Bayesian Modelling of Functional Whole Brain Connectivity

Røge, Rasmus

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):
Røge, R. (2017). *Bayesian Modelling of Functional Whole Brain Connectivity*. Technical University of Denmark. DTU Compute PHD-2017 No. 445

Bayesian Modelling of Functional Whole Brain Connectivity

Rasmus Erbou Røge

Kongens Lyngby 2017
IMM-PhD-2017-DTU-IMM

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Richard Petersens Plads, building 324,
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
compute@compute.dtu.dk
www.compute.dtu.dk
IMM-PhD-2017-DTU-IMM

Summary (English)

This thesis deals with parcellation of whole-brain functional magnetic resonance imaging (fMRI) using Bayesian inference with mixture models tailored to the fMRI data. In the three included papers and manuscripts, we analyze two different approaches to modeling fMRI signal; either we accept the prevalent strategy of standardizing of fMRI time series and model data using directional statistics or we model the variability in the signal across the brain and across multiple subjects. In either case, we use Bayesian nonparametric modeling to automatically learn from the fMRI data the number of funcional units, i.e. parcels. We benchmark the proposed mixture models against state of the art methods of brain parcellation, both probabilistic and non-probabilistic.

The time series of each voxel are most often standardized using *z-scoring* which projects the time series data onto a hypersphere. This underlying manifold is often ignored and the data is modeled using Gaussian distributions. In one contribution, we show that using a mixture model based on the directional distribution, the von Mises-Fisher distribution, increase the reliability of inferred parcellations.

We develop a mixture model for modeling time-series using a Gaussian Process as a prior that is informed of the temporal dynamics of the data expected from the blood oxygenation level dependent (BOLD) signal. In two contributions, we explore the potential of this modeling framework. In the first, we show that this mixture model can delineate regions of task activation that can then be identified unsupervised. This forms a promising framework for unsupervised identification of task activated when the task design is unknown. In the final contribution, we evaluate the performance of the mixture model on the problem of clustering whole-brain fMRI. Based on both simulations on synthetic data and analysis of two fMRI datasets, we show that the model provides improved reliability of clustering compared to traditional clustering methods. Furthermore, the inferred parcellations provide the foundation for a method for increasing the reliability and sensitivity in analyses of task activation and for determining the networks of functionally connectivity in fMRI.

The proposed mixture models form promising tools for brain parcellation and we hope the methods can provide a nudge towards using probabilistic models for fMRI parcellation.

Resumé (Danish)

Denne afhandling omhandler problemstillingen med at inddele hjernen i funktionelle enheder, det vil sige områder der har samme funktion som målt ved funktionelle magnetisk resonans scanningsbilleder. Til dette formål anvender vi miksturmodeller der er skræddersyede til fMRI signalet. I tre bidrag analyserer vi to forskellige tilgangsvinkler til at modellere fMRI signalet; enten kan man acceptere den gængse metode, hvor man standardiserer fMRI tidsserierne og modellere signalet ved hjælp af sandsynlighedsfordelinger over retninger eller man kan forsøge at modellere variabiliteten der findes i støj og signal hen over hjernen og mellem forskellige personer. I hvert tilfælde har vi brug for at sammenligne de modeller vi foreslår med både probabilistiske og ikke-probabilistiske metoder til at parcellere hjernen.

Når voxeltiddserier standardiseres har de ikke længere information om magnituden af observationerne men er projiceret ned på overfladen af en hyperkugle. Alligevel anvendes modeller baseret på Gaussiske miksturer ofte til at modellere fMRI signalet og i et bidrag viser vi at en miksturmodel der baserer sig på von Mises-Fisher fordelingen, der er en fordeling overfladen af hyperkugler, øger pålideligheden af de identificerede parcelleringer.

Derudover udvikler vi en miksturmodel til modellering af tidsseriedata der bruger en Gaussisk Proces som en informeret prior der er informeret om den temporale dynamik der forventes at være grundet ændringer i hjernens blodgennemstrømning. I to bidrag undersøger vi potentialet for denne konstruktion af miksturmodeller. Først viser vi at modellen kan afgrænse områder af hjernen der er aktiveret mens personer udfører en opgave i MR scanneren og disse områder kan identificeres ved at se på korrelation af parcel tidsserier over forskellige individder. Denne metode er lovende til at finde områder af hjernen der aktiveres når individer udsættes for stimuli der ikke udføres i overensstemmelse med et veldefineret skema. I det andet bidrag udfører vi en grundig analyse af miksturmodellens evne til at parcellere fMRI tidsserier fra alle voxels i hjernens grå substans. Baseret på både en syntetisk analyse og på analyser af to fMRI datasæt ser vi at den foreslåede metode finder parcelleringer der er mere pålidelige i forhold til eksisterende metoder, både probabilistiske og ikke-probabilistiske. Derudover viser vi, at de lærte parcelleringer medfører øget følsomhed og pålidelighed i analyser af aktiverede hjerneområder samt i netværksanalyser.

De foreslåede miksturmodeller viser et stort potentiale og vi håber at vores bidrag kan give et skub i retningen mod at bruge probabilistiske metoder til at adressere problemet med opdeling af hjernen i funktionelle enheder.

Preface

This thesis was prepared at the Section for Cognitive Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark in partial fulfillment of the requirements for acquiring the Ph.D. degree in engineering. The project was funded partly ($\frac{2}{3}$) by the Lundbeck foundation through the *Non-parametric Relational Modeling of Functional and Structural Brain Connectivity* project (see also `https://brainconnectivity.compute.dtu.dk/`) and partly by DTU Compute. My main supervisor was Associate Professor Morten Mørup, DTU Compute, Technical University of Denmark. I had two co-supervisors; Associate Professor Mikkel N. Schmidt from DTU Compute and Associate Professor Kristoffer H. Madsen from Danish Research Centre for Magnetic Resonance, Copenhagen University Hospital.

The thesis consists of a summary report, a paper published in a conference proceeding, one submitted paper, and a paper included as manuscript draft. The work has been carried out between February 2014 and January 2017.

Lyngby, 31-January-2017

Rasmus Erbou Røge

List of Publications

Papers included in the thesis

Conference proceedings

A Røge, R. E., Madsen, K. H., Schmidt, M. N., Mørup, M. (2015), 'Unsupervised Segmentation of Task Activated Regions in fMRI'. 2015 IEEE International Workshop on Machine Learning for Signal Processing.

Submitted

B Røge, R. E., Madsen, K. H., Schmidt, M. N., Mørup, M. (2015), 'Non-parametric von-Mishes Fisher based clustering of whole-brain resting state fMRI'. *Submitted 22 Nov. 2016, invited for resubmission within three months 06 Jan. 2017.*

Manuscript

C Røge, R. E., Schmidt, M. N., Churchill, N. W., Madsen, K. H., Mørup, M. (2015), 'Functional Whole-Brain Parcellation using Bayesian Non-Parametric Modeling' *Manuscript to be submitted*

Papers not included in the thesis

- Nielsen, S. F. V., Madsen, K. H., Røge, R., Schmidt, M. N., Mørup, M. (2015). 'Nonparametric modeling of dynamic functional connectivity in fmri data'. In *Proceedings of the 5th Nips Workshop on Machine Learning and Interpretation in Neuroimaging* (mlini 2015).
- Hinrich, J. L., Bardenfleth, S. E., Røge, R. E., Churchill, N. W., Madsen, K. H., Mørup, M. (2016). 'Archetypal Analysis for Modeling Multisubject fMRI Data'. *IEEE Journal of Selected Topics in Signal Processing*, 10(7), 1160-1171.

