k)}$ is modeled as

$$\{\mathcal{A}(\mathbf{T}_i)^{(k)}\}_{l,r} \overset{ind.}{\sim} N(\theta^{(k)}_{l,r,h}, \sigma^2), \text{ for } i \in \mathcal{C}_h, \ \theta^{(k)}_{l,r,h} \sim N(\theta_0, \sigma^2/\phi), \ l < r. \tag{6}$$

(6) can be viewed as an approximation to the actual distribution of TFs under the mixture of tensor normal specification of $\mathbf{T}_i$, when tensor dimensions are large. In fact, when $i \in \mathcal{C}_h$ and $\mathbf{T}_i \sim TN(\mathbf{0}, \mathbf{\Sigma}_{1,h}, .., \mathbf{\Sigma}_{K,h})$, $\{\mathcal{A}(\mathbf{T})^{(k)}\}_{l,r}$ is approximately distributed as normal by central limit theorem as $p_k/p \to 0$.

The specification of (6) leads to a closed form marginal distribution of $\mathcal{A}(\mathbf{T}_i)$ in each cluster conditional on the auxiliary parameters $\tilde{\mathbf{\Theta}}_a = (\sigma^2, \phi)'$ by integrating out cluster specific parameters $\tilde{\mathbf{\Theta}}_h = (\theta^{(k)}_{l,r,h} : l < r)'$. More specifically,

$$\begin{aligned}\tilde{m}\left(\{\{\mathcal{A}(\mathbf{T}_i)^{(k)}\}_{l,r} : i \in \mathcal{C}_h\}|\phi, \sigma^2\right) &= \left(2\pi\sigma^2\right)^{\frac{-n_h}{2}} \left[\frac{\phi}{n_h + \phi}\right]^{\frac{1}{2}} \\ &\exp\left\{-\frac{1}{2\sigma^2}\left(\left[\sum_{i\in\mathcal{C}_h}\left(\{\mathcal{A}(\mathbf{T}_i)^{(k)}\}_{l,r} - \{\bar{\mathcal{A}}(\mathbf{T})^{(k)}_{\mathcal{C}_h}\}_{l,r}\right)^2\right] + \phi\left(\{\bar{\mathcal{A}(\mathbf{T})}^{(k)}_{\mathcal{C}_h}\}_{l,r} - \theta_0\right)^2\right)\right\},\end{aligned} \tag{7}$$

where $n_h = |\mathcal{C}_h|$ is the number of samples belonging to the h-th cluster $\mathcal{C}_h$ and $\{\bar{\mathcal{A}}(\mathbf{T})^{(k)}_{\mathcal{C}_h}\}_{l,r} = \frac{1}{n_h+\phi}\left(\sum_{i\in\mathcal{C}_h}\{\mathcal{A}(\mathbf{T}_i)^{(k)}\}_{l,r} + \phi\theta_0\right)$. The marginal distribution of $\mathcal{A}(\mathbf{T}_1), ..., \mathcal{A}(\mathbf{T}_n)$ conditional

on the auxiliary parameters σ^2 and ϕ is of the form

$$\tilde{m}\left(\mathcal{A}(\mathbf{T}_1),...,\mathcal{A}(\mathbf{T}_n)|\phi,\sigma^2\right) = \prod_{h=1}^{n(\mathcal{C})}\prod_{i\in\mathcal{C}_h}\prod_{k=1}^{K}\prod_{1\le l<r\le p_k}\tilde{m}\left(\{\{\mathcal{A}(\mathbf{T}_i)^{(k)}\}_{l,r} : i\in\mathcal{C}_h\}|\phi,\sigma^2\right), \quad (8)$$

where the form of $\tilde{m}\left(\{\{\mathcal{A}(\mathbf{T}_i)^{(k)}\}_{l,r} : i\in\mathcal{C}_h\}|\phi,\sigma^2\right)$ is obtained from (7).

Following Lau and Green (2007), the prior distribution on the partition $\mathcal{C}$ under such a specification assumes the form,

$$\pi(\mathcal{C}|\phi) = \phi^{n(\mathcal{C})+1}\frac{\Gamma(\phi)}{\Gamma(n+\phi)}\prod_{h=1}^{n(\mathcal{C})}\Gamma(n_h), \quad (9)$$

with the prior being dependent on the auxiliary parameter ϕ. Following the Chinese Restaurant analogy, (9) implies that the probability of assigning a new customer to a new table is proportional to ϕ a priori. The prior specification is completed by setting an inverse-gamma prior on σ^2, $\sigma^2 \sim IG(a_\sigma, b_\sigma)$ and a discrete uniform prior on ϕ taking values $\phi_1,...,\phi_F$ each with probability $1/F$.

3 Posterior Computation

With likelihood and prior distributions specified as in Section 2.4.1, the full posterior distribution of partitions and auxiliary variables is given by,

$$p(\mathcal{C},\phi,\sigma^2|\mathcal{A}(\mathbf{T}_1),...,\mathcal{A}(\mathbf{T}_n)) \propto \tilde{m}\left(\mathcal{A}(\mathbf{T}_1),...,\mathcal{A}(\mathbf{T}_n)|\phi,\sigma^2\right)\times\phi^{n(\mathcal{C})+1}\frac{\Gamma(\phi)}{\Gamma(n+\phi)}\prod_{h=1}^{n(\mathcal{C})}\Gamma(n_h)$$
$$\times\frac{\beta_\sigma^{\alpha_\sigma}}{\Gamma(\alpha_\sigma)}(\sigma^2)^{-\alpha_\sigma-1}\exp\left(-\frac{\beta_\sigma}{\sigma^2}\right).$$

The posterior computation proceeds following the general algorithm described in Section 2.3 with simplifications due to the prior structure. Specifically, the probability of assigning the i-th observation to a new cluster, described in (4), is proportional to

$$\tilde{m}(\mathcal{A}(\mathbf{T}_i)|\phi,\sigma^2)\times\phi.$$

Similarly, the probability of being assigned to the existing j-th cluster $\mathcal{C}_{j,-i}$, described in (5), becomes proportional to

$$\frac{\tilde{m}\left(\{\mathcal{A}(\mathbf{T}_s) : s \in \{i\} \cup \mathcal{C}_{j,-i}\}|\phi, \sigma^2\right)}{\tilde{m}\left(\{\mathcal{A}(\mathbf{T}_s) : s \in \mathcal{C}_{j,-i}\}|\phi, \sigma^2\right)} \times |\mathcal{C}_{j,-i}|.$$

Thus Chinese restaurant process assigns an observation into an existing cluster or to a new cluster depending on the size of the existing clusters, parameter ϕ and similarity of the customers (observations) already in a cluster with the new observation.

