Infinite von Mises-Fisher Mixture Modeling of Whole Brain fMRI Data

Røge, Rasmus; Madsen, Kristoffer Hougaard; Schmidt, Mikkel Nørgaard; Mørup, Morten

Published in:
Neural Computation

Link to article, DOI:
10.1162/neco_a_01000

Publication date:
2017

Document Version
Peer reviewed version

LETTER **Communicated by Zhanyu Ma**

Infinite von Mises–Fisher Mixture Modeling of Whole Brain fMRI Data

Rasmus E. Røge
rasr@dtu.dk
Section for Cognitive Systems, DTU Compute, Technical University of Denmark, DK-2800, Kgs. Lyngby, Denmark

Kristoffer H. Madsen
khma@dtu.dk
Section for Cognitive Systems, DTU Compute, Technical University of Denmark, DK-2800, Kgs. Lyngby, Denmark, and Danish Research Centre for Magnetic Resonance, Centre for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital Hvidovre, DK-2650 Hvidovre, Denmark

Mikkel N. Schmidt
mncs@dtu.dk
Morten Mørup
mmor@dtu.dk
Section for Cognitive Systems, DTU Compute, Technical University of Denmark, DK-2800, Kgs. Lyngby, Denmark

Cluster analysis of functional magnetic resonance imaging (fMRI) data is often performed using gaussian mixture models, but when the time series are standardized such that the data reside on a hypersphere, this modeling assumption is questionable. The consequences of ignoring the underlying spherical manifold are rarely analyzed, in part due to the computational challenges imposed by directional statistics. In this letter, we discuss a Bayesian von Mises–Fisher (vMF) mixture model for data on the unit hypersphere and present an efficient inference procedure based on collapsed Markov chain Monte Carlo sampling. Comparing the vMF and gaussian mixture models on synthetic data, we demonstrate that the vMF model has a slight advantage inferring the true underlying clustering when compared to gaussian-based models on data generated from both a mixture of vMFs and a mixture of gaussians subsequently normalized. Thus, when performing model selection, the two models are not in agreement. Analyzing multisubject whole brain resting-state fMRI data from healthy adult subjects, we find that the vMF mixture model is considerably more reliable than the gaussian mixture model when comparing solutions across models trained on different groups of subjects, and again we find that the two models disagree on the optimal number

Neural Computation **29, 1–30** (2017)
doi:10.1162/NECO_a_01000

of components. The analysis indicates that the fMRI data support more than a thousand clusters, and we confirm this is not a result of overfitting by demonstrating better prediction on data from held-out subjects. Our results highlight the utility of using directional statistics to model standardized fMRI data and demonstrate that whole brain segmentation of fMRI data requires a very large number of functional units in order to adequately account for the discernible statistical patterns in the data.

1 Introduction

In many areas of statistical modeling, data are represented only by a direction, thus setting the stage for directional statistics (Mardia & Jupp, 2009). This is perhaps most easily seen when the data consist of measures of directions in three-dimensional space, such as the directions of radiation beams used for treatment (Bangert, Hennig, & Oelfke, 2010), directions from the earth to stars (Mardia & Jupp, 2009), locating emergency transmitters (Guttorp & Lockhart, 1988), microphone beamforming (Anderson, Teal, & Poletti, 2015), and modeling structure from spherical cameras (Guan & Smith, 2017). One of the most frequently used directional distributions is the von Mises–Fisher distribution (vMF) (Fisher, 1953; Mardia & El-Atoum, 1976). The vMF distribution is specified by a concentration parameter and a mean direction, and because it is part of the exponential family, it has a conjugate prior. Unfortunately, the normalization constant of the conjugate prior is not available in closed form, which makes the vMF distribution more challenging to work with (Nuñez-Antonio & Gutiérrez-Pena, 2005) compared to, say, the gaussian distribution.

Models based on the vMF distribution have been applied to a wide variety of high-dimensional problems on the unit hypersphere. This includes document topic modeling (Banerjee, Dhillon, Ghosh, & Sra, 2005; Gopal & Yang, 2014), the modeling of gene expressions data (Banerjee et al., 2005; Taghia, Ma, & Leijon, 2014), and modeling line spectral frequencies (Ma, Taghia, Kleijn, Leijon, & Guo, 2015). Within the field of neuroscience, normalizing or z-scoring the data is a common step in the preprocessing pipeline for functional magnetic resonance imaging (fMRI) analysis (Craddock, James, Holtzheimer, Hu, & Mayberg, 2012; Hyde & Jesmanowicz, 2012). By z-scoring, the data are transformed such that each voxel time series has zero mean and unit standard deviation—that is, each voxel time series consisting of D brain volumes will be projected onto the hypersphere with radius $\sqrt{D-1}$. Since there is no longer any information in the magnitude of the observations, as all voxels have the same magnitude, the magnitude can be disregarded and the data modeled using directional statistics. This makes the von Mises–Fisher a natural first choice for modeling the standardized fMRI time series. Time series data from several substructures of the brain, including the insula and striatum, were recently modeled

using a mixture model based on the von Mises–Fisher distributions with Markov random field to ensure spatial contiguity (Ryali, Chen, Supekar, & Menon, 2013). The von Mises–Fisher distribution has also been frequently used in modeling fMRI task activations (Lashkari, Vul, Kanwisher, & Golland, 2010; Lashkari & Golland, 2009) and vectors of functional connectivity with a number of regions of interest (Yeo et al., 2011). These studies, however, either focused on low-dimensional representations of high-dimensional time series by extracting task-activated b-maps (Lashkari et al., 2010; Vul, Lashkari, Hsieh, Golland, & Kanwisher, 2012) or considered fMRI time series only within a small region of interest (Ryali et al., 2013). Furthermore, neither of these studies provided a systematic comparison of the vMF with the gaussian distribution assumption when modeling fMRI. It is therefore unclear what the benefits of imposing the more challenging vMF distribution might be, as opposed to applying the well-studied and simpler gaussian distribution. Despite the directional nature of the z-scored fMRI time series data, modeling is still most often based on assumptions of gaussian distributions (Janssen, Jylänki, Kessels, & van Gerven, 2015).

In this letter, we advance the vMF mixture model to large-scale fMRI clustering. We employ collapsed Markov chain Monte Carlo (MCMC) inference and exploit nonparametric Bayesian modeling for model order quantification. We apply the developed framework to multisubject whole brain fMRI segmentation and contrast the performance of the vMF distribution assumption to the conventional gaussian assumption. We thus present a thorough comparison with gaussian mixture models based on identical inference procedures, such that we isolate the differences that are caused by the difference in probabilistic modeling assumptions from what could be caused by potential difference in inference implementation. We investigate the models on synthetic data with ground truth as well as on large-scale multisubject fMRI data and contrast the estimated model order based on nonparametric Bayesian modeling to the model order estimated using the predictive distribution based on finite mixtures.

The letter is structured as follows. In section 2, we introduce the generative models and inference procedure for our nonparametric vMF mixture model. In section 3, we present results regarding the implementation of the vMF models. We apply our model to multisubject resting-state fMRI data and contrast the performance to conventional parametric and nonparametric gaussian mixture modeling. Finally, in section 4, we present our conclusions. In the appendix, we compare our implementation to an existing implementation based on variational inference (Gopal & Yang, 2014).

2 Methods

Clustering using a mixture of vMF distributions was introduced by Banerjee et al. (2005), who proposed an inference procedure using expectation-maximization (EM). Due to the occurrence of the Bessel function in the vMF

probability density function, they relied on an approximation to determine the concentration parameter of the vMF distribution and provided bounds for the accuracy of the approximation. Focusing on the three-dimensional case, Bangert et al. (2010) extended the model to a nonparametric "infinite" vMF mixture, and presented a Markov chain Monte Carlo (MCMC) inference procedure, combining Gibbs sampling and slice sampling. Recently, Taghia et al. (2014) and Gopal and Yang (2014) independently proposed variational inference procedures for finite mixtures of vMF distributions using the gamma distribution and log-normal distribution, respectively, as prior for the concentration parameter. The variational inference method requires some extra work to estimate the concentration parameter, which can be performed using an approximation (Taghia et al., 2014; Gopal & Yang, 2014) or by MCMC sampling (Gopal & Yang, 2014). In contrast to variational inference, MCMC sampling yields an unbiased estimate of the true posterior and may thus have some advantages over variational inference. The downside is that it is computationally demanding and may not converge for larger problems despite providing a useful approximation.

In this letter, we present the Bayesian generative model for clustering directional data based on vMF distributions. Similar to Bangert et al. (2010) we formulate a nonparametric mixture model and base our inference on MCMC sampling; however, we improve on the inference procedure by analytically marginalizing over the mean parameter, as opposed to sampling it, and we apply the model to high-dimensional problems, where Bangert et al. (2010) considered only the three-dimensional case. We carefully investigate the effect of using only few samples to approximate the integration of the concentration parameter in this collapsed distribution, leading to a computationally more efficient inference procedure.

2.1 The von Mises–Fisher Mixture Model. In this section, after a brief review of the vMF distribution, we present the vMF mixture model along with the numerical approximations, a description of the inference procedure, and posterior quantities used for the subsequent analyses.

2.1.1 The von Mises–Fisher Distribution. The vMF distribution is a distribution over unit vectors on the hypersphere and is defined by a mean direction parameter $\boldsymbol{\mu} \in \mathbb{S}^{D-1}$, where $\mathbb{S}^{D-1} = \{\boldsymbol{x} \in \mathbb{R}^D : \|\boldsymbol{x}\| = 1\}$ and a concentration parameter $\tau \in (0, \infty)$. For a given unit vector $\boldsymbol{x} \in \mathbb{S}^D$, the vMF probability density is given by

$$\text{vMF}(\boldsymbol{x} \mid \boldsymbol{\mu}, \tau) = C_D(\tau) \exp(\tau \boldsymbol{\mu}^\top \boldsymbol{x}), \tag{2.1}$$

where

$$C_D(\tau) = \frac{\tau^{D/2-1}}{(2\pi)^{D/2} \mathcal{I}_{D/2-1}(\tau)}, \tag{2.2}$$

and $\mathcal{I}_{D/2-1}(\tau)$ is the modified Bessel function of the first kind of order $D/2 - 1$ and argument τ. The vMF distribution with parameters $\{\boldsymbol{\mu}_0, \tau_0\}$ is in itself a conjugate prior for the mean direction.