Acknowledgements

There are several people I need to thank for without those, this thesis would not have been possible. First of all I want to thank my supervisors, Morten Mørup, Mikkel N. Schmidt, and Kristoffer H. Madsen for letting me be a part of the Brainconnectivity project and for all the guidance, support and discussions that has formed the foundations for this thesis.

I also want to thank the Lundbeck foundation for funding the Brainconnectivity project therein my Ph.D. and my thank also goes to H.I. Hannover, P.A. Fiskers, Otto Mønsted and the Augustinus foundation for travel grants. These travel grants made my stay with the Stanford Cognitive and Systems Neuroscience Lab possible. A thank also goes to all the friends I made while visiting the lab.

I would like to thank everyone in the section for Cognitive systems and especially my fellow Ph.D. students for a pleasant atmosphere and for making the frequent coffee breaks so enjoyable.

Finally, and most importantly, I want to express my gratitude to my friends and family for continued support and encouragement when the struggle was hardest.

Contents

Chapter 1

Introduction

Understanding the human brain is one of the most fundamental questions science has tried to answer as it is crucial to understanding ourselves and possibly to understanding the concept of intelligence in general. Historically, this quest has been driven by few examples where damage has been inflicted to the brain, either by accident or by experimenting with inflicting lesions to the human or animal brain. In the past few decades, the methods of studying the brain, has shifted towards noninvasive studies using modern techniques for brain imaging including electroencephalogram (EEG), magnetoencephalogram (MEG), positron emission tomography (PET), and methods of magnetic resonance imaging (MRI) including structural MRI, diffusion based MRI (dMRI) and the focus of this thesis, functional MRI (fMRI).

Increased neuronal activity indirectly causes an increase in the local level of oxygenated blood and, due to differences between the magnetic properties of oxygenated and deoxygenated hemoglobin, this gives rise to the blood oxygenation level dependent (BOLD) signal. With fMRI, a structural image of the brain is recorded every few seconds and, focusing on the BOLD signal, it provides an indirect measure of the level of neural activity across the brain. The spatial resolution of fMRI scans, i.e. the size of the volumetric pixels, called voxels, is typically between 1 and 4^3 cubic mm for modern MR scanners and scanning sequences. In this resolution, every voxel contains millions of neurons and the BOLD signal is therefore an aggregated measure of the neuronal activity within each voxel. Still, the signal-to-noise level is so high that the single voxel level might not be sufficiently aggregated for the sensitivity of current generation fMRI (Chumbley and Friston, 2009) but gathering voxels in regions of homogenous functional activity can further improve the sensitivity and reliability and better characterize the function and connectivity in the brain.

A prominent view describes the human brain as organized into segregated functional units that function together in a network (Tononi et al., 1994). Identifying these functional units is very much an open problem, and historically this has been done using labor intensive parcellation techniques based on structural properties such as myelin thickness and the tissue types from in vitro brains or structural scans. Brain atlases such as the automated anatomical labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002) and the Broadmann regions are examples of such atlases. A recent exploration of this approach is presented in Glasser et al. (2016b) where multi modal data from a very large number of subjects are used to find a brain parcellation in a mix of labor intensive and data driven parcellation methods. Alternative to these approaches are the purely data driven methods for parcellating the brain that use tools from machine learning to perform clustering on fMRI data.

Prominent classical methods of clustering applied to brain parcellation include K-means clustering, hierarchical agglomerative clustering, divisive clustering, techniques based on growing seed regions, and techniques focusing on boundaries between areas of consistent functional connectivity. These methods share a number of disadvantages in that they require the pre-specification of a number of clusters and they are not based on probabilistic models and thus unable to explicitly model the variability in physiological and scanner related noise that is apparent in fMRI.

Recent approaches have applied probabilistic mixture models to the problem of parcellating substructures of the brain (Ryali et al., 2013; Janssen et al., 2015). Inspired by these results and others, we aim with this thesis to construct generative models that scale to the problem of whole-brain fMRI time-series data and that explicitly incorporates domain knowledge. Furthermore, the models we employ are able to adapt to and quantify the complexity of the data using Bayesian non-parametrics. The produced software comprises a generic framework for efficient sampling in Bayesian mixture models such that the models can be further developed and adapted. The developed clustering models are compared to the state of the art non-probabilistic clustering models based on established methods of comparing whole-brain parcellation techniques as well as against the end goal of fMRI analysis such as network analysis and identifying regions of task activation.

This thesis includes three research contributions; in paper A, we explore the use of a parcellation model as a means to unsupervised extraction of information on the task relevant regions in fMRI. In papers B and C we analyze two approaches to modeling fMRI time series data. A frequently used preprocessing step is to standardize the voxel time series using z-scoring, which means that each voxel will have zero mean and unit standard deviation. Considering the standardized time-series as a point in a vector space, only the direction in space remains and this is properly modeled using directional distributions. We explore this approach to modeling fMRI data using a nonparametric mixture model based on von Mises-Fisher distributions in paper B. The other approach we consider is to model the variability in the data explicitly using a model that account for the spatial heteroscedasticity of the noise and signal amplitude that is present in fMRI data. Furthermore, the proposed model uses Gaussian Processes as an informed prior to focus on the part of the fMRI signal that is caused by changes in the cerebral blood flow. We analyze the model built on these assumptions for parcellating fMRI data in paper C.

The rest of the thesis is organized as follows.

Chapter 2 presents a brief description of fMRI, the traditional approaches to fMRI parcellation, and methods of validating brain parcellation methods.

Chapter 3 describes the theoretical basis for the Bayesian mixture models as well as the inference methods applied in this thesis.

Chapter 4 presents the software tool for sampling based inference in Bayesian mixture models that was developed as part of the thesis.

Chapter 5 summarizes the three research contributions.

Chapter 6 presents a discussion and conclusion on the thesis as a whole.

The three papers included in this thesis can be found in appendices A, B, and C.

Chapter 2

Functional Magnetic Resonance Imaging

In the following sections, I will first introduce functional magnetic resonance imaging (fMRI) and the preprocessing pipeline that was applied to the datasets used for this thesis. In section §2.2 I will introduce whole-brain parcellation along with some of the most commonly applied parcellation methods and ways of evaluating and comparing clusterings.

2.1 fMRI

The history behind the modern MR scanner dates to the initial discovery of the principle of nuclear magnetic resonance by Isidor Rabi around 1938 for which he was awarded the 1944 Nobel Prize in Physics. Felix Block and Edward Purell independently observed the NMR phenomenon in liquids and solids in 1946 for which they shared the 1952 Nobel Prize in Physics. The 2003 Nobel Prize in Medicine was, under some controversy, awarded to Paul Lauterbur and Peter Mansfield who independently described the use of gradients to localize NMR signals and thus setting the foundations for the way MRI is being performed today (Ai et al., 2012). The first demonstrations that the MR scanner could be used to measure the *in vivo* blood oxygenation level dependent (BOLD) response, and thus changes in hemodynamics, was performed in 1990 for the rodent brain (Ogawa et al., 1990) and a few years later, the human brain (Ogawa et al., 1992; Kwong et al., 1992; Bandettini et al., 1992).

In magnetic resonance imaging, the subject that is to be scanned is placed in a strong magnetic field and nuclei with a magnetic dipole moment, such as the hydrogen nuclei, are excited by oscillating a weak magnetic field at the resonance frequency. This causes the magnetic field of the hydrogen nuclei to oscillate while emitting a weak oscillating magnetic field that is measured (Faro and Mohamed, 2006). The magnetic properties of matter vary depending on perturbations in the magnetic field in the vicinity of the hydrogen atom, and in particular there is a slight difference between the magnetic properties of oxygenated and deoxygenated hemoglobin. This gives rise to the blood oxygenation level dependent (BOLD) contrast which is indirectly coupled to neural activity (Faro and Mohamed, 2006). In fMRI, the entire brain is scanned every few seconds resulting in a 3-dimensional image, often described as a brain volume, and from the BOLD signal it is possible to achieve the indirect measure of brain activity of each volumetric pixel, denoted voxels, of the brain.