Finally, the full conditional distribution to sample σ^2 in step 3 of the algorithm is given by $\text{IG}(a_{\sigma|-}, b_{\sigma|-})$ distribution with the values of $a_{\sigma|-}$ and $b_{\sigma|-}$ are given by

$$a_{\sigma|-} = a_\sigma + \frac{n\sum_{k=1}^{K} p_k(p_k-1)}{2}$$

$$b_{\sigma|-} = b_\sigma + \frac{\sum_{h=1}^{n(\mathcal{C})}\sum_{k=1}^{K}\sum_{1\le l<r\le p_k}\left[\sum_{i\in\mathcal{C}_h}\left(\{\mathcal{A}(\mathbf{T}_i)^{(k)}\}_{l,r} - \{\bar{\mathcal{A}}(\mathbf{T})^{(k)}_{\mathcal{C}_h}\}_{l,r}\right)^2 + \phi\left(\{\bar{\mathcal{A}}(\mathbf{T})^{(k)}_{\mathcal{C}_h}\}_{l,r} - \theta_0\right)^2\right]}{2}.$$

ϕ is sampled in each iteration from a discrete uniform distribution taking values ϕ_f with probability proportional to $\tilde{m}\left(\mathcal{A}(\mathbf{T}_1), ..., \mathcal{A}(\mathbf{T}_n)|\phi_f, \sigma^2\right) \times \phi_f^{n(\mathcal{C})+1}\frac{\Gamma(\phi_f)}{\Gamma(n+\phi_f)}$, for $f = 1, ..., F$. We fix $F = 20$ throughout our empirical investigation.

4 Numerical Illustration

This section studies the clustering performance, i.e., estimation of true clusters and uncertainty in clustering, of our proposed Bayesian Tensor Clustering (BTC) approach vis-a-vis its competitors. We begin by showing performances of competitors when the true data generating distribution of tensors does not allow any difference in the mean level between clusters, i.e., the data generating distribution satisfies **Assumption A**. In Section 4.4 we show performance of the competitors when this assumption is violated.

4.1 Simulated Data Generation

We simulate $n = 100$ tensors $\mathbf{T}_1$,...,$\mathbf{T}_n$ from a finite mixture of tensor normal models with H^* mixing components given by

$$\mathbf{T}_i \sim \sum_{h=1}^{H^*} \pi_h TN(\mathbf{0}, \mathbf{\Sigma}_{1,h}, ..., \mathbf{\Sigma}_{K,h}), \ \sum_{h=1}^{H^*} \pi_h = 1. \tag{10}$$

The data generation scheme ensures that the tensors in different cluster differ only in their variability, satisfying **Assumption A**. Further, each simulated tensor is assumed to have $K = 3$ modes of dimensions $p_1 = 10$, $p_2 = 20$ and $p_3 = 30$. While our approach is scalable for a much bigger tensor size, we kept the tensor dimensions moderate in simulations to facilitate comparison with the full Bayesian model-based clustering approach. The probability of inclusion in every mixture component is set to be identical $\pi_h = 1/H^*$, resulting in clusters of similar size. The tensor mode-specific covariance matrices $\mathbf{\Sigma}_{k,h}$ are generated as sparse to aid accurate estimation of them following Lemma 2.1. More specifically, each mode-specific covariance matrix of dimension $p_k \times p_k$, $k = 1, ..., K$, is generated following the strategy described below.

1. A symmetric matrix $\mathbf{E}_k$ is constructed by setting its non-diagonal entries equal to 1 with probability α, and 0 with probability $(1 - \alpha)$. All diagonal elements are set to 0.

2. Construct $\mathbf{D}_k = \mathbf{E}_k/2 + \delta\mathbf{I}$ with δ chosen so that $\mathbf{D}_k$ has a condition number of p_k. The sparsity of $\mathbf{D}_k$ is determined by α and we refer to $(1 - \alpha)$ as the "sparsity parameter."

3. The matrix $\mathbf{\Sigma}_{k,h}$ is obtained by sampling from a G-Wishart distribution with degrees of freedom equal to $p_k + 3$ and scale matrix equal to $\mathbf{D}_k$.

We consider seven simulation cases by varying the number of clusters H^* and the sparsity parameter $(1 - \alpha)$ of covariance matrices, as shown in Table 1. The simulation results will develop understanding of how the interplay between number of clusters and the sparsity in the covariance matrices affects performance of the competitors.

4.2 Competitors and Metrics of Evaluation

Following the idea proposed in Rousseau and Mengersen (2011), we choose the number of mixture components H so that there are unoccupied cells. If H is chosen to be too small and none of the clusters is unoccupied, the analysis should be repeated for larger H. We observed $H = 20$ to be sufficient for the empirical investigation. As a competitor to BTC, we employ a static version of the Dynamic Tensor Clustering algorithm (DTC) (Sun and Li, 2019b) and Doubly-Enhanced EM algorithm (DEEM) proposed for tensor mixture models (Mai et al., 2021b) using only a modified enhanced M-Step. While our Bayesian approach allows simultaneous model-based determination of cluster number and composition, frequentist clustering techniques fix the number of clusters before implementing the clustering. In the simulation studies, we implement both DTC and DEEM by fixing the number of clusters at the truth H^*. Although this leads to somewhat unfair comparison for BTC, it is nonetheless instructive to investigate performance of BTC when the simulation design offers advantage to its competitors. Finally, we employ (10) after fixing the true number of clusters and the true values of $\mathbf{\Sigma}_{k,h}$'s for each tensor normal mixture component. This competitor is referred to as the Oracle Bayesian tensor clustering approach, where the only parameters left to estimate are the weights of the mixture components. Oracle is naturally expected to perform better than all the approaches and is used to assess the loss in performance due to various approximations in our approach. Notably, Oracle competitor is only available for simulation studies.

To assess inference on clusters from BTC, we investigate (i) the point estimate of cluster membership indicators denoted by $\hat{\mathbf{c}}$, and (ii) a heatmap of the posterior probability of any two samples belonging to the same cluster, or the co-clustering matrix $\mathbf{G}$ with the (i, j)th entry $G_{i,j} = P(c_i = c_j|\text{Data})$ (which provides a measure of the uncertainty associated with the clustering). An empirical estimate of the co-clustering matrix $\mathbf{G}$ can be obtained from the post burn-in MCMC samples of the cluster membership indices $\mathbf{c}$. With the information on true cluster configuration in simulation studies, we evaluate the quality of point estimate of clustering using the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) of the posterior cluster configurations with respect to the known cluster configuration. ARI ranges between

Table 1: Adjusted Rand Index (ARI) for competitors (BTC, DTC, DEEM, Oracle) when tensor-valued observations are simulated following (10). We consider different simulation configurations by changing the true number of clusters (H^*) and true sparsity of mode-specific covariance matrices ($1-\alpha$).