For N observations from a vMF distribution with concentration τ, the marginal likelihood for τ is given by

$$p(\boldsymbol{x}_{1:N} \mid \tau) = \int \text{vMF}(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, \tau) \prod_{i=1}^{N} \text{vMF}(\boldsymbol{x}_i \mid \boldsymbol{\mu}, \tau) d\boldsymbol{\mu}$$
$$= \frac{C_D(\tau)^{N+1}}{C_D\left(\tau \left\| \boldsymbol{\mu}_0 + \sum_{i=1}^{N} \boldsymbol{x}_i \right\|\right)}. \tag{2.3}$$

Therefore, if we apply a prior given by

$$f(\tau \mid a, b) \propto \frac{C_D(\tau)^a}{C_D(b\tau)}, \tag{2.4}$$

with parameters a and b, where $a > b > 0$, it corresponds to having seen a observations from a vMF distribution that has the combined length b (cf. Hornik & Grün 2013). The normalization constant for this prior is not available in closed form due to the dependence on the modified Bessel functions. Previous implementations have used either the log-normal or gamma distribution (Taghia et al., 2014; Gopal & Yang, 2014) as priors, but as Taghia et al. (2014) showed, the gamma distribution very closely resembles the above prior we have chosen for our implementation. For our implementation, there is no computational advantage in using the gamma or log-normal distributions, and we therefore use that of equation 2.4.

2.1.2 Prior Distributions for Cluster Assignments. A natural choice for a probability distribution for the cluster assignments, which we denote by $\boldsymbol{z}$, is the compound Dirichlet-categorical distribution, also known as the Pólya distribution: it posits that each observation belongs to cluster k with probability π_k and that the cluster proportions π_k are generated from a symmetric Dirichlet distribution with parameter $\frac{\alpha}{K}$. When the cluster proportions are marginalized, the resulting Pólya distribution with parameter $\alpha > 0$ is given by

$$\text{Pólya}(\boldsymbol{z} \mid \alpha) = \frac{\Gamma(\alpha)}{\Gamma(N + \alpha)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})}, \tag{2.5}$$

where N is the number of observations, K is the number of clusters, and n_k is the number of observations that belong to cluster k. Taking the limit $K \to \infty$

of the Pólya distribution yields the so-called Chinese restaurant process (CRP) (Aldous, 1985; Pitman, 2006),

$$\mathrm{CRP}(z \mid \alpha) = \frac{\Gamma(\alpha)\alpha^K}{\Gamma(N+\alpha)} \prod_{k=1}^{K} \Gamma(n_k), \tag{2.6}$$

where now K denotes the number of nonempty clusters. Utilizing the nonparametric nature of the CRP, the number of clusters can be directly inferred from data, while it must be fixed in advance when using the Pólya distribution. Both distributions enforce the rich-get-richer principle in which higher-probability mass is assigned to large clusters, to a degree controlled by the parameter α. In this work, we have implemented both variants to assess if the theoretical advantage of the CRP is also apparent in practice.

2.1.3 Mixture Model Specification. Modeling data with a mixture of multiple vMF distributions is the classical mixture model. To complete model specification for the finite or the infinite case, we need to include the Pólya distribution or Chinese restaurant process as prior on the clustering. The vMF mixture model is then given by the following generative process:

$$\tau_k \mid a, b \quad \sim f(a, b) \qquad k = 1, \ldots, K, \tag{2.7}$$

$$\boldsymbol{\mu}_k \mid \boldsymbol{\mu}_0, \tau_0 \quad \sim \mathrm{vMF}(\boldsymbol{\mu}_0, \tau_0) \qquad k = 1, \ldots, K, \tag{2.8}$$

$$\boldsymbol{x}_i \mid \boldsymbol{\mu}_{z(i)}, \tau_{z(i)} \sim \mathrm{vMF}(\boldsymbol{\mu}_{z(i)}, \tau_{z(i)}) \quad i = 1, \ldots, N, \tag{2.9}$$

where $\boldsymbol{x}_i$, $\boldsymbol{\mu}_k$, and $\boldsymbol{\mu}_0$ are vectors on the D-dimensional hypersphere and $f(a, b)$ the normalized prior for the concentration parameter from equation 2.4. The joint probability of the generative model is given by

$$p(\boldsymbol{x}_{1:N}, \boldsymbol{\mu}_{1:K}, \tau_{1:K} \mid \boldsymbol{z}, \boldsymbol{\mu}_0, \tau_0, a, b) = \left[\prod_{i=1}^{N} \mathrm{vMF}(\boldsymbol{x}_i \mid \boldsymbol{\mu}_{z(i)}, \tau_{z(i)})\right] \left[\prod_{k=1}^{K} \mathrm{vMF}(\boldsymbol{\mu}_k \mid \boldsymbol{\mu}_0, \tau_0)\right] f(\tau_k \mid a, b). \tag{2.10}$$

To marginalize the cluster mean direction parameters, we turn our attention to the terms of the joint distribution related to cluster k, which are given by

$$p(\boldsymbol{x}_{\mathcal{Z}_k}, \boldsymbol{\mu}_k \mid \boldsymbol{z}, \boldsymbol{\mu}_0, \tau_0, \tau_k) = \left[\prod_{i \in \mathcal{Z}_k} \mathrm{vMF}(\boldsymbol{x}_i \mid \boldsymbol{\mu}_k, \tau_k)\right] \mathrm{vMF}(\boldsymbol{\mu}_k \mid \boldsymbol{\mu}_0, \tau_0) \tag{2.11}$$

$$= C_D(\tau_0)C_D(\tau)^{n_k} \exp(\lambda_k \boldsymbol{m}_k^\top \boldsymbol{\mu}_k), \tag{2.12}$$

where $\mathcal{Z}_k = \{i \in 1, \ldots, N : z_i = k\}$ is the index set of observations in cluster k and

$$\lambda_k = \left\| \tau_0 \boldsymbol{\mu}_0 + \tau_k \sum_{i \in \mathcal{Z}_k} \boldsymbol{x}_i \right\|, \quad \boldsymbol{m}_k = \frac{1}{\lambda_k}\left(\tau_0 \boldsymbol{\mu}_0 + \tau_k \sum_{i \in \mathcal{Z}_k} \boldsymbol{x}_i \right). \tag{2.13}$$

By conjugacy, we can now marginalize $\boldsymbol{\mu}_k$ analytically,

$$p(\boldsymbol{x}_{\mathcal{Z}_k} \mid \boldsymbol{z}, \boldsymbol{\mu}_0, \tau_0, \tau_k) = \int p(\boldsymbol{x}_{\mathcal{Z}_k}, \boldsymbol{\mu}_k \mid \boldsymbol{z}, \boldsymbol{\mu}_0, \tau_0, \tau_k) d\boldsymbol{\mu}_k = \frac{C_D(\tau_0)C_D(\tau_k)^{n_k}}{C_D(\lambda_k)}, \tag{2.14}$$

where n_k is the number of elements in cluster k. We further marginalize the concentration parameter τ_k:

$$\begin{aligned} p(\boldsymbol{x}_{\mathcal{Z}_k} \mid \boldsymbol{z}, \boldsymbol{\mu}_0, \tau_0, a, b) &= \int p(\boldsymbol{x}_{\mathcal{Z}_k} \mid \boldsymbol{z}, \boldsymbol{\mu}_0, \tau_0, \tau_k) f(\tau_k \mid a, b) d\tau_k \\ &= C_D(\tau_0) \int \frac{C_D(\tau_k)^{n_k}}{C_D(\lambda_k)} f(\tau_k \mid a, b) d\tau_k. \end{aligned} \tag{2.15}$$

As this unidimensional integral is analytically intractable, we approximate it using MCMC integration. One approach could be to perform joint MCMC inference of the cluster labels and the concentration parameters; however, numerically marginalizing the concentration parameters significantly simplifies the MCMC inference for the cluster labels by allowing for a standard Gibbs sampling approach. Therefore, we take the approach of marginalizing the concentration parameters in a separate step, as we discuss next.

2.1.4 MCMC Approximation for the Concentration Parameter. If we simulate S samples, $\{\tau_k^{(s)}\}$, from $f(\tau_k \mid a, b)$, then the integral in equation 2.15 can be approximated as

$$\int \frac{C_D(\tau_k)^{n_k}}{C_D(\lambda_k)} f(\tau_k \mid a, b) d\tau_k \approx \frac{1}{S} \sum_{s=1}^{S} \frac{C_D(\tau_k^{(s)})^{n_k}}{C_D(\lambda_k^{(s)})}. \tag{2.16}$$

It is possible to use a number of different sampling techniques to simulate independent samples from the prior. In our implementation, we used Metropolis-Hastings sampling, discarded the first 200 samples as burn-in, and used a thinning factor of 20 to get approximately independent samples. Note that Metropolis-Hastings sampling does not require the distribution to be normalized.

Only when the values of the hyperparameters a or b change, the prior $f(\tau \mid a, b)$ will change and thus require sampling a new set of of $\tau_k^{(s)}$'s. The number of samples used in the approximation will affect the overall accuracy of the algorithm. If only a few samples are used for approximating the integral, there is a higher risk of accepting a poor proposal for a or b. Similarly, if few samples are used, we might not recover the correct clustering. However, the computational complexity of the inference procedure scales linearly with the number of samples used to approximate the integral, and it is thus beneficial to use as few samples as possible that still provide accurate inference.