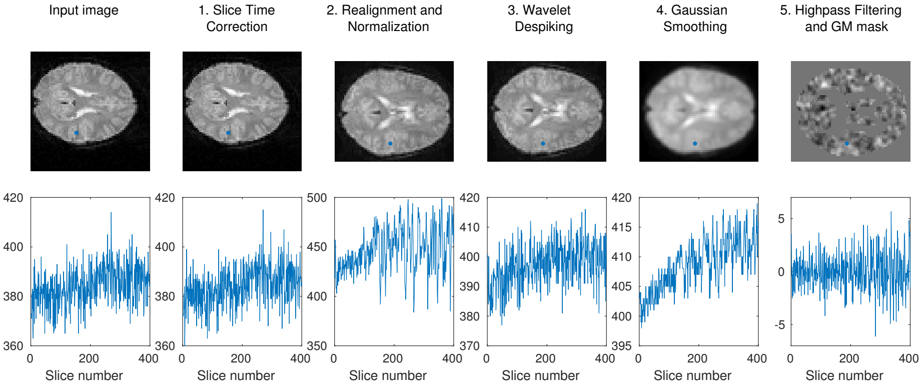


Figure 2.1: Illustration of the preprocessing pipeline applied for the data in this thesis. Above we show an axial slice of the brain where a voxel is marked with a blue dot and below are the the time series for the first 400 brain volumes for the marked voxel. Note, that the voxel time series before and after realignment and normalization is not necessarily from the same voxel.

It is common practice to perform a number of preprocessing steps to the recorded fMRI signal (Faro and Mohamed, 2006). In this thesis, we used two datasets recorded at the Danish Research Centre for Magnetic Resonance, consisting of 29 healthy subjects scanned in a resting state session and a finger tapping task session. In the resting state dataset, the subjects were asked to rest and not think of anything while 480 brain volumes were recorded. For the finger tapping task dataset, subjects were instructed to follow a task paradigm consisting of 10 repetitions of the stimulation cycle: 20 s right handed finger tapping, 10 s rest, 20 s left handed finger tapping, and 10 s rest resulting in 240 brain volumes in total. Both the fMRI datasets used a TR=2.49 s and recorded 3 mm isotropic voxels. For further description of the scanning parameters, see Rasmussen et al. (2012); Andersen et al. (2014) or the included papers A-C. The preprocessing pipeline used for the datasets was performed using the SPM12 software (SPM12, Wellcome Trust Centre for Neuroimaging, `http://www.fil.ion.ucl.ac.uk/spm/software/spm12/`) and the `BrainWavelet` software tool (Patel et al., 2014). The pipeline is illustrated in Figure 2.1 and consists of the following steps:

1. **Slice Time Correction** The brain volumes are typically recorded one slice at a time and each slice will thus have a slightly different acquisition time. Under the assumption of critical sampling, this can to some degree be corrected for by interpolating the time series.

2. **Realignment and Normalization** The brain volumes are highly susceptible to even minor movement in the scanner and the volumes must therefore be realigned using a six-parameter rigid body transformation. Furthermore, the volumes of individual subjects are often normalized to a space shared over subjects, such as the MNI152 brain (Montreal Neurological Institute, average of 152 coregistered structural images). All fMRI volumes were registered to the mean volume and normalized to the MNI152 brain.

3. **Wavelet despiking** Spikes (sudden high amplitude signals) are typically present in the fMRI signal, often caused by head movement. It is to some degree possible to correct for this using some form of despiking and we used the software package `BrainWavelet` for this purpose (Patel et al., 2014).

4. **Gaussian Smoothing** For most of the datasets we applied spatial smoothing using a 4mm FWHM Gaussian kernel. While this step severely reduces the spatial resolution it increases the signal to noise ratio by reducing the non-spatially distributed noise and may help to reduce the anatomical differences between subjects.

5. **High pass filtering and GM mask** To remove low frequency fluctuations unrelated to the signal we applied a high-pass filter with a cut-off period at 128 s. We then subtracted the mean time series from each voxel time series, and used the SPM tissue probability map with a threshold set to 0.25 pct. for a rough grey matter mask.

The preprocessing pipeline is quite minimal and often further steps are taken to correct for motion (Friston et al., 1996) and physiological noise such as respiration and pulse (Lund et al., 2006). Furthermore, the fMRI data might be projected onto the cortical surface to get a 2-dimensional representation of the data (Glasser et al., 2016b).

Traditional analysis of task-fMRI, where the subject is asked to perform a task per instructions, is performed using a mass-univariate general linear model (GLM). GLM performs a univariate test for each voxel against a design matrix creating a statistical parametric map of the brain, often correcting for autocorrelation. The design matrix consists of regressors for the task paradigm along regressors correcting for motion and possibly other nuisance regressors. The regressors in the design matrix are compared by setting up contrasts, i.e. testing whether one task is more explanatory compared to rest or another task. These contrast maps can be used as input for studies in brain parcellation, see for instance (Goutte et al., 1999; Thirion et al., 2014). To test if a region of activation is statistically significant, the tests against the task paradigm is corrected for multiple comparisons using either Bonferroni correction or controlling for family wise errors (FWE) using Gaussian random fields (Worsley et al., 1996). Alternatives to correction using FWE include permutation test (Nichols and Holmes, 2002), controlling the false discovery rate (Chumbley and Friston, 2009), and cluster based inference as we apply in paper C.

2.2 Functional brain parcellation

Functional brain parcellation divides the brain into regions that are functionally homogenous with respect to the voxel time series or to the connectivity map of each voxel. The benefits of a good parcellation are many, Glasser et al. (2016a) phrases it in the following form:

> Accurate parcellation provides a map of where we are in the brain, enabling efficient comparison of results across studies and communication among investigators; as a foundation for illuminating the functional and structural organization of the brain; and as a means to reduce data complexity while improving statistical sensitivity and power for many neuroimaging studies.
>
> Glasser et al. (2016a)

While Glasser et al. (2016a) argues for an atlas of the brain based on both machine learning methods and manual labor on a multi modal magnetic resonance imaging dataset consisting of several hundred subjects, we focus in this thesis on the purely data driven parcellation methods. An atlas based on a purely data driven parcellation would optimally be able to characterize the salient features of the dataset.

There have been numerous studies searching for an optimal method to compute a data driven parcellation of the human brain. These studies typically face a number of questions such as: How can spatial constraints be incorporated in the model? How should data be aggregated from multiple subjects for an optimal group level clustering? How can different clustering techniques be evaluated and compared? How should the number of parcels be selected? In the following sections, we present an overview of some of the methods that are applied to the problem of functional brain parcellation and how they answer the three questions phrased above.

2.2.1 Methods of functional brain parcellation

Brain parcellations are typically computed from one of three types of fMRI derived data:

(a) fMRI time series.

(b) Graphs of functional connectivity (FC).

(c) Statistical parametric maps or Z-maps against task designs.