Cases	$1-\alpha$	H^*	*BTC*	*Oracle*	*DEEM*	*DTC*
1	0.9	3	0.877	1.000	0.089	0.002
2	0.9	4	0.934	1.000	0.035	0.002
3	0.8	3	0.958	1.000	0.101	0.000
4	0.8	4	0.984	1.000	0.044	0.002
5	0.7	3	0.992	1.000	0.113	-0.001
6	0.7	4	0.996	1.000	0.061	0.001
7	0.6	4	0.999	1.000	0.052	0.001

-1 and 1, with larger values indicating more agreement between cluster configurations. Notably, ARI is only available for simulation studies when the true clusters are known.

4.3 Simulation Results

Table 1 provide insights into the point estimates of the cluster structure by displaying the discrepancy between the true and the estimated clusters. BTC shows excellent clustering accuracy under all cases with ARI being close to 1. DEEM often clubs multiple clusters to a single cluster which naturally yields an under-estimation in the number of clusters and consequently, a drop in ARI values. Table 1 shows that the clustering accuracy of DEEM plummets when true number of clusters in the data increases, though sparsity does not seem to have any major impact on the clustering performance of DEEM. DTC performs clustering based on the low-rank decomposition of the mean structure of each tensor which is not conducive to capturing the cluster patterns in the present scenario, since data generating clusters mainly differ in terms of their variability. In fact, the tensors simulated from (10) are not likely to be approximated well by a low-rank decomposition, which presumably leads to the less satisfactory performance of DTC. In contrast, the "gold standard" Oracle is provided with the true covariance structure of the tensors as well as the true number of clusters; hence it identifies true clusters accurately in every simulation.

The uncertainty in clustering is displayed using the heat maps of posterior probabilities

Table 2: Adjusted Rand Index (ARI) for competitors (BTC, DTC, DEEM, Oracle) when tensor-valued observations are simulated following (11) and $\Delta = 0.3$. We consider different simulation configurations by changing the true number of clusters (H^*) and true sparsity of mode-specific covariance matrices ($1 - \alpha$).

Cases	$1 - \alpha$	H^*	*BTC*	*Oracle*	*DEEM*	*DTC*
1	0.9	3	0.878	1.000	0.124	0.007
2	0.9	4	0.998	1.000	0.157	0.010
3	0.8	3	0.961	1.000	0.114	0.005
4	0.8	4	0.979	1.000	0.081	0.005
5	0.7	3	0.991	1.000	0.132	0.008
6	0.7	4	0.998	1.000	0.061	0.002
7	0.6	4	0.920	1.000	0.062	0.004

of pairs of subjects belonging to the same cluster, or the co-clustering matrix. Figures 1 and 2 show co-clustering matrices for all competitors (except DTC) under all the simulation scenarios. Since DTC only offers point estimate of clusters, co-clustering matrix corresponding to DTC is not available. To facilitate visualization in Figures 1 and 2, subjects are ordered according to their true cluster configurations in the heatmap. Under all cases, BTC successfully recovers the true cluster structure, with little uncertainty associated with the estimator. As stated before, DEEM underestimates the number of clusters, with a very little uncertainty in the clusters. Oracle also recovers true clusters with very little uncertainty. Importantly, unlike existing model-based tensor clustering approaches, high dimensionality of tensors is a blessing rather than a curse for BTC as, with high dimensions, the transformed features can more accurately estimate the true mode-specific covariance matrices.

4.4 Clustering Performance under Mis-specification

While Sections 4.1-4.3 show excellent performance of BTC when **Assumption A** is satisfied and Lemma 2.1 holds, this section evaluates BTC when the data generation scheme violates **Assumption A**. To evaluate the performance of BTC under such mis-specification, when the no difference in means between clusters is violated, we simulate tensor-valued

observations from

$$\mathbf{T}_i \sim \sum_{h=1}^{H^*} \pi_h TN(\boldsymbol{M}_h, \boldsymbol{\Sigma}_{1,h}, ..., \boldsymbol{\Sigma}_{K,h}), \ \sum_{h=1}^{H^*} \pi_h = 1, \tag{11}$$

with $\boldsymbol{\Sigma}_{1,h}, ..., \boldsymbol{\Sigma}_{K,h}$ constructed using steps 1-3, outlined in Section 4.1. 80% entries of the cluster-specific mean entries $\boldsymbol{M}_1, ..., \boldsymbol{M}_{H^*}$ are set to 0, while the rest of the entries are simulated to ensure $||\boldsymbol{M}_h - \boldsymbol{M}_{h'}||_F/p = \Delta$ for any $1 \leq h \neq h' \leq H^*$. Two simulation settings are considered with $\Delta = 0.3$ and $\Delta = 2$, which correspond to "small" and "big" difference between cluster means. For both $\Delta = 0.3$ and $\Delta = 2$, we consider seven simulation cases by varying H^* and $(1 - \alpha)$, as shown in Table 2 and Table 3, respectively.

Table 2 demonstrates superior performance of BTC and Oracle, while DEEM and DTC struggle to identify the true clusters. As Δ increases, the performance edge of BTC and Oracle over their competitors is maintained (see Table 3). However, with higher values of Δ, clustering performance of DTC improves substantially. This is presumably due to the fact that the mean structure of the tensor-valued data plays a more significant role in clustering for larger values of Δ, which is conducive to the clustering architecture of DTC. The performance of DEEM deteriorates substantially (see Table 3) as H^* increases, though sparsity does not seem to play an important role in the performance of DEEM. Overall, The simulated data reveals excellent performance of BTC when the true clusters of observations differ substantially in terms of their variability, rather than only in their mean structure.