The numerical integration requires the evaluation of $C_D(\tau)$, which in turn requires the evaluation of $\mathcal{I}_\nu(x)$ for some values of ν and x. Using the Matlab function `besseli`, we noted that issues with overflow or underflow would sometimes arise. To avoid this issue, we use a large-order approximation for $\nu > 10$ (Hornik & Grün, 2014):

$$\log \mathcal{I}_\nu(x) \approx \sqrt{x^2 + (\nu+1)^2} + (\nu + 1/2) \log \frac{x}{\nu + 1/2 + \sqrt{x^2 + (\nu+1)^2}} + \frac{1}{2} \log x/2 + (\nu + 1/2) \log \frac{2\nu + 3/2}{2(\nu+1)} - \frac{\log 2\pi}{2}. \tag{2.17}$$

Using this numerical integration, we obtain the following expression for the collapsed joint distribution (disregarding the prior on the clustering parameter $\boldsymbol{z}$):

$$p(\boldsymbol{x}_{1:N} \mid \boldsymbol{z}, \boldsymbol{\mu}_0, \tau_0, a, b) = \prod_k \frac{C_D(\tau_0)}{S} \sum_{s=1}^{S} \frac{C_D(\tau_k^{(s)})^{n_k}}{C_D(\lambda_k^{(s)})}. \tag{2.18}$$

2.1.5 Inference. Having analytically marginalized $\boldsymbol{\mu}_k$ and numerically integrated τ_k, inference reduces to standard Gibbs sampling for the cluster assignments $\boldsymbol{z}$ combined with updates for the hyperparameters τ_0, a, and b. For the infinite model with the CRP as a prior for the clustering, the posterior distribution for assigning the ith element to the kth component using Gibbs sampling is (up to proportionality) given by

$$p(\boldsymbol{z}_i = k \mid \boldsymbol{z}_{\setminus i}, \ldots) \propto n_k \frac{\sum_{s=1}^{S} \frac{C_D(\tau_k^{(s)})^{n_k+1}}{C_D\left(\left\| \tau_0 \boldsymbol{\mu}_0 + \tau_k^{(s)} [\boldsymbol{x}_i + \sum_{j \in \mathcal{Z}_k} \boldsymbol{x}_j] \right\|\right)}}{\sum_{s=1}^{S} \frac{C_D(\tau_k^{(s)})^{n_k}}{C_D\left(\left\| \tau_0 \boldsymbol{\mu}_0 + \tau_k^{(s)} \sum_{j \in \mathcal{Z}_k} \boldsymbol{x}_j \right\|\right)}}, \tag{2.19}$$

with the convention that observation i has been removed from $\mathcal{Z}_{z_i}$. This is derived using Bayes' theorem and combining equation 2.18 with equation

2.6. For assigning observation i to a new cluster, this gives (again up to proportionality) the following posterior:

$$p(z_i = K+1 \mid \boldsymbol{z}_{\setminus i}, \ldots) \propto \frac{\alpha C_D(\tau_0)}{S} \sum_{s=1}^{S} \frac{C_D(\tau_k^{(s)})}{C_D\big(\|\tau_0 \boldsymbol{\mu}_0 + \tau_k^{(s)} \boldsymbol{x}_i\|\big)}. \tag{2.20}$$

We further apply the split-merge algorithm (Jain & Neal, 2004) with accelerated merge moves (Røge, Madsen, Schmidt, & Mørup, 2015) for faster convergence.

The version of the model with the Pólya distribution is identical except that in the posterior conditional distribution for each cluster in equation 2.19, the factor n_k must be replaced by $n_k + \frac{\alpha}{K}$. The split-merge algorithm is not applicable to finite mixture models, and the procedure is thus omitted from the inference in that case.

To infer the hyperparameters τ_0, a, and b, we use Metropolis-Hastings sampling. The parameter τ_0 is required to be positive, and we therefore use a log transform to facilitate the use of the symmetric normal distribution as a proposal distribution. Furthermore, the parameters a and b have the constraint that $a > b > 0$, and we therefore apply the appropriately truncated gaussian proposal distributions. We impose the improper and relatively uninformative prior $p(\theta) = \theta^{-1}$ on each of the hyperparameters $\tau_{0,s}$, a, and b, with the additional constraint that $a \geq b$. We keep $\boldsymbol{\mu}_0$ parameter fixed at the mean of the data.

2.1.6 Multiple Data Set Analysis. The models can be straightforwardly extended to multiple data sets that share the clustering configuration. To construct the generative model in this case, we use the CRP or Pólya distribution as prior for the clustering configuration and then take the product of the joint distribution in equation 2.10 over the multiple data sets. This approach is frequently used in fMRI data analysis when fMRI scans from multiple subjects are acquired, and it is not unreasonable to assume that the clustering should be the same over subjects after spatial normalization (Craddock et al., 2012). In our implementation, the subjects share the same hyperparameters for τ_0, a, and b, while $\boldsymbol{\mu}_0$ is fixed for each subject as the mean time series of all voxels from the subject.

2.1.7 Posterior Quantities. We can use Bayes' theorem to obtain the posterior probability for the concentration parameter:

$$p(\tau_k \mid \boldsymbol{x}, a, b) = \frac{p(\boldsymbol{x} \mid \tau, a, b) p(\tau \mid a, b)}{\int p(\boldsymbol{x} \mid \tau, a, b) p(\tau \mid a, b) d\tau}. \tag{2.21}$$

This is proportional to

$$p(\tau_k \mid \boldsymbol{x}, a, b) \propto \frac{C_D(\tau_k)^{n_k} C_D(\tau_k)^a}{C_D(\lambda_k) C_D(b\tau_k)}. \tag{2.22}$$

This enables us to compute the radii of the confidence regions and the posterior curves for the concentration parameter.

Similarly, we can obtain the posterior probability for the mean direction conditioned on the concentration:

$$p(\boldsymbol{\mu}_k \mid \boldsymbol{X}, \tau_k, \mu_0, \tau_0) \propto p(\boldsymbol{X} \mid \boldsymbol{\mu}_k, \tau_k) p(\boldsymbol{\mu}_k \mid \tau_0, \mu_0) \propto \exp(\lambda_k \boldsymbol{m}_k^\top \boldsymbol{\mu}_k). \tag{2.23}$$

Since this is the functional form of a vMF distribution, we know the normalization constant and obtain

$$p(\boldsymbol{\mu}_k \mid \boldsymbol{X}, \tau_k, \mu_0, \tau_0) = C_D(\lambda_k) \exp(\lambda_k \boldsymbol{m}_k^\top \boldsymbol{\mu}_k). \tag{2.24}$$

2.2 Gaussian Mixture Model. For comparison, we include two versions of the gaussian mixture model with both the Pólya distribution and CRP as priors for the clustering configuration for comparing the difference between modeling data on the hypersphere and ignoring the underlying manifold. The gaussian mixture model with the CRP prior is known as the infinite Gaussian mixture model and was introduced by Rasmussen (1999).

The multivariate gaussian mixture model can be defined with the covariance matrix being either a scaled identity matrix (spherical), a diagonal matrix (elliptical), or a full matrix. The computational complexity of models with the spherical or elliptical covariance scales linearly in D, while the full covariance model scales with D^2, thus rendering it intractable for large problems. The gaussian models with the spherical covariance structure most closely resemble that of the vMF distribution, and for completeness, we include both the spherical and elliptical gaussian mixture models in our analyses.

The generative model for the mixture of gaussians with axis-aligned elliptical covariance structure is given by

$$\sigma^2_{m,k} | \nu, \gamma \sim \mathrm{IG}(\nu, \gamma) \qquad m = 1, \ldots, D \quad k = 1, \ldots, K, \tag{2.25}$$

$$\boldsymbol{\mu}_k | \gamma, \boldsymbol{\sigma}_k^2 \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \frac{1}{\lambda}\mathrm{diag}(\boldsymbol{\sigma}_k^2)\right) \quad k = 1, \ldots, K, \tag{2.26}$$

$$\boldsymbol{x}_i | \boldsymbol{\mu}_{z_i}, \boldsymbol{\sigma}_k^2 \sim \mathcal{N}(\boldsymbol{\mu}_{z_i}, \mathrm{diag}(\boldsymbol{\sigma}_k^2)) \qquad i = 1, \ldots, N, \tag{2.27}$$

where IG is the inverse gamma distribution and $\mathrm{diag}(\boldsymbol{\sigma}_k^2)$ the diagonal matrix with the elements of $\boldsymbol{\sigma}_k^2$ on the diagonal. The collapsed joint distribution is, in concordance with the procedure for the vMF-based model, obtained

by marginalizing over the mean $\boldsymbol{\mu}_k$ and noise $\boldsymbol{\sigma}_k^2$ parameters

$$\begin{aligned} p(\boldsymbol{x}_{1:N}, \boldsymbol{z} \mid \boldsymbol{\theta}) &= \prod_{k=1}^{K} \iint \prod_{i \in k} p(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2) p(\boldsymbol{\mu}_k \mid \boldsymbol{\sigma}_k^2) p(\boldsymbol{\sigma}_k^2) d\boldsymbol{\mu}_k d\boldsymbol{\sigma}_k^2 \\ &= \prod_{k=1}^{K} \prod_{m=1}^{D} \frac{(\lambda/[n_k + \lambda])^{1/2} \gamma^{\nu} \Gamma(n_k/2 + \nu)}{(2\pi)^{n_k/2} \Gamma(\nu) R_{mk}^{n_k/2+\nu}}, \end{aligned} \tag{2.28}$$

where

$$R_{mk} = \gamma + \frac{1}{2} \left(\bar{\sigma}_{mk}^2 + \lambda \mu_{0_m}^2 - \frac{(\bar{x}_k + \lambda \mu_{0_m})^2}{n_k + \lambda} \right), \tag{2.29}$$

and $\bar{\sigma}_{mk}^2 = \sum_{n \in \mathcal{Z}_k} x_{mn}^2$ and $\bar{x}_k = \sum_{n \in k} \boldsymbol{x}_n$. For the spherical gaussian mixture model, the generative model is given by

$$\sigma_k^2 | \nu, \gamma \sim \mathrm{IG}(\nu, \gamma) \qquad k = 1, \ldots, K, \tag{2.30}$$