For the following let $\boldsymbol{x}_i$ for $i = 1, \ldots, N$ denote either the (a) fMRI time series of the i'th voxel, (b) the N dimensional vector of correlations with all voxels in the brain (c) The P dimensional vector consisting of the values in the parametric maps for each of the P contrasts. Let the voxels be partitioned into K clusters according to the cluster assignment vector $\boldsymbol{z}$ such that $z_i = k$ if the i'th voxel is in the k'th cluster. Furthermore, let $\mathcal{Z}_k$ be the indices of voxel in cluster k, i.e. $\mathcal{Z}_k = \{i \mid z_i = k\}$ and $\boldsymbol{\mu}_k$ the k'th centroid. Using this notation, we can describe the following list of clustering methods that recently have gained traction:

K-Means is perhaps the most commonly used clustering algorithm and has been used on Z-maps by Goutte et al. (1999); Thirion et al. (2014), and functional connectivity by Yeo et al. (2011). K-means is a greedy algorithm for minimizing the following cost function

$$\frac{1}{K}\sum_{k=1}^{K}\sum_{i\in\mathcal{Z}_k}||\boldsymbol{x}_i-\boldsymbol{\mu}_k||^2. \tag{2.1}$$

The centroids $\boldsymbol{\mu}_k$ are initialized either randomly or using the ++ algorithm (Arthur and Vassilvitskii, 2007) and refined using greedy updates that iterates between assigning clusters to the closest centroid and updating the centroids of each cluster to the empirical centers.

Ward Ward's hierarchical agglomerative clustering algorithm (Ward Jr, 1963) has been used to cluster both Z-maps (Thirion et al., 2014) and functional connectivity (Baldassano et al., 2015). The hierarchical clustering method initializes with voxels in singleton clusters and at each step joins the two clusters with the shortest Euclidean distance between the two centroids $||\boldsymbol{\mu}_l-\boldsymbol{\mu}_k||^2$ until the desired number of clusters are left.

Normalized Cut has been frequently used in neuroimaging and perhaps most prominently by Craddock et al. (2012). Ncut (Shi and Malik, 2000) represents the fMRI data as a fully connected graph with voxels as nodes and where the edges between two voxels is weighted by their similarity, s, for any measure of similarity such as the Pearson correlation between time series or functional connectivity profiles. At any step in the divisive clustering method, Ncut divides a cluster $\mathcal{Z}_k$ into two cluster $\mathcal{Z}_{k_1}$ and $\mathcal{Z}_{k_2}$ such that it minimizes the Ncut cost function:

$$\text{Ncut}(\mathcal{Z}_{k_1},\mathcal{Z}_{k_2})=\frac{\text{cut}(\mathcal{Z}_{k_1},\mathcal{Z}_{k_2})}{\text{assoc}(\mathcal{Z}_{k_1},\mathcal{Z}_k)}+\frac{\text{cut}(\mathcal{Z}_{k_1},\mathcal{Z}_{k_2})}{\text{assoc}(\mathcal{Z}_{k_2},\mathcal{Z}_k)}, \tag{2.2}$$

where

$$\text{cut}(\mathcal{Z}_{k_1},\mathcal{Z}_{k_2})=\sum_{a\in\mathcal{Z}_{k_1},b\in\mathcal{Z}_{k_2}}s(a,b) \tag{2.3}$$

is the sum of the total edge connections of the two clusters and

$$\text{assoc}(\mathcal{Z}_{k_1},\mathcal{Z}_k)=\sum_{a\in\mathcal{Z}_{k_1},v\in\mathcal{Z}_k}s(a,v) \tag{2.4}$$

is the sum of the total edge connections between the nodes in $\mathcal{Z}_{k_1}$ to all the nodes in $\mathcal{Z}_k$. Advantages include that it is robust to outliers due to the normalization of the cut criteria. This same normalization does, however, also means that it produces clusters of similar size and that the clusterings might be driven by size instead of the data as illustrated in paper C.

Region Growing has recently been proposed by Blumensath et al. (2013) to parcellate surface based FC and is a technique where a number of seed regions are selected in regions of stable functional connectivity. The seeds are then grown into non-overlapping initial regions. These regions are finally merged using a hierarchical clustering technique, such as Ward's, until the desired number of clusters is left.

Boundary mapping was proposed by Cohen et al. (2008) and later refined by Wig et al. (2014); Gordon et al. (2014) and applied to a large scale multi modal dataset by Glasser et al. (2016a). Boundary mapping is applied to surface based functional connectivity fMRI data: For each vertex, a map of the level local spatial similarity of the vertex by vertex functional connectivity map is created. Based on this map, areas with high gradients in the map of spatial similarity are identified as areas with high probability of being a boundary between two clusters. This map of probability edges is averaged over all vertices and, for a group analysis, over all subjects. Finally an edge detection algorithm is applied to calculate the clustering.

Network modeling A number of variants of the stochastic block models have been applied to modeling the functional connectivity network in resting state fMRI data. Mørup et al. (2010) applied the infinite relational model (IRM) on thresholded connectivity maps and Andersen et al. (2014) compared the IRM model to several constrained variants of the IRM model. The full connectivity map was modeled by Baldassano et al. (2015) using a probabilistic model that showed increased reliability compared to several non-probabilistic methods including Ncut, Ward, Region growing and boundary mapping.

Mixture modeling Churchill et al. (2016) introduces a mixture model that simultaneously optimizes the within cluster homogeneity and the connectivity to the rest of the brain. Lashkari and Golland (2009); Lashkari et al. (2010) applied a mixture model based on the von Mises-Fisher distribution to task activations. Recently a non-parametric Gaussian mixture model with spatial restrictions have been used to cluster the striatum (Janssen et al., 2015) and Ryali et al. (2013) used a vMF based mixture model to cluster several substructures of the brain including the insula and motor cortex. We elaborate on probabilistic mixture modeling in Chapter 3.

Spatial Constraints

There is a consensus, that the functional units of the brain should be spatially contiguous. Clustering methods based on the voxel by voxel similarity matrix can directly enforce spatial contiguity using a hard constraint such as fixing the similarity to zero outside some neighborhood of the voxels:

$$w_{i,j} = \begin{cases} s(v_i, v_j) & \text{if } d_{i,j} \leq \epsilon \\ 0 & \text{otherwise} \end{cases}, \tag{2.5}$$

where $\epsilon > 0$ and $s(v_i, v_j)$ and $d_{i,j}$ are the similarity and distance between voxel i and j respectively. This is frequently implemented for both Ward's and the N-cut algorithms. Spatial contiguity is, by construction, ensured for both region growing and boundary detection clustering algorithms.

In probabilistic parcellation methods this spatial contiguity can be ensured using the distance dependent Chinese Restaurant Process (dd-CRP) prior on the clustering Blei and Frazier (2011). This has been applied to clustering of fMRI time series by Janssen et al. (2015) and to FC networks by Baldassano et al. (2015). In section 3.2.1 we will introduce the Chinese Restaurant Process (CRP). The dd-CRP modifies the CRP to take the distance between clusters and nodes into account.

Ryali et al. (2013) used a Markov random field to incorporate spatial dependency in a probabilistic model for the time series to parcellate several substructures of the brain including the striatum and the insula. Note, that the use of a Markov random field can be computationally expensive and thus prohibitive for use in whole-brain parcellation.

In our contributions, we omit spatial constraints in the probabilistic models allowing the data to be optimally modeled. The means we are not guaranteed spatially contiguous clusters but we allow for other important aspects of the data such as left right symmetry between the two hemispheres of the brain. Note, that even though the models do not guarantee spatial contiguity, we would expect the models to infer clusters that are spatially localized if this structure is supported by the data. Note also that spatially smoothing the data in preprocessing also, however indirectly, encourages spatially contiguous clusters.

Group level clustering

To obtain a reliable clustering that is able to handle the level of noise in fMRI data, it is necessary to incorporate data from multiple scanning sessions. Recent studies have incorporated information from several hundred subjects, such as in Yeo et al. (2011) who uses 500 subjects and Glasser et al. (2016a) uses information from 210 subjects.