5 Application to the 'Eyes-Open' Paradigm Data

We illustrate performance of BTC using a dataset on EEG signals for 58 children aged 25 to 126 months with autism spectrum disorder (ASD). For each subject, EEG signals were sampled at 500 HZ for two minutes from a 128-channel HydroCel Geodesic sensor Net. EEG recordings were collected during an 'eyes-open' paradigm in which bubbles were presented on a computer screen in a sound-attenuated room to ASD children at rest. More details related to pre-processing and data acquisition can be found at Scheffler et al. (2019). The EEG data for each subject is reduced via model-based interpolation to a standard $10 - 20$ system 25 electrode montage, as described in Perrin et al. (1989), resulting in 25 electrodes

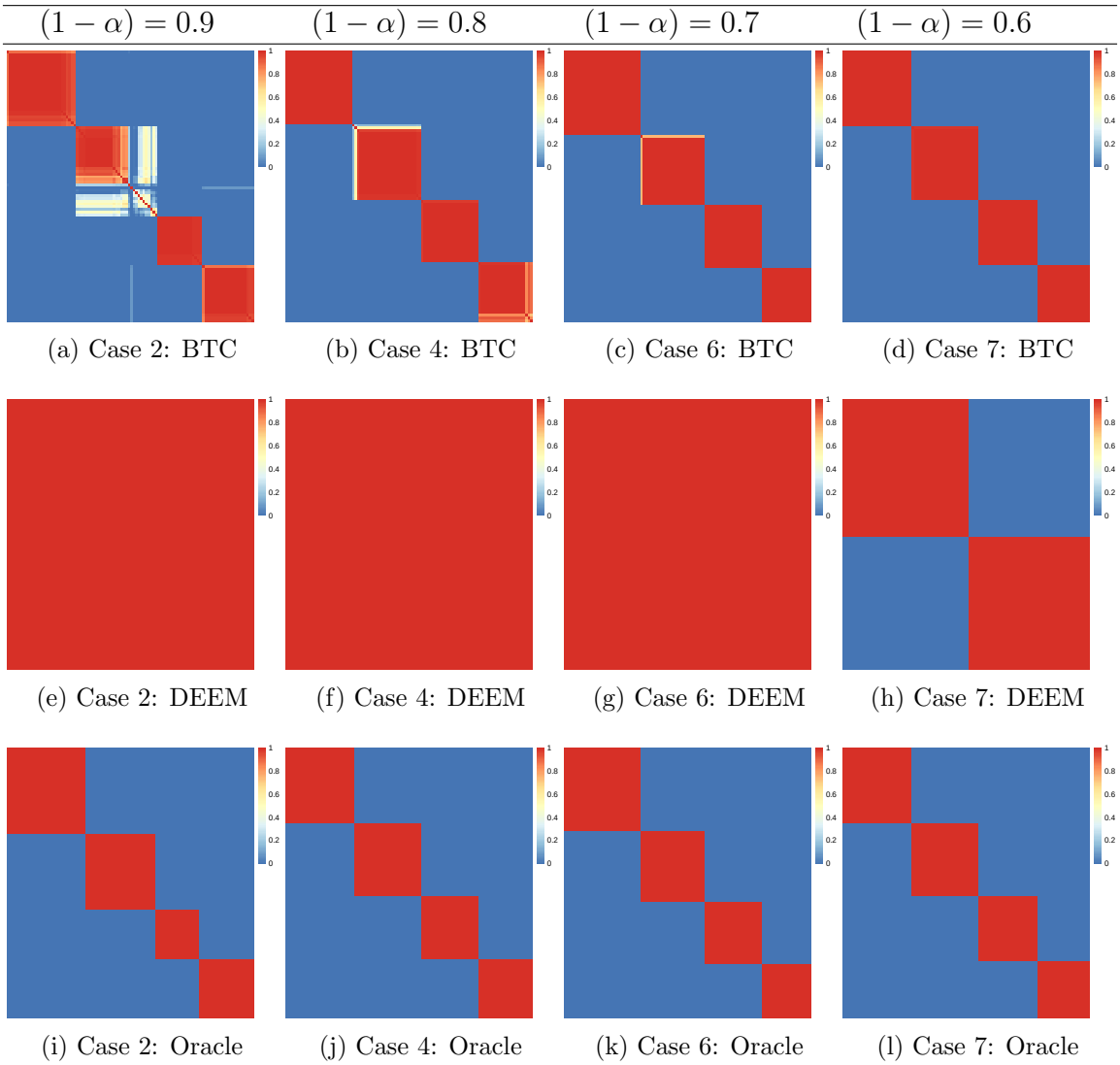
Table 3: Adjusted Rand Index (ARI) for competitors (BTC, DTC, DEEM, Oracle) when tensor-valued observations are simulated following (11) and with $\Delta = 2$. We consider different simulation configurations by changing the true number of clusters (H^*) and true sparsity of mode-specific covariance matrices ($1 - \alpha$).

Cases	$1-\alpha$	H^*	*BTC*	*Oracle*	*DEEM*	*DTC*
1	0.9	3	0.981	1.000	0.164	0.870
2	0.9	4	0.996	1.000	0.060	0.734
3	0.8	3	0.991	1.000	0.203	0.883
4	0.8	4	0.996	1.000	0.060	0.734
5	0.7	3	0.998	1.000	0.288	0.863
6	0.7	4	1.000	1.000	0.088	0.745
7	0.6	4	1.000	1.000	0.149	0.868

with continuous EEG signal. We obtained spectral density estimates on the first 38 seconds of artifact free EEG data for each electrode using the Fast Fourier Transform described in Welch (1967) with two second Hanning windows and 50 percent overlap. In our analysis, we consider only the alpha spectral band ($\Omega = (6\text{Hz}, 14\text{Hz})$) which due to the sampling scheme has a frequency resolution of 0.25Hz resulting in 33 grid points. Finally, we normalize this band to a unit area to better facilitate comparisons across electrodes and subjects. As a result we end up with 58 two-way tensors (or matrices) of dimensions 25×33. As discussed in Section 1, prior evidence suggests patients with ASD can be clustered based on patterns of EEG spectral covariation. Furthermore, previous findings on this data (Scheffler et al., 2019) reveal a common alpha spectral mean structure across development in ASD patients 2-12 years old but substantial subject-level heterogeneity in terms of alpha spectral dynamics across the scalp. Thus, in this application, it is of interest to determine how ASD patients cluster in terms of patterns of spectral covariation across development. Potential subgroups with cluster-specific covariances will be investigated for links to observed characteristics such as verbal and non-verbal intelligence quotients (VIQ and NVIQ, respectively).

While the prior scientific knowledge on the dataset indicates similar mean structure among subjects, it is instructive to further explore for more empirical evidence before we proceed to clustering with BTC. While it is hard to verify such an assumption in high dimensional objects, an exploratory analysis is presented to investigate this issue.

Figure 1: Heatmap of the posterior probability of any two samples belonging to the same cluster (co-clustering matrix) for the cases with $H^* = 4$. The tensor-valued observations are simulated following (10).