$$\boldsymbol{\mu}_k | \sigma_k^2, \lambda \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \frac{\sigma_k^2}{\lambda} \boldsymbol{I} \right) \quad k = 1, \ldots, K, \tag{2.31}$$

$$\boldsymbol{x}_i | \boldsymbol{\mu}_k, \sigma_k^2 \sim \mathcal{N}(\boldsymbol{\mu}_{z_i}, \sigma_k^2 \boldsymbol{I}) \quad i = 1, \ldots, N, \tag{2.32}$$

and the collapsed joint distribution is given by

$$p(\boldsymbol{x}_{1:N}, \boldsymbol{z} \mid \boldsymbol{\theta}) = \prod_{k=1}^{K} \frac{(\lambda/[n_k + \lambda])^{D/2} \gamma^{\nu} \Gamma(Dn_k/2 + \nu)}{(2\pi)^{Dn_k/2} \Gamma(\nu) R_k^{Dn_k/2+\nu}}, \tag{2.33}$$

where, with $\bar{\sigma}_k^2 = \sum_{n \in k} \|\boldsymbol{x}\|$ and $\bar{\boldsymbol{x}}_k = \sum_{n : z_n = k} \boldsymbol{x}_n$,

$$R_k = \boldsymbol{\gamma} + \frac{1}{2} \left(\bar{\sigma}_k^2 + \lambda \|\boldsymbol{\mu}_0\|^2 - \frac{\|\bar{\boldsymbol{x}}_k + \lambda \boldsymbol{\mu}_0\|^2}{n_k + \lambda} \right). \tag{2.34}$$

We apply the same inference procedure as with the vMF mixture model with suitable priors and transformations on the hyperparameters.

2.2.1 Predictive Analysis. To evaluate how well the model, when estimated on training data, is able to characterize unseen test data, we evaluate the predictive likelihood, which in general is given by

$$p(\boldsymbol{x}^* \mid \boldsymbol{X}) = \int p(\boldsymbol{x}^* \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{X}) d\boldsymbol{\theta}, \tag{2.35}$$

where $\boldsymbol{X}$ is training data, $\boldsymbol{x}^*$ is test data, and $\boldsymbol{\theta}$ are the parameters of the model.

In case we are given a test data set that shares the same clustering and has a one-to-one correspondence with the training data, such that each test observation has a known corresponding training observation, the predictive likelihood can be computed directly from the MCMC approximation. After generating a sample of M parameter sets from the posterior, $\{\boldsymbol{\theta}^{(m)}\} \sim p(\boldsymbol{\theta}|\boldsymbol{X})$, we can compute the Monte Carlo estimate:

$$p(\boldsymbol{x}^* \mid \boldsymbol{X}) \approx \frac{1}{M} \sum_{m=1}^{M} p(\boldsymbol{x}^* \mid \boldsymbol{\theta}^{(m)}). \tag{2.36}$$

In case we are given a new observation but no information regarding which cluster it belongs to, we can compute the predictive likelihood by the following procedure. First, we sum over each cluster,

$$p(\boldsymbol{x}^* \mid \boldsymbol{X}) = \sum_{k=1}^{K} p(z_{\boldsymbol{x}^*} = k \mid \boldsymbol{X}) p(\boldsymbol{x}^* \mid \boldsymbol{X}, z_{\boldsymbol{x}^*} = k), \tag{2.37}$$

where $p(z_{\boldsymbol{x}^*} = k \mid \boldsymbol{X})$ is the posterior predictive distribution of the clustering. For the infinite models, we need to sum over all populated clusters, as well as one unpopulated cluster. We evaluate the expression by approximation using samples drawn from the posterior distribution during inference,

$$\begin{aligned} & p(\boldsymbol{x}^* \mid \boldsymbol{X}, z_{\boldsymbol{x}^*} = k) \\ &= \int p(\boldsymbol{x}^* \mid \boldsymbol{X}, z_{\boldsymbol{x}^*} = k, \tau_0, \boldsymbol{z}, a, b) p(\tau_0, \boldsymbol{z}, a, b \mid \boldsymbol{X}) d\{\tau_0, \boldsymbol{z}, a, b\} \\ &= \frac{1}{M} \sum_{m=1}^{M} p(\boldsymbol{x}^* \mid \boldsymbol{X}, z_{\boldsymbol{x}^*} = k, \tau_0^{(m)}, \boldsymbol{z}^{(m)}, a^{(m)}, b^{(m)}). \end{aligned} \tag{2.38}$$

We can compute part of this expression analytically and part with MCMC samples from the posterior of τ_k as given by equation 2.22:

$$\begin{aligned} & p(\boldsymbol{x}^* \mid \boldsymbol{X}, z_{\boldsymbol{x}^*} = k, \tau_0, \boldsymbol{z}, a, b) \\ &= \iint p(\boldsymbol{x}^* \mid \boldsymbol{\mu}_k, \tau_k) p(\boldsymbol{\mu}_k \mid \boldsymbol{X}, \boldsymbol{\mu}_0, \tau_0) d\boldsymbol{\mu}_k p(\tau_k \mid \boldsymbol{X}, a, b) d\tau_k \\ &= \int \frac{C_D(\tau_k) C_D(\lambda_k^{(s)})}{C_D(\lambda_k^{*(s)})} p(\tau_k \mid \boldsymbol{X}, a, b) d\tau_k, \end{aligned} \tag{2.39}$$

where $\lambda_k^{*(s)} = \|\lambda_k \boldsymbol{m}_k + \tau_k \boldsymbol{x}^*\|$.

2.3 Initialization. It is not clear how the hyperparameters of the models are best initialized. If the parameters initially set such that the level of noise in the model is much too high compared to the variance of the data, the MCMC sampler will often collapse everything into one cluster and learn hyperparameters that reinforce that solution. Similarly, if the level of noise is too low, all elements are often placed in singleton clusters. In both cases, the model is initialized near a bad local posterior mode, which the MCMC sampler struggles to escape from.

We therefore investigate several different initialization strategies that build on the idea of providing an appropriate initialization of the clustering followed by Metropolis-Hastings proposals to infer reasonable values for the hyperparameters before running the full inference procedure.

2.4 Measures of Cluster Validity. We use two methods to quantify and compare the quality of the clustering methods on fMRI data: reliability of inferred clustering and the homogeneity and cluster separation of the inferred clustering.

To quantify the reliability of the inferred clusters, we use three frequently used measures of similarity between clusterings: normalized mutual information (NMI; Strehl & Ghosh, 2002), adjusted mutual information (AMI; Vinh, Epps, & Bailey, 2010), and the adjusted Rand index (AR; Hubert & Arabie, 1985). The NMI and AMI measures have several variants, and we have used the following:

$$\mathrm{NMI} = \frac{\mathrm{MI}(z_z, z_2)}{\sqrt{H(z_1)H(z_2)}} \tag{2.40}$$

and

$$\mathrm{AMI} = \frac{\mathrm{MI}(z_z, z_2) - \mathrm{E}[\mathrm{MI}(z_z, z_2)]}{\max(H(z_1), H(z_2)) - \mathrm{E}[\mathrm{MI}(z_z, z_2)]}, \tag{2.41}$$

where $\mathrm{MI}(z_z, z_2)$ is the mutual information between clusterings z_z and z_2, H is the entropy, and E[MI] is the expected mutual information, which is the expectation for random clusterings of the given number of clusters. The adjusted Rand index is given by

$$\mathrm{AR} = \frac{\mathrm{RI} - \mathrm{E}[\mathrm{RI}]}{\max(\mathrm{RI}) - \mathrm{E}[\mathrm{RI}]}, \tag{2.42}$$

where RI is the Rand index and E[RI] is the expected Rand index. These adjusted measures are a way of compensating for the fact that two random clusterings tend to have a higher Rand index and normalized mutual information as the number of clusters increases and should therefore be a better

measure for comparing the reliability of two clusterings that have a different number of clusters.

The silhouette index (SI) is a measure of both the homogeneity of clusters and intercluster distance (Goutte, Toft, Rostrup, Nielsen, & Hansen, 1999; Craddock et al., 2012). For observation i, the silhouette value is given by

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)}, \tag{2.43}$$

where a_i is the average distance from observation i to the observations in the same cluster and b_i is the minimum average distance from observation i to the remaining clusters. In this letter, we use the correlation of the time series as the distance measure similar to Craddock et al. (2012) and report the silhouette index averaged over all observations.

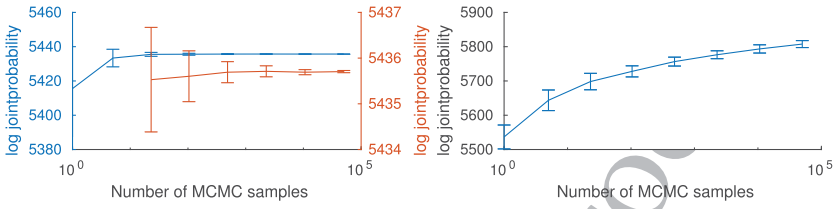
2.5 Implementation. Both the gaussian and the vMF mixture models have been implemented in `Matlab` in an object-oriented framework. This means that the code for the Gibbs and split-merge sampling can be reused and that the framework is easily extendable with additional statistical clustering models. Our code and examples are available at https://brainconnectivity.compute.dtu.dk.

Variational-inference-based vMF clustering models have previously been applied to a variety of document topic modeling data sets (Gopal & Yang, 2014). The code is not available online, and we therefore compare the results of our implementation to theirs on the publicly available CNAE-9 data set (Gopal & Yang, 2014) based on normalized mutual information between the inferred clusters and ground truth. This comparison can be found in the Appendix, where we observe that our implementation is at least on par with the variation inference based procedure.

3 Results and Discussion

To analyze aspects of the proposed vMF model related to the MCMC integration technique and initialization strategy and to illustrate and compare the model to the GMM, we first applied the models to synthetic data simulated from the generative model such that ground truth about the clustering was known. Next, we applied the model to a multisubject resting-state fMRI data set and compared the results with the GMM approach.