In many studies of functional connectivity, the connectivity matrix is averaged across subjects (Baldassano et al., 2015; Glasser et al., 2016a; Craddock et al., 2012). Other approaches calculate subject specific parcellations for each subject in the group and subsequently uses these parcellations for a second level group clustering (Craddock et al., 2012). In probabilistic clustering models, each subject can be modeled independently allowing for explicitly modeling the variability across subjects (Janssen et al., 2015) and in this thesis, we follow that approach.

2.2.2 Evaluating clusterings

While there is no consensus on the details of the methods that are used to evaluate methods for brain parcellation, it is quite well established that for a parcellation method to be *good*, the parcellation method must satisfy the following three criteria in some form:

Compliancy The inferred parcellations must be in compliance with brain structures known from studies of human brain anatomy.

Parsimony The description of the data should capture important aspects and otherwise use as few components as possible as stated by the principle of parsimony or Occam's razor.

Reliability If the inferred parcellation is to be interpreted as a presentation of the structure of the human brain in a way that generalizes among subjects it is necessary for the inferred parcellation to be robust.

Compliancy

It is important that a parcellation of the brain complies with the wealth of knowledge that has been accumulated on the structure of the brain. Areas of task activation should align with parcel borders (Wig et al., 2014; Gordon et al., 2014; Glasser et al., 2016a) and the parcellation of specific structures of the brain should be similar to those found by analyzing the architecture of the brain, for instance based on myelin thickness Ryali et al. (2013); Janssen et al. (2015); Glasser et al. (2016a). This is, of course, only true under the assumption that the brain volumes are properly normalized and that there is no individual differences in the function of brain regions. This is an assumption that is often made for studies of groups of healthy subjects but limits the use of compliancy as a method of evaluating parcellation methods.

It is, however, difficult to quantify the level of compliancy and in most studies the evaluation based on this criterion is primarily based on a visual comparison on the cortical surface. Note, however, that the criterion of compliancy is closely related with the functional homogeneity of the parcels, since parcels of good functional homogeneity should also align with areas of task activation.

Inspired by single subject test-retest reliability of task activation (Gorgolewski et al., 2013), in paper C we propose the use of the reliability of inferred regions of task activation as a measure that combines the evaluation criteria compliancy, parsimony, and reliability. For a parcellation based task analysis to be reliable it requires the parcellation to be reliably inferred, that the regions are functionally homogeneous, and that the parcels align with regions of task activation.

Parsimony

The principle of parsimony states that the explanation to any phenomenon should make as few assumptions as possible. For parcellation methods this means that the inferred parcels should be homogenous while balancing homogeneity and number of parcels. Most of the conventionally applied methods of brain parcellation requires the prespecification of the number of components that are required by the clustering model. The evaluation of how well the parcellation models can explain the data can be measured by evaluating the homogeneity of the inferred parcels.

There is a long list of methods that address the issues of parsimony and reliability. In the following we present some of the most prominent and a more extensive list of methods of evaluating clusterings can be found in Eickhoff et al. (2015). A frequently applied measure of functional homogeneity is the average similarity (Craddock et al., 2012) often using either the similarity between time-series or similarity between functional connectivity maps.

Gordon et al. (2014) propose the use of the percent of the parcel variance in functional connectivity that can be explained by the most prominent connectivity pattern in the parcel. This can be computed in the following way: First compute the N by N matrix with the vectors of functional and for each cluster select the vectors that corresponds to voxels resulting in a n_k by N matrix. Compute the singular value decomposition of this matrix, then the percentage of variance from the first eigenvalue will denote the homogeneity from this cluster. The measure averages the homogeneity from all clusters and Gordon et al. (2014) notes that it produces similar results to the average similarity of Craddock et al. (2012).

Regardless of which measure of homogeneity is chosen, it is difficult to compare the homogeneity of clusterings with a different number of clusters. For surface based clustering methods, this can be addressed using the relative homogeneity against a null model with randomly placed parcels of the same size, shape, and relative position to each other. These null model parcellations can be achieved by applying small rotations of the surface based parcellations (Gordon et al., 2014).

For probabilistic models the predictive likelihood can be used to compare models and to quantify the homogeneity. The predictive likelihood balances model complexity and homogeneity and should therefore be better at comparing models that differ in the number of clusters. In the following chapter, we will elaborate on the predictive likelihood. The NPAIRS framework (Strother et al., 2002) applies a split-half analysis evaluating the prediction and reproducibility of a model. It was, for instance, used in Andersen et al. (2014) to compare different variations over the stochastic block model.

Reliability

If a clustering is a way of determining an anatomically meaningful parcellation of the human brain, then the inferred parcellations on different groups of subjects should be similar. This raises the question of how to evaluate the similarity of two parcellations. In the neuroimaging literature, three measures are often applied:

1. The Rand index (RI) or adjusted rand index (AR) (Thirion et al., 2014; Janssen et al., 2015).
2. The mutual information (MI) or normalized mutual information (NMI) (Mørup et al., 2010; Andersen et al., 2014; Baldassano et al., 2015) or adjusted mutual information (AMI) (Thirion et al., 2014).
3. Dice index or averaged Dice index (a-Dice) (Blumensath et al., 2013; Craddock et al., 2012).

Vinh et al. (2009) provides a theoretical discussion why it is necessary to adjust these measures for chance. There are several ways ways of doing this and we follow Vinh et al. (2010). The mutual information between two clusterings $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are given by

$$\mathrm{MI}(\boldsymbol{z}_1, \boldsymbol{z}_2) = \sum_{k=1}^{K_1} \sum_{l=1}^{K_2} p(\boldsymbol{z}_1 = k, \boldsymbol{z}_2 = l) \log \left(\frac{p(\boldsymbol{z}_1 = k, \boldsymbol{z}_2 = l)}{p(\boldsymbol{z}_1 = k) p(\boldsymbol{z}_2 = l)} \right) \tag{2.6}$$

The normalized mutual information normalizes the mutual information such that it is in the range between 0 and 1 and we use the following way of normalizing the mutual information:

$$\mathrm{NMI} = \frac{\mathrm{MI}(\boldsymbol{z}_z, \boldsymbol{z}_2)}{\sqrt{H(\boldsymbol{z}_1)H(\boldsymbol{z}_2)}}, \tag{2.7}$$

where H is the entropy and the entropy of a clustering $\boldsymbol{z}$ can be computed as $\mathrm{MI}(\boldsymbol{z}, \boldsymbol{z})$. The NMI between two random clusterings is highly dependent on the number of components that are in the two clusterings. Therefore, the adjusted mutual information corrects for this, and we use the following formulation

$$\mathrm{AMI} = \frac{\mathrm{MI}(\boldsymbol{z}_z, \boldsymbol{z}_2) - \mathrm{E}[\mathrm{MI}(\boldsymbol{z}_z, \boldsymbol{z}_2)]}{\max(H(\boldsymbol{z}_1), H(\boldsymbol{z}_2)) - \mathrm{E}[\mathrm{MI}(\boldsymbol{z}_z, \boldsymbol{z}_2)]}, \tag{2.8}$$

where $\mathrm{E}[\mathrm{MI}(\boldsymbol{z}_z, \boldsymbol{z}_2)]$ is the expected mutual information between two random clusterings with the same number of clusters.

The Rand index is given by $RI(\boldsymbol{z}_1, \boldsymbol{z}_2) = (N_{00} + N_{11})/\binom{N}{2}$, where N_{00} is the number of pairs $i, j \in 1, \ldots, N$ such that $z_i \neq z_j$ in both $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ and N_{11} are the pairs where i and j are in the same component in both clusterings. Again, it is most often used in the form where it is adjusted for chance:

$$\mathrm{AR}(\boldsymbol{z}_1, \boldsymbol{z}_2) = \frac{\mathrm{RI}(\boldsymbol{z}_1, \boldsymbol{z}_2) - \mathrm{E}[\mathrm{RI}(\boldsymbol{z}_1, \boldsymbol{z}_2)]}{\max(\mathrm{RI}(\boldsymbol{z}_1, \boldsymbol{z}_2)) - \mathrm{E}[\mathrm{RI}(\boldsymbol{z}_1, \boldsymbol{z}_2)]}, \tag{2.9}$$

where $\mathrm{E}[\mathrm{RI}(\boldsymbol{z}_1, \boldsymbol{z}_2)]$ again is the expected rand index between two random clusters with the same number of clusters.