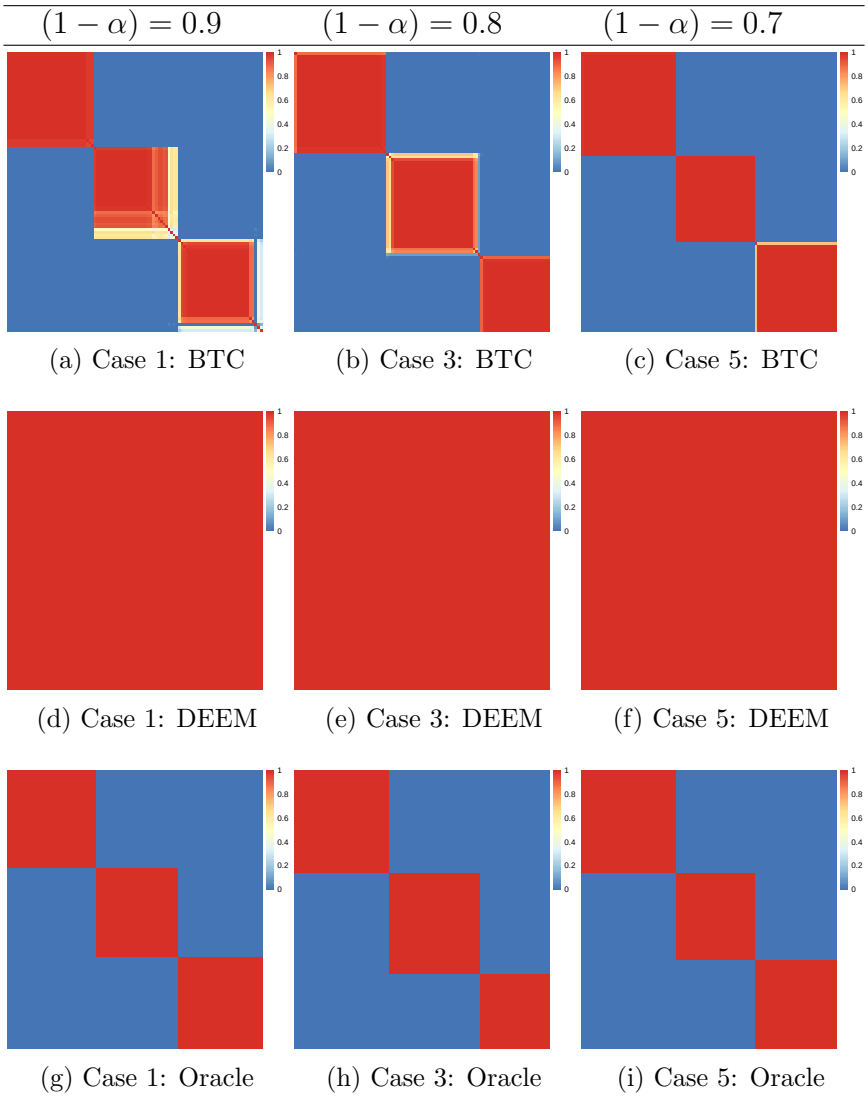
(a) Case 2: BTC (b) Case 4: BTC (c) Case 6: BTC (d) Case 7: BTC

(e) Case 2: DEEM (f) Case 4: DEEM (g) Case 6: DEEM (h) Case 7: DEEM

(i) Case 2: Oracle (j) Case 4: Oracle (k) Case 6: Oracle (l) Case 7: Oracle

As part of our exploratory analysis, we vectorize each 25×33 tensor to a long vector of 825 co-ordinates and perform k-means clustering, with $k = 2, 3, 4, 5$, separately on each of these co-ordinates. If several of the coordinates show similar clustering pattern, then one might intuitively expect a meaningful difference in the cluster means. We compute the similarity of coordinate clustering by computing the ARI of every coordinate cluster against every other coordinate cluster, resulting in $\binom{825}{2}$ ARI values. Table 4 presents the 5th, 25th, 50th, 75th and 95th percentile values for ARI corresponding to $k = 2, 3, 4, 5$. The results

Figure 2: Heatmap of the posterior probability of any two samples belonging to the same cluster (co-clustering matrix) for the cases with $H^* = 3$. The tensor-valued observations are simulated following (10).



(a) Case 1: BTC (b) Case 3: BTC (c) Case 5: BTC

(d) Case 1: DEEM (e) Case 3: DEEM (f) Case 5: DEEM

(g) Case 1: Oracle (h) Case 3: Oracle (i) Case 5: Oracle

demonstrate the distribution of the ARI is concentrated around 0 for all choices of k, offering no evidence that a significant number of coordinates results in similar clusters. Choice of higher values of k leads to even lower degree of concordance between clustering of samples along different dimensions.

With the preliminary exploration suggesting no difference in clusters in terms of mean, we proceed to identify clusters with differences in their variability using BTC. BTC shows rapid convergence and is run for 400 iterations, with inference is based on the last 300 post

Table 4: Summary statistics for the similarity of the coordinate clustering computed by the Adjusted Rand Index (ARI).

k-Means	*5th percentile*	*25th percentile*	*Median*	*75th percentile*	*95th percentile*
$k = 2$	-0.06940	-0.023240	-0.003562	0.06126	0.2623
$k = 3$	-0.02938	-0.010196	0.015847	0.06482	0.1857
$k = 4$	-0.02837	-0.005866	0.019221	0.05750	0.1400
$k = 5$	-0.02692	-0.003716	0.018981	0.05024	0.1162

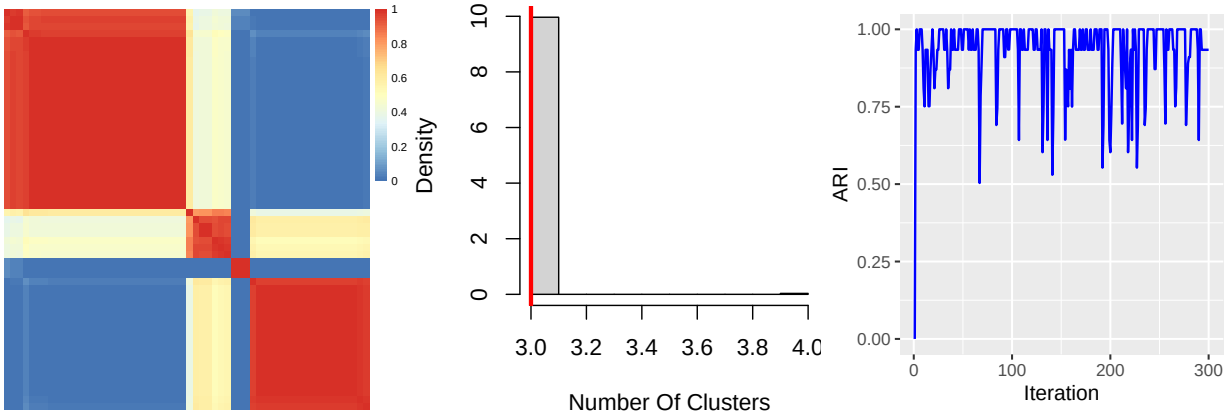
burn-in iterates. The posterior distribution of the number of clusters in Figure 3b shows a clear mode at 3, indicating three clusters among subjects. The co-clustering matrix shown in Figure 3a suggests four clusters with a high degree of uncertainty in the cluster membership for elements in the first two clusters. Indeed, the result indicates that the elements in the second cluster are often included as part of the first cluster in post burn-in iterates, which is consistent with the posterior mode of the number of clusters being identified as three. To demonstrate the stability of clusters in the post burn-in iterations, we plot (Figure 3c) ARI of clusters in any two successive post burn-in iterations. The plot indicates that most of the partitions in successive iterations are identical or have high overlaps. The nominal degree of fluctuations in the ARI stems mainly from the fact that elements in the second cluster are entirely part of the first cluster in many of the iterations.