3.1 Analysis of MCMC Integration. The computational complexity of the inference procedure scales linearly with the number of samples used to approximate the integral described in section 2.4; however, if an insufficient number of samples is used, the inference procedure will not provide a good data fit.



(a) $\text{avg}(\tau_k) = 30$, $\text{std}(\tau_k) = 10$ (b) $(\tau_1, \tau_2) = (0.1, 400)$, $\text{avg}(\tau_{3:10}) = 30$, $\text{std}(\tau_{3:10}) = 10$

Figure 1: The likelihood of the model as a function of the number of MCMC samples used to approximate the integral. The data used for this comparison are generated according to the vMF model with parameters $D = 50$, $N = 200$, $K = 10$. The red graph is a close-up of the blue.

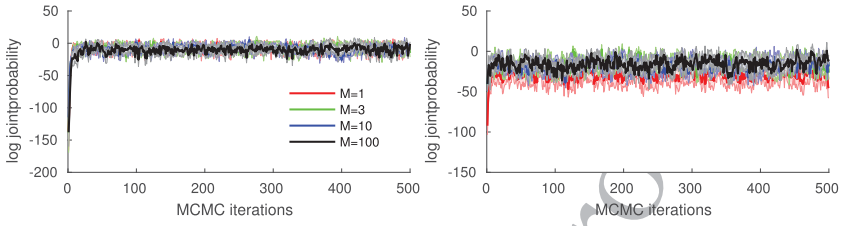
To analyze how many samples are needed, we generated a small data set according to the model (we fixed the clustering to $K = 10$ with 20 elements in each cluster), generated $\tau_k \sim \mathcal{N}(\tau_{\text{avg}}, \tau_{\text{std}})$ for $(\tau_{\text{avg}}, \tau_{\text{std}}) = (30, 2)$, and finally generated $\boldsymbol{x}_i \sim \text{vMF}(\boldsymbol{e}_1, \tau_k)$, where $\boldsymbol{e}_1$ is the first canonical unit vector for each $i = 1, \ldots, N$. This procedure was used for the generation of each of the synthetic data sets used for the analyses in this section. We then varied the number of samples used to approximate the integral. The results are presented in Figure 1. With only one sample used to approximate the integral, the standard deviation of the approximated integral is less than 3% of the actual value.

In order to answer the question of how many samples are needed for inference to converge, we generated a number of data sets and ran the inference procedure with a varying number of samples used to approximate the integral. From the results given in Figures 2a and 2b, we observe that on data sets with low variance in the concentration parameter between clusters, it is sufficient with only one sample, whereas increasing variance also increases the required number of samples.

For each of the following applications, we used 30 samples for the approximation of the integral based on ad hoc tests on each of the data sets.

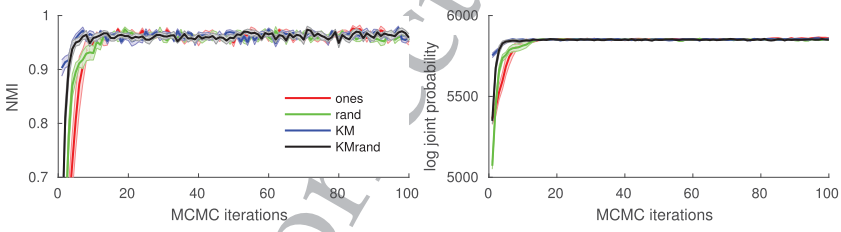
3.2 Analysis of Initialization. To investigate the impact of initialization, we compared four initialization strategies on synthetic data:

ones: Initializing all elements to the same cluster followed by the evaluation of 100 MCMC proposals for each hyperparameter
rand: Initializing each label at random among K clusters followed by the evaluation of 100 MCMC proposals for each hyperparameter



(a) std$(\tau_k) = 2$ (b) std$(\tau_k) = 20$

Figure 2: The effect of different numbers of samples used in approximating the integral on inference. As the clusters differ more in concentration parameter, the more samples are needed for sufficient inference. The colored regions are $\pm$ the standard deviation over six restarts on different data sets. The data used for this comparison are generated according to the vMF mixture model with parameters $D = 50$, $N = 200$, $K = 10$, avg$(\tau_k) = 30$.



(a) (b)

Figure 3: Comparing four initialization strategies. The solid lines are the mean of the runs, and the colored areas are $\pm$ the standard deviation. The data used for this comparison are generated according to the vMF model with parameters $D = 50$, $N = 200$, $K = 10$, $\tau_{\text{avg}} = 35$, $\tau_{\text{std}} = 2$.

KM: Initializing the clustering to a K-means solution followed by the evaluation of 100 MCMC proposals for each hyperparameter

KMrand: Like KM but assigning each label at random after learning hyperparameters

We initialized the model with each of the four initialization strategies and performed 200 MCMC iterations to infer the clustering and parameters. We repeated this six times, and the results are given in Figure 3. We achieve must faster convergence with the K-means initialization but also observe that the other initialization strategies reach similar solutions when the models have converged. For the remainder of the letter, we have used the *KMrand* initialization strategy as it avoids initializing to a local minimum.

3.3 A Three-Dimensional Example. To illustrate the model, we generated a small three-dimensional data set of six clusters with 40 elements in each cluster. We generated the mean directions from a vMF distribution with mean direction as the third canonical unit vector e_3 and $\tau_0 = 0.01$. For each cluster, we generated the concentration parameter randomly from $\mathcal{N}(50, 20^2)$.

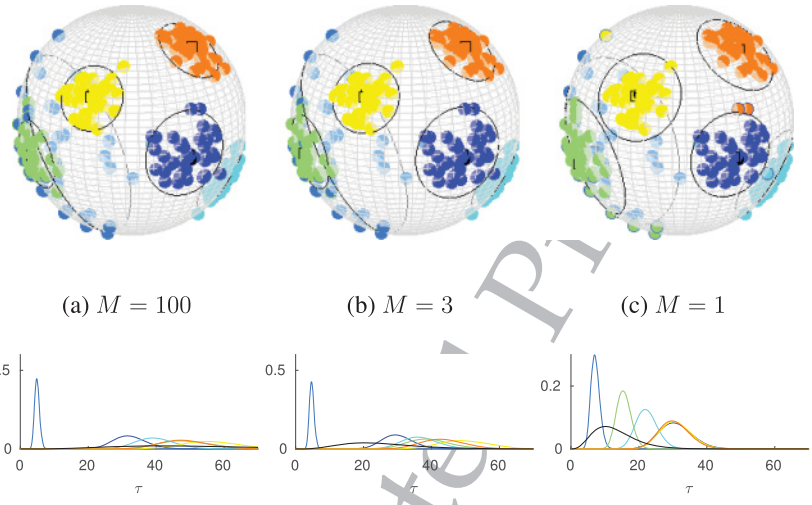
We ran three samplers using either 1, 3, or 100 samples to estimate the integral and stopped the inference chains after 200 Monte Carlo iterations. For each of the three runs, we present the clustering from the highest likelihood sample in Figures 4a to 4c. The circles on the spheres represent the 95% credibility regions. To emphasize the difference, we plot the posterior distribution for the concentration parameters for the prior and for each of the clusters in Figure 4d.

Finally, we present the log joint probability and the NMI for each iteration of the inference chains in Figures 4e and 4f. There are significant differences in the inference using only a single sample, while the difference between using 3 and 100 samples is negligible. For the chain with a single sample, we see that the mode of the posterior densities is concentrated too heavily around the prior compared to the other two chains, and therefore the confidence regions are either too small or too large.

3.4 Comparison of GMM and vMF. The comparison of mixture models based on different probability distributions on simulated data sets is inherently dependent on the parameters used for generating the data, such as the probability distribution used for generating the data, number of observations, number of clusters, and the temporal dimension of the data. The parameters used for generating the data set in this section are selected to illustrate the differences between using gaussian and vMF-based mixture models.

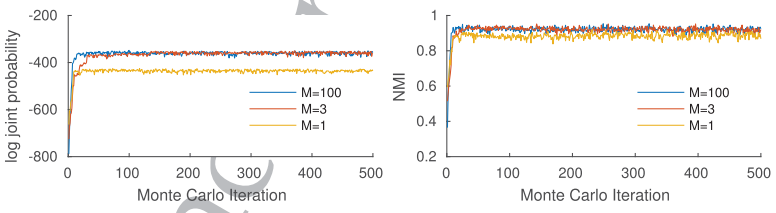
To analyze the differences between using a gaussian and vMF-based mixture model, we generated several data sets according to the generative model for both the mixture of vMF distributions and the mixture of gaussians that is subsequently normalized to the unit sphere. Each data set is generated with $N = 1000$ observations with $D = 240$ divided into $K = 50$ clusters of equal size and vary the average and standard deviation of the noise of each cluster such that $\text{avg}(\sigma_k) = \text{std}(\sigma_k)$ and $\text{avg}(\tau) = \text{std}(\tau)$ for each of the $k = 1, \ldots, K$ clusters for the gaussian and vMF-based data, respectively. We apply standard K-means and run inference in the vMF, GMMs, and GMMd models for 100 MCMC iterations with identical initializations, which the plots of the log-likelihood (not included) indicate are sufficient for convergence. For each level of noise, we repeat the experiment 20 times.

The highest likelihood sample is selected for each sampling chain, and the AMI with the true clustering in given in Figure 5. We see a clear advantage for the vMF model on data generated from a mixture of vMFs for



(a) $M = 100$ (b) $M = 3$ (c) $M = 1$

(d) Posterior density for each τ_k matched in colors to the picture above and in black the prior density.



(e) The log joint probability over MCMC iterations.

(f) The NMI over MCMC iterations.