The Dice score is normally used as a measure of overlap in two sets X and Y:

$$\mathrm{Dice}(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|}. \tag{2.10}$$

This is extended to the average Dice score between two clusterings $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ the following way: Each parcel in $\boldsymbol{z}_1$ is matched with a parcel in $\boldsymbol{z}_2$ in descending order according to their Dice overlap. The two clusters with largest overlap are matched and the process is repeated omitting the previously considered parcels Blumensath et al. (2013).

Chapter 3

Bayesian Mixture Modeling

Cluster analysis or clustering is the task of dividing a set of objects into groups such that the objects within groups are more similar than objects between two groups. This can be done using a multitude of different algorithms, both probabilistic mixture models and classical clustering algorithms. Opposed to classical clustering algorithms, probabilistic mixture models can quantify uncertainty in the inferred components as illustrated in Figure 3.1.

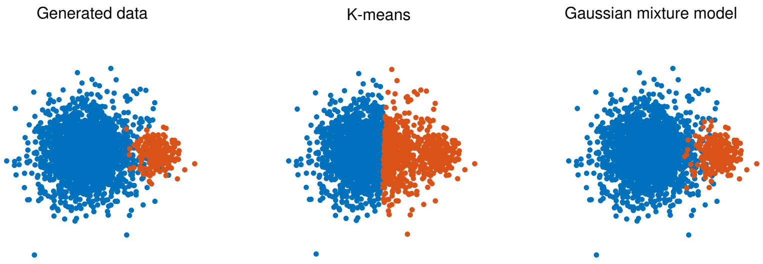


Figure 3.1: Mixture models are able to handle uncertainty and each component is able to have a different noise. This is easy to illustrate in 2 or 3 dimensions but it is perhaps even more important in the high dimensional spaces where fMRI time series reside.

In this chapter, we develop the theory behind probabilistic mixture models and the sampling based inference that we have employed to the problem of whole-brain parcellation in the papers included in this thesis. We follow this with a presentation and discussion of the models we have applied in this thesis.

3.1 Bayesian modeling

The following description of Bayesian modeling is loosely based on Gelman et al. (2014) and Bishop (2006) which are good introductions to the topic. Bayes theorem follows directly from basic probability theory and states that

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{p(x)}, \tag{3.1}$$

and provides the possibility to invert conditional probabilities. Often the theorem is used to invert the relation between observations and the parameters governing the distribution of the observations. In this setting, x might be an observation and θ the parameters governing the likelihood of the data, denoted $p(x \mid \theta)$, $p(\theta)$ is the prior distribution for the parameters, and $p(\theta \mid x)$ is the posterior of the parameters given the data. The final part of the expression, $p(x)$, is constant with regards to the parameters and referred to as the evidence, marginal likelihood, or normalizing constant. It is computed by integrating over the space of possible parameters, i.e. by marginalizing the parameter:

$$p(x) = \int p(x \mid \theta)p(\theta)d\theta. \tag{3.2}$$

In general, this integral is often intractable analytically. The likelihood and the prior together forms the joint distribution as $p(x \mid \theta)p(\theta) = p(x, \theta)$.

In traditional statistical modeling with data $\mathcal{D} = \{\boldsymbol{x}_2, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ the specification of the likelihood $p(\mathcal{D} \mid \boldsymbol{\theta})$ with $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_M\}$ denotes the statistical model and the task of inference is to find the set of parameters $\hat{\boldsymbol{\theta}}$ that optimizes the likelihood of the observed data. A *Bayesian statistical model* is specified by the likelihood distribution along with a prior distribution on the parameters $p(\boldsymbol{\theta})$ or by the joint distribution $p(\mathcal{D}, \boldsymbol{\theta})$. Note that often the prior distribution for $\boldsymbol{\theta}$ includes some parameters and these parameters are called hyperparameters, and are either considered fixed or given an additional layer of priors, possibly with fixed parameters used. For Bayesian inference, we are interested in the posterior distribution of the parameters given the data and not just a single set of parameters. This must be considered when evaluating how well a Bayesian statistical model is able to predict a new observation, $\boldsymbol{x}^*$. This is quantified by the predictive likelihood given by

$$p(\boldsymbol{x}^* \mid \mathcal{D}) = \int p(\boldsymbol{x}^* \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D}) d\boldsymbol{\theta}. \tag{3.3}$$

Here, every set of parameters contribute to the predictive likelihood proportional to how likely they are given the observed data.

A similar question is which of two statistical models are best able to fit an observed dataset. For this calculation we need to further condition on the Bayesian model such that the joint distribution for model m is given by $p(\mathcal{D}, \boldsymbol{\theta} \mid m)$. This means that we can compare two statistical models using the posterior distribution for the model given the data calculated by Bayes theorem

$$p(m \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid m) p(m)}{p(\mathcal{D})}, \tag{3.4}$$

where $p(\mathcal{D} \mid m)$ is the likelihood of the model or the evidence of the data in the setting with a single Bayesian model and $p(\mathcal{D})$ can be calculated by summing over the considered models. This is, for instance, used in Bayesian hierarchical clustering (Heller and Ghahramani, 2005).

One effect of using a prior distribution is that it regularizes the posterior distribution of the parameters and thereby embodies Occam's razor in reducing the complexity of the parameters that are likely and therefore reduce the effect of a few extreme observations. A good review of the effects of using Bayesian modeling in contrast to frequentist modeling can be found in Robert (2007).

3.1.1 Exponential family and conjugate priors

All the distributions of this thesis are members of the exponential family (Bishop, 2006) which means the probability density function can be described as

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = h(\boldsymbol{x}) g(\boldsymbol{\theta}) \exp(\theta^\top u(\boldsymbol{x})), \tag{3.5}$$

where u is a function mapping the data to it's sufficient statistics, $\boldsymbol{\theta}$ are the canonical or natural parameters associated to the vector of sufficient statistics and $g(\boldsymbol{\theta})$ is the partition function that ensures that the probability distribution integrates to one over the domain of $\boldsymbol{x}$.

What is most important for this thesis is that the members of the exponential family have natural conjugate priors. A prior $p(\boldsymbol{\theta})$ is conjugate if it leads to a posterior $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ with the same functional form as the prior. The prior for a member of the exponential family can then be written in the form

$$p(\boldsymbol{\theta} \mid \boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu) g(\boldsymbol{\theta})^{\nu} \exp(\nu \boldsymbol{\theta}^{\top} \boldsymbol{\chi}). \tag{3.6}$$

This means that, provided the prior has a closed form, it is possible to evaluate the integral for the marginal likelihood analytically and for Bayesian mixture models it allows us to marginalize the parameters for easier inference.

3.2 Mixture models

In mixture models, data is modeled independently from a mixture of several different probability distributions or components such that each of these components contribute to the likelihood of the data. This can be described by the following probabilistic model:

$$p(\boldsymbol{x}_{1:N} \mid \boldsymbol{\theta}_{1:K}, \boldsymbol{\pi}) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}_i \mid \boldsymbol{\theta}_k), \tag{3.7}$$

where N is the number of observations, K the number of components, and π_k for $k = 1, \ldots, K$ are mixing coefficients that satisfies $\sum_{k=1}^{K} \pi_k = 1$ such that it is a proper distribution. It is easy to determine parameters that optimize a single component with respect to the observations in the component but the dependencies between the mixing coefficient and component parameters makes it difficult to jointly optimize the parameters for both.