We examine the resulting mode-specific correlation structures (electrode and frequency, respectively) for each of the clusters. Recall that the transformed features estimate mode-specific covariance matrices up to a constant following Lemma 2.1. Thus, we construct estimates of mode-specific correlation matrices from transformed features, which are comparable across clusters. We present the resulting correlation matrices for the most frequent clustering assignment which accounts for 35.33% of the post burn-in samples. Figure 4 displays the (a-c) electrode and (d-f) frequency mode-specific correlations for the most frequent clusters which show distinct patterns of variation that allow for interpretation. In the electrode dimension, clusters 1 shows stronger overall positive correlations among electrode pairs, followed by cluster 2 and 3, respectively. In the frequency dimension, clusters 1 and 2

display positive correlations among alpha oscillatory dynamics at higher frequencies (11-14 Hz) while cluster 3 shows positive correlations over a wider frequency range (8-14 Hz). Overall, the electrode and frequency mode-specific correlations for clusters 1 and 2 show similar patterns of correlation, distinct in magnitude but not direction, while cluster 3 is distinct from clusters 1 and 2 in both the magnitude and direction for the frequency mode-specific correlations. However, cluster 3 displays similar positive electrode mode-specific correlations to clusters 1 and 2.

While the clusters are identified by clustering on the resting state EEG data, it is also of interest to determine if cluster membership is associated with non-EEG clinical covariates. To this end, we investigate the three clusters identified by BTC by performing separate one-way analysis of variance (ANOVA) for the two covariates measured on the subjects, VIQ, and NVIQ, to test the null hypothesis that the cluster means are equal. The three clusters varied significantly with respect to NVIQ (p-value = 0.021) and revealed borderline significance with respect to VIQ (p-value = 0.065). Ultimately, an unsupervised tensor clustering analysis is inherently exploratory, and the identified clusters form the basis of identifying ASD phenotypes which may not be captured by the two non-EEG clinical covariates measured.

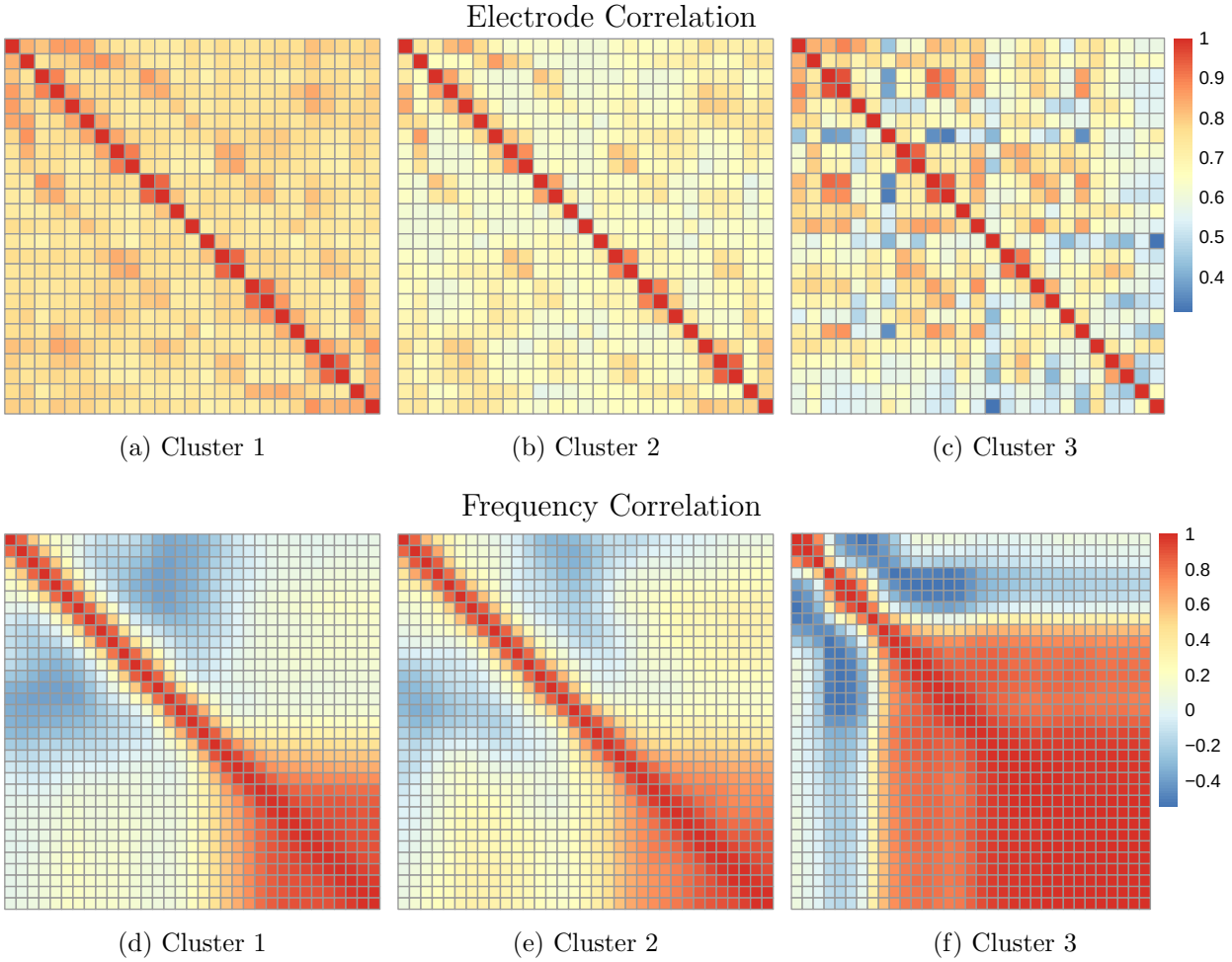
Figure 3: Figures (a) and (b) show co-clustering matrix and histogram for the number of identified clusters by BTC, respectively. Figure (c) presents ARI between any partitions for any two successive iterations for the post burn-in iterates.