Figure 4: The 3D example. The data on the sphere are presented on the top. The center color denotes the actual clustering, and the border is the clustering inferred. The 95% credibility region is marked by the black circle. For the generated data, the model is able to infer the correct clustering.

average concentration parameters of the generated data between $\text{avg}(\tau_k) = 50$ and 85 and a minor advantage to the vMF model when the noise is higher than $\text{avg}(\sigma_k) = 10$ for data generated from a mixture of gaussians.

Next, we explored how the gaussian and vMF nonparametric models handled data generated from a mixture of vMFs with parameters $N = 100$, $K = 5$, $D = 30$, $\tau_0 = 30$, $\tau_{\text{std}} = 25$, and $\tau_{\text{mean}} = \{20, 25, \text{ and } 30\}$ for high noise, medium noise, and low noise data sets, respectively. For each of the

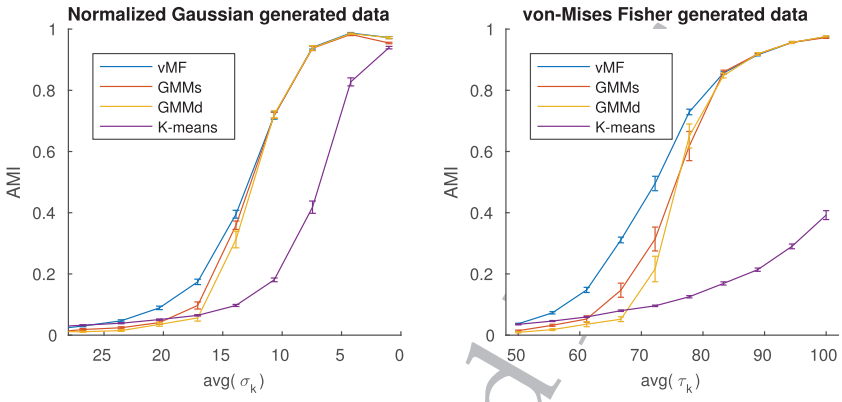
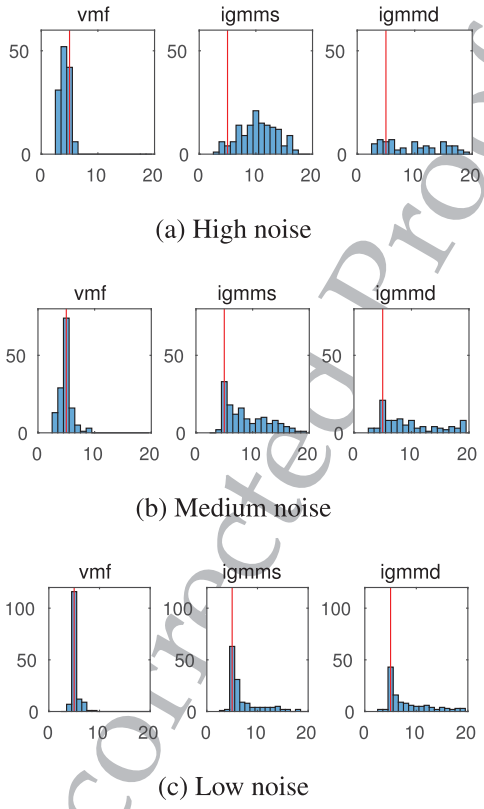


Figure 5: Results on the two simulated data sets based on adjusted mutual information (AMI).

three settings, we generated $\tau_k \sim \mathcal{N}(\tau_{\text{mean}}, \tau_{\text{std}})$ and then generated the data set. We ran the infinite vMF, GMMs, and GMMd models for 200 MCMC iterations for each dataset.

Results, based on the number of clusters in the highest-likelihood sample, are presented in Figure 6. It is clear that the vMF-based nonparametric models infer a number of clusters much closer to the truth compared to both the spherical and elliptical gaussian mixture models. This emphasizes the importance of modeling data using directional statistics in determining the complexity of a data set.

3.5 Resting-State fMRI Analysis. Functional brain connectivity can be assessed by analyzing fluctuations in the blood oxygenation level–dependent signal (BOLD). Statistical dependencies across brain areas are typically measured by correlation such that highly correlated regions constitute estimates of functional networks. Resting state (i.e., fMRI recorded during rest, without explicit task) has become prominent for probing functional connectivity in the resting brain (Biswal et al., 2010). Often these functional networks are extracted by defining a seed region and evaluating correlation to this region throughout the brain (Biswal, Yetkin, Haughton, & Hyde, 1995). Rather than specifying seeds, clustering methods extract prominent latent activation profiles and identify corresponding brain networks (Craddock et al., 2012). These latent class models are useful as they do not rely on a priori specification of seeds and can provide an overview of the functional organization across large high-dimensional data sets. The interpretation of these networks hinges on their reliability. However, latent variable models can be plagued by issues of reproducibility across data splits; thus, reliability is an important issue to address for their utility



(a) High noise

(b) Medium noise

(c) Low noise

Figure 6: Results from nonparametric synthetic analysis. In the top image of each, the NMI between the inferred solutions and truth for each repetition of the experiment can be seen, and on the bottom, histograms of the inferred number of clusters in the solutions. We see that the vMF-based nonparametric mixture models in general find solutions closer to the truth in terms of both NMI and the inferred number of clusters.

(Strother et al., 2002; Thirion, Varoquaux, Dohmatob, & Poline, 2014; Churchill, Madsen, & Mørup, 2016). As correlation is formed by the inner product of standardized fMRI time series, thus naturally complying with the vMF distribution assumptions, the vMF mixture model is attractive for clustering resting-state fMRI data as clusters are explicitly formed by their correlation to the extracted latent activation profiles.

In this study, we apply the clustering models to a resting-state fMRI data set consisting of 30 healthy subjects scanned on a Siemens 3T MRI scanner. The data set has been previously used in Andersen et al. (2014). During the functional scans, the participants were instructed to keep their eyes closed

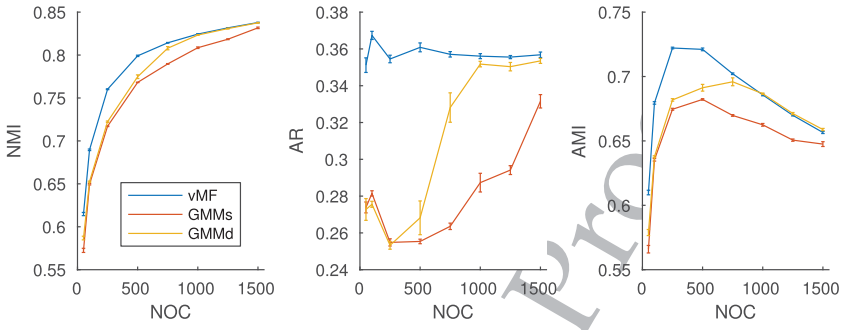


Figure 7: The normalized mutual information (NMI), rand index (AR), and adjusted mutual information (AMI) between groups of five subjects.

and refrain from any voluntary motor or cognitive activity while the 480 brain volumes were scanned over 20 minutes with a repetition time of 2.49 s.

Data were preprocessed using the SPM12 software package (SPM12, Wellcome Trust Centre for Neuroimaging, http://www.fil.ion.ucl.ac.uk/spm/software/spm12/) with the following steps: (1) rigid body realignment, (2) coregistration, (3) spatial normalization to the Montreal Neurological Institute (MNI) 152 template, (4) reslicing of images into MNI space at 3 mm isotropic voxels, and (5) spatial smoothing was applied with a 6 mm full-width, half maximum isotropic gaussian filter. Finally, a rough gray matter mask consisting of 48,799 voxels was applied.

We divide the data set into two groups of five subjects, and for each group we select the first 240 brain volumes, allowing us to quantify the generalizability of the clustering to new subjects. Then we apply the parametric models (vMF, GMMs, and GMMd) with number of clusters, $K = \{50, 100, 250, 500, 750, 1000, 1250, 1500\}$ to the time series data using the KMrand initialization strategy. For each model, we perform 100 MCMC iterations and repeat the process four times on each of the two data sets for each of the three models and for each of the four settings of K, resulting in a total of 96 runs. Note that although we apply sampling-based inference, the solutions found will be subject to local maxima and suffer from poor mixing due to the size of the problem.

We evaluate the results based on three different metrics of similarity between the clusterings inferred across the two groups of five subjects: normalized mutual information (NMI), adjusted mutual information (AMI), and the adjusted Rand index (AR). The results are presented in Figure 7, and in Figures 8 and 9, the best likelihood sample from the vMF, GMMs, and GMMd models are visualized with axial slices and surface plots for the solutions with 100 and 250 regions of interest. The GMMd model in a slightly different formulation with a distance-dependent Chinese restaurant process prior on the clustering has produced promising results in

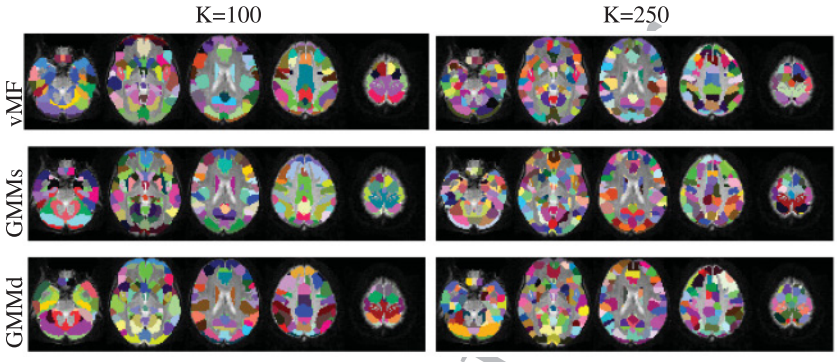


Figure 8: Visual comparison of axial slices of the solutions from the vMF, GMMs, and GMMd clustering methods for $K = 100$ and $K = 250$.