In a Bayesian mixture model, the parameters are also considered stochastic and thus require the specification of priors. Furthermore, the latent variable $\boldsymbol{z} = \{z_1, \ldots, z_N\}$ is introduced such that $z_i = k$ if the i'th observation is part of the k'th component. The joint distribution of the Bayesian mixture model can thus be specified as

$$p(\boldsymbol{x}_{1:N}, \boldsymbol{\theta}_{1:K}, \boldsymbol{\pi}, \boldsymbol{z}) = p(\boldsymbol{\pi}) p(\boldsymbol{\theta}) \prod_{i=1}^{N} p(\boldsymbol{x}_i \mid \boldsymbol{\theta}_{z_i}) p(z_i \mid \boldsymbol{\pi}), \tag{3.8}$$

where $p(z_i \mid \boldsymbol{\pi})$ are the probabilistic distribution that corresponds to the mixing coefficients determined by parameter $\boldsymbol{\pi}$ and with prior $p(\boldsymbol{\pi})$.

3.2.1 Prior on the latent variable, $\boldsymbol{z}$

The multinomial distribution can be used as prior on the latent variable, $p(\boldsymbol{z} \mid \boldsymbol{\pi})$. It is part of the exponential family with the Dirichlet distribution the associated conjugate prior. The multinomial distribution is specified by mixing parameter $\boldsymbol{\pi}$:

$$p(\boldsymbol{z} \mid \boldsymbol{\pi}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \pi_k^{\delta(k,z_i)}, \tag{3.9}$$

such that $\sum_{k=1}^{K} \pi_k = 1$ and $\delta(a, b)$ is the delta function which is one if $a = b$ and zero otherwise. The Dirichlet distribution is given by

$$p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)^K} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}. \tag{3.10}$$

In general, we are not interested in the mixing parameter $\boldsymbol{\pi}$ and using conjugacy it is easy to evaluate the marginal likelihood analytically:

$$p(\boldsymbol{z} \mid \boldsymbol{\alpha}) = \int p(z_i \mid \boldsymbol{\pi}) p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) d\boldsymbol{\pi} = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\Gamma(N + \sum_{k=1}^{K} \alpha_k)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \alpha_k)}{\Gamma(\alpha_k)}, \tag{3.11}$$

where $n_k = \sum_{i=1}^{N} \delta(z_i, k)$ is the number of elements in component k. The marginalized distribution is known as the Dirichlet-multinomial distribution or the Pólya distribution. Furthermore, since we have no prior knowledge on the mixing, the parameter $\boldsymbol{\alpha}$ of the Dirichlet distribution is often fixed such that $\alpha_k = \alpha/K$ for $i = 1, \ldots, K$.

It is useful to consider the probability of adding element i to cluster k:

$$p(z_i = k \mid \boldsymbol{z}_{\setminus i}, \alpha) = \frac{n_k + \alpha/K}{N - 1 + \alpha}, \tag{3.12}$$

where n_k is the number of elements in cluster k. This illustrates the effect of marginalizing the mixing parameter; now the equivalent of the mixing parameter is $n_k + \alpha/K$ and the probability of being added to a cluster increases with its size enforcing the rich-get-richer principle. It is possible to take the limit $K \to \infty$ and for ease of notation let D be the number of occupied clusters. The probability that element i is added to cluster k is then given by

$$p(z_i = k \mid \boldsymbol{z}_{\setminus i}, \alpha) = \frac{n_k + \alpha/K}{N - 1 + \alpha} \to \frac{n_k}{N - 1 + \alpha} \text{ as } K \to \infty, \tag{3.13}$$

and for one of the $K - D$ empty clusters

$$p(z_i = k \mid \boldsymbol{z}_{\setminus i}, \alpha) = \frac{\alpha(K - D)/K}{N - 1 + \alpha} \to \frac{\alpha}{N - 1 + \alpha} \text{ as } K \to \infty. \tag{3.14}$$

Gathering the joint distribution $p(\boldsymbol{z} \mid \alpha) = p(z_1)p(z_2 \mid z_1)\ldots p(z_N \mid \boldsymbol{z}_{\setminus N})$ yields the distribution known as the Chinese Restaurant Process (Aldous, 1985) and is given by

$$\mathrm{CRP}(\boldsymbol{z} \mid \alpha) = \frac{\Gamma(\alpha)\alpha^K}{\Gamma(N+\alpha)} \prod_{k=1}^{K} \Gamma(n_k). \tag{3.15}$$

That this is the case can be illustrated by constructing the joint distribution for the elements of cluster k by Equations (3.13) and (3.14):

$$p(\{z_i\}_{i \in \mathcal{Z}_k} \mid \alpha) = \frac{\alpha}{\alpha} \prod_{j=2}^{n_k} \frac{j-1}{j+\alpha} = \frac{\alpha\Gamma(\alpha)\Gamma(n_k)}{\Gamma(n_k+\alpha)}, \tag{3.16}$$

where $\mathcal{Z}_k$ is the list of elements in cluster k. Employing the CRP in mixture models has several advantages. Primarily the mixture model is now a distribution over any possible clustering and not just clusterings with K components. The posterior distribution $p(\boldsymbol{z}, \boldsymbol{\theta} \mid \boldsymbol{x}_{1:N})$ will therefore have contributions from clusterings with many different number of components, the mode will be able to select the number of clusters that best describes the data and it is possible to compute the posterior distribution for the number of clusters in the data. Finally, since all empty clusters are considered together inference is more efficient.

The CRP is an alternative view of the Dirichlet process and the mixture models with CRP prior are also known as Dirichlet process mixtures as introduced by (Ferguson, 1973). The prior on the CRP prior, α, controls the prior distribution on the number of clusters. Miller and Harrison (2013) criticizes Dirichlet process mixtures providing both theoretical and experimental evidence that they tend to overestimate the number of clusters, with a fixed CRP parameter α. We briefly address this issue in §3.5.1.

Typically, a Bayesian mixture model is described by the generative process:

$$\boldsymbol{z} \mid \alpha \sim p(\boldsymbol{z} \mid \alpha) \tag{3.17}$$

$$\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_0 \overset{iid}{\sim} p(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_0) \tag{3.18}$$

$$\boldsymbol{x}_i \mid \boldsymbol{z}, \boldsymbol{\theta} \overset{ind}{\sim} p(\boldsymbol{x}_i \mid \boldsymbol{\theta}_{z_i}), \tag{3.19}$$

here with marginalized mixing parameter $\boldsymbol{\pi}$ and where $\overset{iid}{\sim}$ and $\overset{ind}{\sim}$ denote independently and identically and just independently drawn samples respectively. The prior distribution for the parameters, $p(\boldsymbol{\theta}_k)$ is typically requires specification of a number of hyperparameters denoted here by $\boldsymbol{\theta}_0$. This formulation explicitly denotes the dependencies in the statistical model and provides an easy recipe for generating data; first sample a clustering from the prior $p(\boldsymbol{z} \mid \alpha)$, then components $\boldsymbol{\theta}_k$ iid. from $p(\boldsymbol{\theta}_k)$ for each k, and finally observations independently from their respective component distributions. Note that this formulation is equivalent to the specification of the joint distribution of Equation (3.8).