(a) Cluster Probability Matrix (b) Histogram Number of Clusters (c) ARI between successive partitions

Since the size of the tensors in the EEG data application is smaller than the simulation studies, they allow fitting a full Bayesian mixture model analysis of the data using tensor

Figure 4: The (a-c) electrode and (d-f) frequency mode-specific correlations for the most frequent cluster assignments (modal clustering).

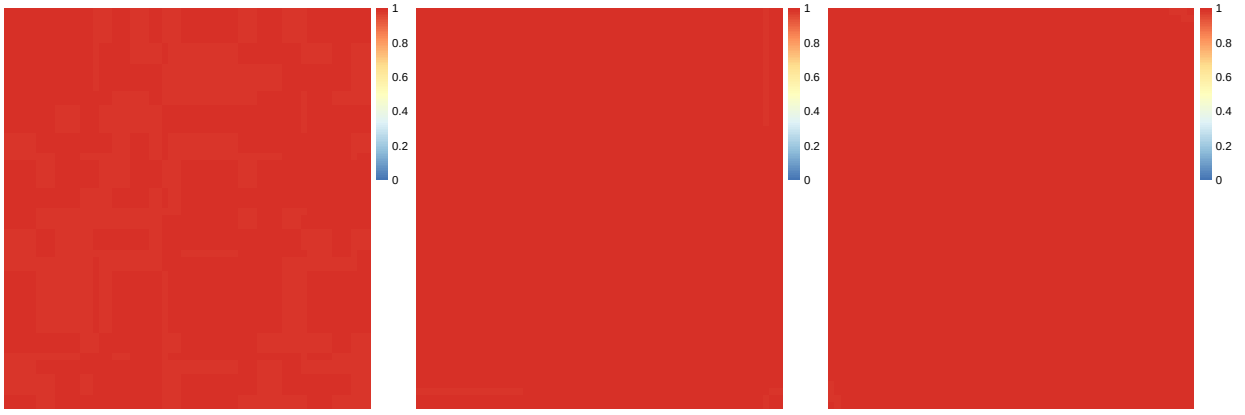


(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

(d) Cluster 1 (e) Cluster 2 (f) Cluster 3

normal distributions with zero mean as mixture components. Figure 5 presents co-clustering matrices for the full Bayesian implementation for a mixture of $H = 3, 4, 5$ tensor normal distributions. The figure demonstrates unsatisfactory performance of the full Bayesian clustering approach, identifying only one cluster. This somewhat confirms the underestimation in the number of clusters demonstrated by DEEM in the simulation studies, given that DEEM is a frequentist analogue to the Bayesian mixture model. The Bayesian mixture modeling approach should ideally offer better clustering performance than DTC, since DTC clusters tensors based on the difference in their centers. As the true model parameters are not available for the real data, the results from Oracle are not available.

Figure 5: Cluster structure for EEG data on 58 ASD children using a full Bayesian mixture of tensor normals.



(a) $H = 5$ (b) $H = 4$ (c) $H = 3$

6 Conclusion

This article presents a two-step unsupervised clustering technique for tensor-valued observations when clusters show difference mainly in their covariances, rather than in their means. Our proposed approach identifies true clusters with accurate uncertainty in misclassification, when tensor dimensions are large and sample size is moderate. Our approach aids in meaningful identification of subgroups in the ASD patients based on their EEG recordings.

An immediate future work is to extend our approach in clustering large time-varying undirected networks (represented by symmetric matrices) which frequently occur in econometric applications. The methodology developed in this article does not find straightforward extension in such a scenario since symmetric restriction in the undirected network-matrix requires modification in our proposed framework. We are currently developing novel strategies to solve this problem.

7 Acknowledgements

Rajarshi Guhaniyogi acknowledges funding from National Science Foundation Grant DMS-2220840, DMS-2210672 and Office of Naval Research Grant N00014-18-1-274. Autism Speaks Meixner Postdoctoral Fellowship in Translational Research(9292, PI: DiStefano, Men-

tor: Jeste) and the National Institutes of Health Autism Center of Excellence (2P50HD055784-08, PI: Bookheimer, Co-I: Jeste) supported the EEG data collection and maintenance.

References

Anderlucci, L., C. Viroli, et al. (2015). Covariance pattern mixture models for the analysis of multivariate heterogeneous longitudinal data. *The Annals of Applied Statistics 9*(2), 777–800.

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

Banerjee, A., S. Merugu, I. Dhillon, and J. Ghosh (2004, April). Clustering with Bregman Divergences. In *Proceedings of the 2004 SIAM International Conference on Data Mining*, pp. 234–245. Society for Industrial and Applied Mathematics.

Celeux, G., K. Kamary, G. Malsiner-Walli, J.-M. Marin, and C. P. Robert (2019). Computational solutions for Bayesian inference in mixture models. In *Handbook of Mixture Analysis*, pp. 73–96. Chapman and Hall/CRC.

Chi, E. C., B. J. Gaines, W. W. Sun, H. Zhou, and J. Yang (2020). Provable convex co-clustering of tensors. *J. Mach. Learn. Res. 21*, 214–1.

Chi, E. C. and T. G. Kolda (2012). On tensors, sparsity, and nonnegative factorizations. *SIAM Journal on Matrix Analysis and Applications 33*(4), 1272–1299.

Dickinson, A., C. DiStefano, D. Senturk, and S. S. Jeste (2018). Peak alpha frequency is a neural marker of cognitive function across the autism spectrum. *European Journal of Neuroscience 47*(6), 643–651.

Duffy, F. H. and H. Als (2012, June). A stable pattern of EEG spectral coherence distinguishes children with autism from neuro-typical controls - a large case control study. *BMC Med. 10*, 64.

Duffy, F. H. and H. Als (2019, February). Autism, spectrum or clusters? an EEG coherence study. *BMC Neurol. 19*(1), 27.

Duffy, F. H., A. Shankardass, G. B. McAnulty, and H. Als (2013, July). The relationship of asperger's syndrome to autism: a preliminary EEG coherence study. *BMC Med. 11*, 175.

Dunson, D. B. and C. Xing (2009). Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association 104*(487), 1042–1051.