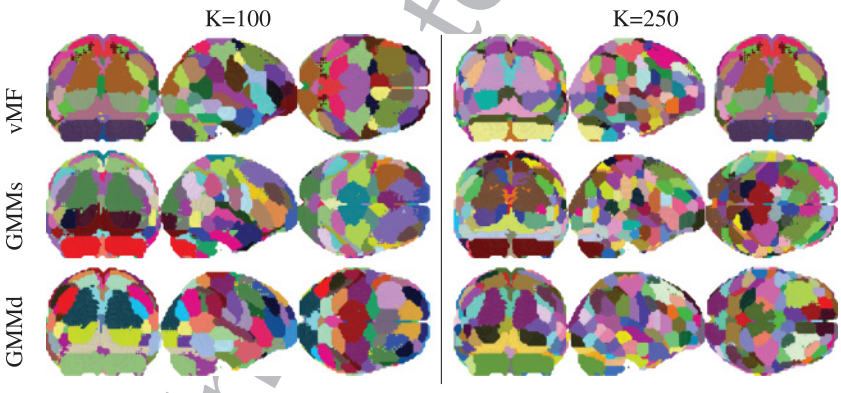


Figure 9: Visual comparison of the surface of the clustering solutions from the vMF, GMMs, and GMMd clustering methods for $K = 100$ and $K = 250$.

parcellating the Striatum (Janssen, Jyländki, Kessels, & van Gerven, 2015). The results here clearly show that the vMF-based model outperforms both the models based on gaussian densities in terms of all three measures of similarity of the obtained clusterings, thus providing a more reliable whole brain segmentation.

The results with silhouette score emphasize the utility of the vMF-based mixture model, as seen in Figure 10.

As an example, we inspect the segmentation of the striatum and the insula as delineated by the automated anatomical labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002) by visualizing the highest likelihood sample of the $K = 100$ and the $K = 250$ vMF parcellations. (See Figure 11.) For both the striatum and the insula, we see a clear symmetry across hemispheres

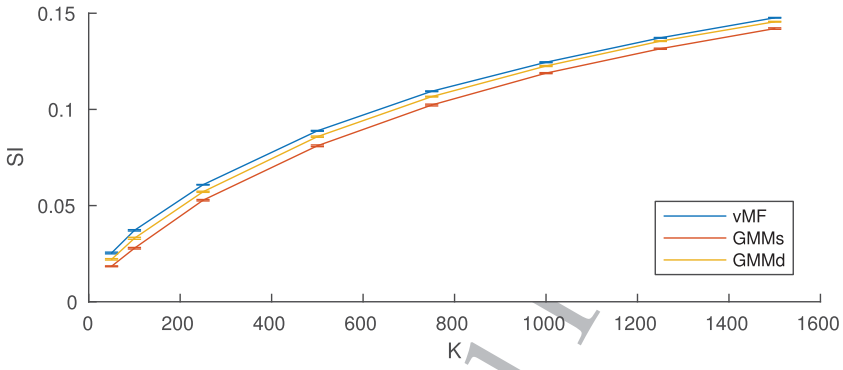
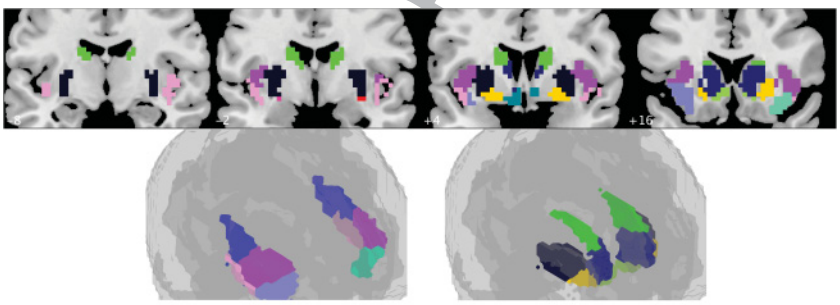
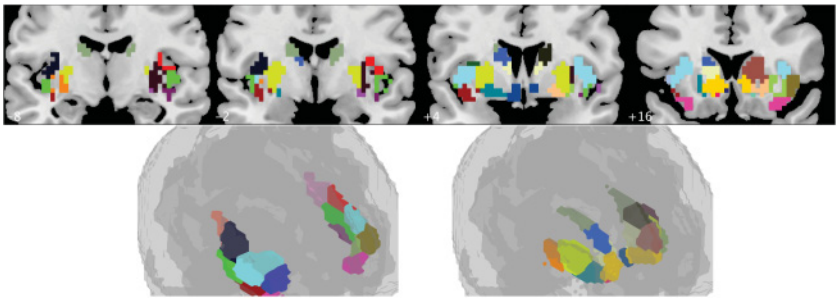


Figure 10: The silhouette index for the brain clusterings from the parcelations on the groups of five subjects.

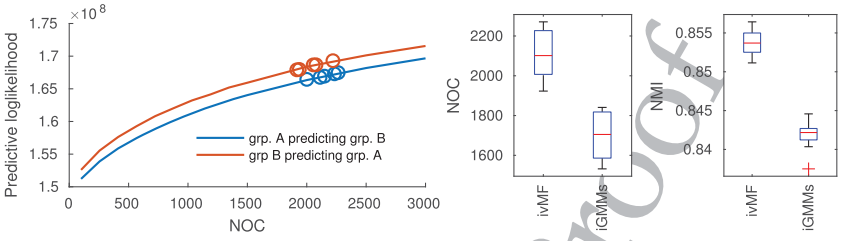


(a) K=100



(b) K=250

Figure 11: Coronal brain slices and a surface visualization of the segmentation of the insula (left) and the striatum (right) for the highest-likelihood sample of the vMF clustering.



(a) (b)

Figure 12: The predictive likelihood as a function of the number of clusters is shown in panel a for both with the parametric models and noparametric vMF mixture models. In panel b is a comparison between the repetitions of the ivMF and iGMMs models.

for the $K = 100$ with a compelling delineation of both the putamen and the caudate. The insula is divided into four clusters: one for the posterior part, two for the central transitional region, and one for the anterior (Menon & Uddin, 2010). The putamen is subdivided into two regions and the caudate into three. The subdivision of the caudate and the putamen appears to be in general agreement with that reported by Janssen et al. (2015), where a rostral and caudal part of the putamen can be consistently identified, as well as a dorsal and rostral part of the caudate. Possibly due to the coarser-grained clustering at $K = 100$, a subdivision of the dorsal putamen was not seen, and the division of the ventral caudate appears slightly different from that of Janssen et al. (2015). For the $K = 250$, as expected, we see further subdivision of both the striatum and the insula. Some of the clusters are now separated across hemispheres, but the symmetry is still clear, and the division of the striatum is still in general agreement with Janssen et al. (2015).

We then applied the vMF model on the two data sets with number of clusters varying between 200 and 3000, again with the KMrand initialization strategy. After 100 MCMC iterations, we stopped the sampler and computed the predictive likelihood on the left-out group of subjects based on the hyperparameters and clustering configuration from the highest likelihood sample. In Figure 12, the results of this predictive analysis are given and show that the nonparametric models require on the order of a few thousand parcels to explain the data. These results are consistent with the analysis in Thirion et al. (2014), where Ward clustering of task-activated b-maps evaluated based on goodness of fit showed support for up to 5000 clusters.

Finally, we employ the nonparametric models, again using the KMrand initialization strategy assigning the voxels to 1000 clusters at random after the hyperparameters have been learned for 100 MCMC iterations. Each

run of the vMF-based model is presented as a circle in Figure 12a whereas box plots of the number of clusters inferred as well as NMI across the two groups of subjects are presented in Figure 12b. These results do not solve the problem of determining the number of functional units in the brain but suggest that whole brain fMRI segmentation requires on the order of a few thousand clusters to adequately account for the functional organization of the fMRI data and that the nonparametric models by identifying a large number of clusters are not overfitting to the data.

4 Conclusion

In this letter, we presented a thorough comparison of the effect of modeling directional data using vMF-based distributions in comparison to assuming the data are gaussian distributed considering both synthetic data and large-scale clustering of resting-state whole brain fMRI time series. We demonstrated a significant improvement in terms of the stability of solutions across groups of subjects when correctly imposing that the data reside on a hypersphere over the standard assumption of gaussian distributed observations. We have further shown that it is computationally feasible to apply sampling-based inference on multisubject whole brain fMRI time series data.

The predictive analysis shows that employing Bayesian nonparametrics can be a cheap substitute for using the computationally expensive, predictive cross-validation in determining the complexity of the data. Both the predictive cross-validation analysis and the Bayesian nonparametric analysis show that the resting-state fMRI data set supports a number of clusters on the order of a few thousand, which is in correspondence with recent findings (Thirion et al., 2014).

A variational inference–based implementation of the vMF mixture model has been proposed in two recent contributions (Taghia et al., 2014; Gopal & Yang, 2014). It could be interesting to compare the MCMC- and VI-based model implementations on both synthetic data and real problems based on ability to model the problem and computational complexity. We suspect there is a trade-off between VI being vulnerable to local minima with the MCMC implementation being more computationally expensive.

Modeling directional data using the appropriate directional distributions shows great promise, and this is an area worth more attention. A natural extension of this work would be to employ more advanced distributions on the hypersphere that has a more complex covariance structure, such as the Kent or Fisher Bingham distribution. We considered mixture modeling applications; however, we anticipate that the use of the vMF distribution may be useful in general when modeling standardized fMRI time series. The developed vMF clustering algorithm has been implemented in Matlab and is available from the authors at https://brainconnectivity.compute.dtu.dk.

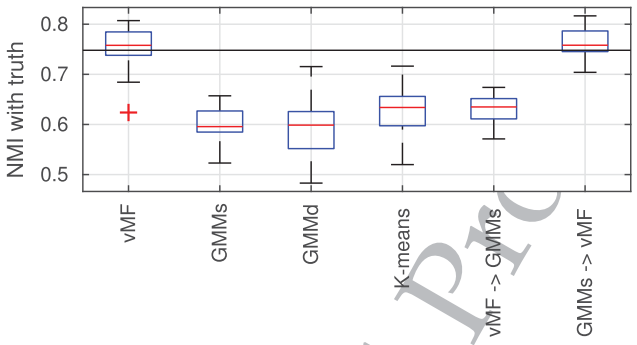


Figure 13: NMI with truth of the methods implemented. The solid black line is the result of averaging 10 repetitions of the VI implementation of the mixture of vMF model reported by Gopal et al. (2014).