Edgar, J. C., M. Dipiero, E. McBride, H. L. Green, J. Berman, M. Ku, S. Liu, L. Blaskey, E. Kuschner, M. Airey, J. L. Ross, L. Bloy, M. Kim, S. Koppers, W. Gaetz, R. T. Schultz, and T. P. L. Roberts (2019). Abnormal maturation of the resting-state peak alpha frequency in children with autism spectrum disorder. *Human Brain Mapping 40*(11), 3288–3298.

Edgar, J. C., K. Heiken, Y.-H. Chen, J. D. Herrington, V. Chow, S. Liu, L. Bloy, M. Huang, J. Pandey, K. M. Cannon, S. Qasmieh, S. E. Levy, R. T. Schultz, and T. P. L. Roberts (2015, Mar). Resting-state alpha in autism spectrum disorder and alpha associations with thalamic volume. *Journal of Autism and Developmental Disorders 45*(3), 795–804.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics 1*, 209–230.

Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association 97*(458), 611–631.

Fröhwirth-Schnatter, S. and S. Kaufmann (2008). Model-based clustering of multiple time series. *Journal of Business & Economic Statistics 26*(1), 78–89.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media.

Gao, X., W. Shen, L. Zhang, J. Hu, N. J. Fortin, R. D. Frostig, and H. Ombao (2020). Regularized matrix data clustering and its application to image analysis. *Biometrics 77*(3), 890–902.

Gopalan, R. and D. A. Berry (1998). Bayesian multiple comparisons using dirichlet process priors. *Journal of the American Statistical Association 93*(443), 1130–1139.

Guha, S. and R. Guhaniyogi (2020). Bayesian generalized sparse symmetric tensor-on-vector regression. *Technometrics 63*, 1–11.

Guhaniyogi, R., S. Qamar, and D. B. Dunson (2017). Bayesian tensor regression. *The Journal of Machine Learning Research 18*(1), 2733–2763.

Guhaniyogi, R. and D. Spencer (2018). Bayesian tensor response regression with an application to brain activation studies. Technical report, Technical report, UCSC.

Hallac, D., S. Vare, S. Boyd, and J. Leskovec (2018). Toeplitz inverse covariance-based clustering of multivariate time series data. http://arxiv.org/abs/1706.03161.

Hartigan, J. A. and M. A. Wong (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 28*(1), 100–108. Publisher: [Wiley, Royal Statistical Society].

Huang, H., C. Ding, D. Luo, and T. Li (2008). Simultaneous tensor subspace selection and clustering: the equivalence of high order svd and k-means clustering. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pp. 327–335.

Hubert, L. and P. Arabie (1985, December). Comparing partitions. *Journal of Classification 2*(1), 193–218.

Ieva, F., A. Paganoni, and N. Tarabelloni (2016). Covariance-based clustering in multivariate and functional data analysis. *Journal of Machine Learning Research 17*, 1–21.

Jegelka, S., S. Sra, and A. Banerjee (2009). Approximation algorithms for tensor clustering. In *International Conference on Algorithmic Learning Theory*, pp. 368–383. Springer.

Kolda, T. G. and B. W. Bader (2009). Tensor decompositions and applications. *SIAM review 51*(3), 455–500.

Lau, J. W. and P. J. Green (2007, September). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics 16*(3), 526–558.

Lee, J., P. Müller, Y. Zhu, and Y. Ji (2013). A nonparametric Bayesian model for local clustering with application to proteomics. *Journal of the American Statistical Association 108*(503), 775–788.

Lee, M., H. Shen, J. Z. Huang, and J. Marron (2010). Biclustering via sparse singular value decomposition. *Biometrics 66*(4), 1087–1095.

Lock, E. F. (2018). Tensor-on-tensor regression. *Journal of Computational and Graphical Statistics 27*(3), 638–647.

Mai, Q., X. Zhang, Y. Pan, and K. Deng (2021a). A doubly enhanced EM algorithm for model-based tensor clustering. *Journal of the American Statistical Association 0*(0), 1–15.

Mai, Q., X. Zhang, Y. Pan, and K. Deng (2021b, March). A Doubly Enhanced EM Algorithm for Model-Based Tensor Clustering. *Journal of the American Statistical Association 0*(0), 1–15. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2021.1904959.

Medvedovic, M. and S. Sivaganesan (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics 18*(9), 1194–1206.

Müller, P., F. A. Quintana, A. Jara, and T. Hanson (2015). *Bayesian nonparametric data analysis*. Springer.

Oh, M.-S. and A. E. Raftery (2007). Model-based clustering with dissimilarities: A Bayesian approach. *Journal of Computational and Graphical Statistics 16*(3), 559–585.

Pan, W. and X. Shen (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research 8*(5), 1145–1164.

Perrin, F., J. Pernier, O. Bertrand, and J. F. Echallier (1989, February). Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology 72*(2), 184–187.

Raftery, A. E. and N. Dean (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association 101*(473), 168–178.

Rousseau, J. and K. Mengersen (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(5), 689–710.

Scheffler, A. W., D. Telesca, C. A. Sugar, S. Jeste, A. Dickinson, C. DiStefano, and D. Şentürk (2019, December). Covariate-adjusted region-referenced generalized functional linear model for EEG data. *Statistics in medicine 38*(30), 5587–5602.

Schwartz, S., R. Kessler, T. Gaughan, and A. W. Buckley (2017, February). Electroencephalogram coherence patterns in autism: An updated review. *Pediatr. Neurol. 67*, 7–22.

Spencer, D., R. Guhaniyogi, and R. Prado (2020). Joint Bayesian estimation of voxel activation and inter-regional connectivity in fMRI experiments. *Psychometrika 85*, 1–25.

Sun, W. W. and L. Li (2019a). Dynamic tensor clustering. *Journal of the American Statistical Association 114*(528), 1894–1907.

Sun, W. W. and L. Li (2019b, October). Dynamic Tensor Clustering. *Journal of the American Statistical Association 114*(528), 1894–1907.

Tan, K. M. and D. M. Witten (2014). Sparse biclustering of transposable data. *Journal of Computational and Graphical Statistics 23*(4), 985–1008.

Viroli, C. (2011). Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing 21*(4), 511–522.

Wang, M. and Y. Zeng (2019). Multiway clustering via tensor block models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 32. Curran Associates, Inc.

Wang, S. and J. Zhu (2008). Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics 64*(2), 440–448.

Welch, P. (1967, June). The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions*

on Audio and Electroacoustics 15(2), 70–73. Conference Name: IEEE Transactions on Audio and Electroacoustics.

Zhong, S. and J. Ghosh (2003). A unified framework for model-based clustering. *The Journal of Machine Learning Research 4*, 1001–1037.

Zhou, H., L. Li, and H. Zhu (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association 108*(502), 540–552.