Appendix: Document Topic Modeling

Document topic modeling is an application where variational inference–based vMF models have shown great promise (Gopal & Yang, 2014), and to confirm that our implementation of sampling-based inference is at least on par with the VI vMF, we apply our clustering method to the CNAE-9 data set.

The CNAE-9 data set consists of 1080 documents, and each document is a vector of the frequency of occurrence for 857 words: $N = 1080$ and $D = 857$. The documents are divided into nine categories, and the true clustering is thus available. Before clustering, we perform term frequency–inverse document frequency (tf-idf) on the data set, a standard preprocessing step for topic modeling and known to increase performance (Salton & McGill, 1983). First, we use the parametric models with the number of clusters set to $K = 10$ and apply the initialization method KMrand such that we use the K-means solution only to compute reasonable hyperparameters and then continue with a random initialization of the clustering. We repeat this process 60 times, and in each repetition, all models are initialized to the same K-means solution for the initial parameter estimation and the same random initialization afterward. We perform 500 MCMC iterations for each model and repetition and select the highest likelihood sample for comparison.

In order to confirm that the difference between the vMF- and GMM-based models is not a question of mixing, we continue a spherical GMM model from each of the vMF clustering solutions and perform another 500 MCMC iterations. Similarly, for each of the spherical GMM solutions, we continue in a vMF model for 500 MCMC iterations. The results are presented and compared to Gopal and Yang (2014) in Figure 13.

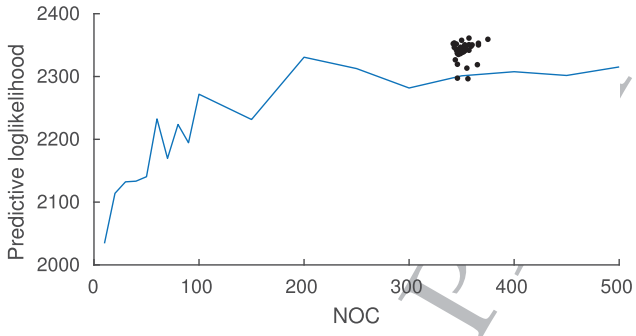


Figure 14: The predictive likelihood on the 10% hold-out data. The line is the result of running the vMF models with a fixed number of clusters averaged over 36 repetitions, and the black dots are the result of 36 repetitions of the nonparametric vMF models.

We use the same initialization procedure for the nonparametric vMF model and observe that it converges to around 300 clusters. In order to validate that the data have support for that number of clusters, we ran finite models with the number of clusters varying from 10 to 300 on a training set that consists of 90% of the data and computed the predictive likelihood on the hold-out set. These results are in Figure 14. We see that the nonparametric implementations of the vMF model can use the more advanced inference steps in split-merge to increase the predictive performance and that the inferred number of clusters is in a regime also supported by the predictive likelihood on hold-out test data.

Acknowledgments

This work was supported by the Lundbeckfonden (fellowship grant R105-9813 to M.M.). K.H.M. was supported by Lundbeckfonden (Grant of Excellence "ContAct" R59-5399 to Hartwig Roman Siebner) and by a Novo Nordisk Foundation Interdisciplinary Synergy Grant (NNF14OC0011413). The Magnetom Trio MR scanner was donated by the Simon Spies Foundation.

References

Aldous, D. J. (1985). *Exchangeability and related topics*. New York: Springer.

Andersen, K. W., Madsen, K. H., Siebner, H. R., Schmidt, M. N., Mørup, M., & Hansen, L. K. (2014). Non-parametric Bayesian graph models reveal community structure in resting state FMRI. *NeuroImage*, *100*, 301–315.

Anderson, C. A., Teal, P. D., & Poletti, M. A. (2015). Spatially robust far-field beamforming using the von Mises (-Fisher) distribution. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, *23*(12), 2189–2197.

Banerjee, A., Dhillon, I. S., Ghosh, J., & Sra, S. (2005). Clustering on the unit hypersphere using von Mises–Fisher distributions. *Journal of Machine Learning Research*, *6*, 1345–1382.

Bangert, M., Hennig, P., & Oelfke, U. (2010). Using an infinite von Mises–Fisher mixture model to cluster treatment beam directions in external radiation therapy. In *Proceedings of the Ninth International Conference on Machine Learning and Applications* (pp. 746–751). Piscataway, NJ: IEEE.

Biswal, B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., . . . Milham, M. D. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, *107*(10), 4734–4739.

Biswal, B., Yetkin, F. Z., Haughton, V. M., & Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn. Reson. Med.*, *34*, 537–541.

Churchill, N. W., Madsen, K., & Mørup, M. (2016). The functional segregation and integration model: Mixture model representations of consistent and variable group-Level connectivity in fMRI. *Neural Computation*, *28*, 1–41.

Craddock, R. C., James, G. A., Holtzheimer, P. E., Hu, X. P., & Mayberg, H. S. (2012). A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human Brain Mapping*, *33*, 1914–1928.

Fisher, R. (1953). Dispersion on a sphere. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *217*, 295–305.

Gopal, S., & Yang, Y. (2014). Von Mises–Fisher clustering models. In *Proceedings of the 31st International Conference on Machine Learning* (pp. 154–162). Red Hook, NY: Curran.

Goutte, C., Toft, P., Rostrup, E., Nielsen, F. Å., & Hansen, L. K. (1999). On clustering fmri time series. *NeuroImage*, *9*(3), 298–310.

Guan, H., & Smith, W. A. (2017). Structure-from-motion in spherical video using the von Mises–Fisher distribution. *IEEE Transactions on Image Processing*, *26*, 711–723.

Guttorp, P., & Lockhart, R. A. (1988). Finding the location of a signal: A Bayesian analysis. *Journal of the American Statistical Association*, *83*(402), 322–330.

Hornik, K., & Grün, B. (2013). On conjugate families and Jeffreys priors for von Mises–Fisher distributions. *Journal of Statistical Planning and Inference*, *145*(5), 992–999.

Hornik, K., & Grün, B. (2014). movmf: An r package for fitting mixtures of von Mises–Fisher distributions. *Journal of Statistical Software*, *58*(10), 1–31.

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

Hyde, J. S., & Jesmanowicz, A. (2012). Cross-correlation: an fMRI signal-processing strategy. *NeuroImage*, *62*(2), 848–851.

Jain, S., & Neal, R. M. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, *13*(1), 158–182.

Janssen, R. J., Jylänki, P., Kessels, R. P. C., & van Gerven, M. A. J. (2015). Probabilistic model-based functional parcellation reveals a robust, fine-grained subdivision of the striatum. *NeuroImage*, *119*, 398–405.

Lashkari, D., & Golland, P. (2009). Exploratory FMRI analysis without spatial normalization. In *Proceedings of the International Conference on Information Processing in Medical Imaging* (pp. 398–410). New York: Springer.

Lashkari, D., Vul, E., Kanwisher, N., & Golland, P. (2010). Discovering structure in the space of FMRI selectivity profiles. *NeuroImage*, *50*(3), 1085–1098.

Ma, Z., Taghia, J., Kleijn, W. B., Leijon, A., & Guo, J. (2015). Line spectral frequencies modeling by a mixture of von Mises–Fisher distributions. *Signal Processing*, *114*, 219–224.

Mardia, K. V., & El-Atoum, S. A. M. (1976). Bayesian inference for the von Mises–Fisher distribution. *Biometrika*, *63*(1), 203–206.

Mardia, K., & Jupp, P. E. (2009). *Directional statistics*. Hoboken, NJ: Wiley.

Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: A network model of insula function. *Brain Structure and Function*, *214*(5–6), 655–667.

Nunez-Antonio, G., & Gutiérrez-Pena, E. (2005). A Bayesian analysis of directional data using the von Mises–Fisher distribution. *Communications in Statistics Simulation and Computation*, *34*(4), 989–999.

Pitman, J. (2006). Combinatorial stochastic processes. *Lecture Notes in Mathematics*, *1875*, 1–247.

Rasmussen, C. E. (1999). The infinite gaussian mixture model. In S. A. Solle, T. K. Leen, & K. Müller (Eds.), *Advances in neural information proceesing systems, 12* (pp. 554–560). Cambridge, MA: MIT Press.

Røge, R., Madsen, K. H., Schmidt, M. N., & Mørup, M. (2015). Unsupervised segmentation of task activated regions in FMRI. In *Proceedings of the 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing* (pp. 1–6). Piscataway, NJ: IEEE.

Ryali, S., Chen, T., Supekar, K., & Menon, V. (2013). A parcellation scheme based on von Mises–Fisher distributions and Markov random fields for segmenting brain regions using resting-state fMRI. *NeuroImage*, *65*, 83–96.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Strehl, A., & Ghosh, J. (2002). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, *3*, 583–617.

Strother, S. C., Anderson, J., Hansen, L. K., Kjems, U., Kustra, R., Sidtis, J., . . . Rottenberg, D. (2002). The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework. *NeuroImage*, *15*(4), 747–71.

Taghia, J., Ma, Z., & Leijon, A. (2014). Bayesian estimation of the von–Mises Fisher mixture model with variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(9), 1701–1715.

Thirion, B., Varoquaux, G., Dohmatob, E., & Poline, J. (2014). Which FMRI clustering gives good brain parcellations? *Frontiers in Neuroscience*, *8*, 167.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., . . . Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*(1), 273–289.

Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, *11*, 2837–2854.

Vul, E., Lashkari, D., Hsieh, P., Golland, P., & Kanwisher, N. (2012). Data-driven functional clustering reveals dominance of face, place, and body selectivity in the ventral visual pathway. *Journal of Neurophysiology*, *108*, 2306–2322.

Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., . . . Buckner, R. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, *106*(3), 1125–1165.

Received November 16, 2016; accepted May 1, 2017.