3.3 Inference

From the maximum likelihood perspective we are interested in a set of parameters $\boldsymbol{\theta}_{1:K}$ that optimizes likelihood in Equation (3.7), i.e. the clustering configuration and set of cluster parameters for each that are most likely. By $\boldsymbol{\theta}_{1:K}$ we denote the matrix with $\boldsymbol{\theta}_k$ as column vectors for $k = 1, \ldots, K$. For the Bayesian perspective, the primary object of interest is the posterior distribution of the parameters given the data. With N observations in K possible clusters, there are K^N possible assignments or Bell(N) that are equivalent under permutations of the cluster labels. The dependencies between the component parameters and the clustering labels makes estimating the parameters in both the maximum likelihood problem and for evaluating the posterior distribution for the Bayesian problem NP hard.

The joint distribution with marginalized $\boldsymbol{\pi}$ for a finite Bayesian mixture model is given by

$$p(\boldsymbol{x}_{1:N}, \boldsymbol{\theta}_{1:K}, \boldsymbol{z} \mid \alpha) = p(\boldsymbol{z} \mid \alpha) \left[\prod_{k=1}^{K} p(\boldsymbol{\theta}_k) \right] \prod_{i=1}^{N} p(\boldsymbol{x}_i \mid \boldsymbol{\theta}_{z_i}), \tag{3.20}$$

and to compute the posterior distribution for the parameters, $p(\boldsymbol{\theta}_{1:K}, \boldsymbol{z} \mid \boldsymbol{x}_{1:N}, \alpha)$ using Bayes theorem we are required to compute the evidence, $p(\boldsymbol{x}_{1:N} \mid \alpha)$. Integrating over the parameters $\boldsymbol{\theta}_{1:K}$ and latent variables $\boldsymbol{z}$ gives the following expression for the evidence

$$p(\boldsymbol{x}) = \int p(\boldsymbol{\theta}_{1:K}) \prod_{i=1}^{N} \sum_{z_i=1}^{K} P(z_i = k \mid \alpha) p(\boldsymbol{x}_i \mid z_i, \boldsymbol{\theta}_k) d\boldsymbol{\theta}_{1:K} \tag{3.21}$$

Using conjugate priors it is possible to marginalize the parameters $\boldsymbol{\theta}_k$ if the integrand in Equation (3.21) factors in k but this is not the case here and the time complexity of evaluating the integral is $\mathcal{O}(K^N)$. If we rearrange and distribute over the sum, we can write the evidence as the following:

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{z}} p(\boldsymbol{z}) \int p(\boldsymbol{\theta}_{1:K}) \prod_{i=1}^{N} p(\boldsymbol{x}_i \mid \boldsymbol{\theta}_{1:K}) d\boldsymbol{\theta}_{1:K} \tag{3.22}$$

The integral here factorizes such that each factor only contains a contribution from the k'th component. Now, however, there are K^N possible configurations in the sum over $\boldsymbol{z}$ and computing the evidence therefore remains exponential in N and intractable (Blei et al., 2016).

There are several possible approximations to do inference in mixture models. For a non-Bayesian mixture model, the most popular approximation is the expectation maximization algorithm (EM). EM is an iterative approach that first performs an expectation step that optimizes $p(\boldsymbol{z} \mid \boldsymbol{x}_{1:N}, \boldsymbol{\theta}_{1:K})$ and $p(\boldsymbol{\theta}_{1:K} \mid \boldsymbol{x}_{1:N}, \boldsymbol{z})$ iteratively. This is followed by an expectation step where the configuration is changed to optimize the likelihood based on the parameters. This process is guaranteed to converge to a solution but the solution is highly dependent on the initial configuration and therefore susceptible to local minima. This can be addressed by using multiple restarts from random initializations but that does not solve the issue. For a further discussion on this topic and expectation maximization, see Bishop (2006).

In a Bayesian mixture model, there are two main approaches: Variational or sampling based inference. In variational inference (VI), sometimes also referred to as variational Bayes, the joint distribution is approximated by a factorized distribution q within some variational family. The "distance" or divergence of the two distributions is then minimized, often as measured by the Kullback-Leibler (KL) divergence, $\mathrm{KL}(p||q)$, which for continuous distributions is given by

$$\mathrm{KL}(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx. \tag{3.23}$$

Note that the KL divergence is not a metric since it does not satisfy the triangle inequality. This is a general approach for doing inference and in the setting of a mixture model the only approximation that is required is that the posterior distribution $q(\boldsymbol{\theta}, \boldsymbol{z} \mid \mathcal{D})$ factors into $q(\boldsymbol{z} \mid \mathcal{D})q(\boldsymbol{\theta} \mid \mathcal{D})$. Then the parameters of the distributions can be optimized in an iterative procedure that resembles the EM algorithm; first the expectation of $q(\boldsymbol{z})$ is optimized given the $\boldsymbol{\theta}$ parameters and vice versa. As for the traditional EM algorithm, this procedure is also highly susceptible to local minima. In literature there are many examples of sampling based inference in mixture models and for a general introduction we refer to Bishop (2006) or Jordan et al. (1999); Blei et al. (2016).

Sampling based inference is a very general tool and sampling based inference is also applicable for marginalized mixture models. Using a sufficient number of samples, it is possible to approximate the posterior distribution with arbitrary precision. With no closed from expression for the posterior distribution, we can use the samples to compute the relevant quantities such as the mode, quantiles, mean etc. for the approximated distribution. In theory, and often in practice, sampling based inference will approximate any distribution given sufficient samples and is thus not susceptible to local minima like the VI or

EM algorithms. In practice, however, it is not always feasible to collect enough samples for a good approximation. In this thesis, we have followed the path of sampling based inference and apply state of the art sampling techniques to try combat the downsides. In the following sections I will present the sampling techniques that have been applied in this thesis.

3.4 Sampling based inference

Sampling is a very general method for approximate inference in probabilistic problems and instead of optimizing a parametric distribution, we sample from the distribution and can then use the samples to compute the quantities of interest.

For mixture models, we are interested in the posterior distribution $p(\boldsymbol{\theta}_{1:K}, \boldsymbol{z} \mid \boldsymbol{x}_{1:N}, \alpha)$ or possibly $p(\boldsymbol{z} \mid \boldsymbol{x}_{1:N}, \alpha)$ with marginalized parameters. Note, that since the posterior distribution is proportional to the joint distribution we must resort to the sampling techniques that does not require the distribution, we are sampling from, to be normalized such as many Markov chain Monte Carlo samplers.

3.4.1 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) is a technique for generating samples using a Markov chain, ie. to generate samples $\nu^{(0)}, \nu^{(1)}, \ldots, \nu^{(M)}$ where each generated sample is only dependent on the previous sample and a transition kernel, $T(\nu^{(t)} \mid \nu^{(t-1)}) = p(\nu^{(t)} \mid \nu^{(t-1)})$. Provided the Markov chain is *ergodic* and *irreducible*, then the Markov chain will converge to a unique stationary distribution, denoted *equilibrium* distribution $p^*(\nu)$ (also called *stationary* or *invariant*). p^* is invariant if

$$p^*(\nu) = \sum_{\nu'} T(\nu \mid \nu') p^*(\nu'), \tag{3.24}$$

and it is irreducible if it is possible to go from ν to any other ν' in a finite number of steps, and if the limiting distribution $p(\nu^{(m)}) \to p^*(\nu)$ for $m \to \infty$ regardless of the choice of initial distribution $p(\nu^{(0)})$ it is also ergodic. A sufficient, but not necessary, criterion is for the chain to satisfy *detailed balance*:

$$p^*(\nu) T(\nu' \mid \nu) = p^*(\nu') T(\nu \mid \nu'). \tag{3.25}$$

It is, perhaps surprisingly, quite easy to construct a transition kernel such that the Markov chain satisfies either detailed balance or that it is ergodic, irreducible with the proper equilibrium distribution (Bishop, 2006; Neal, 1993).

3.4.2 Metropolis-Hastings

One of the most widely used Markov Chain Monte Carlo methods is the Metropolis-Hastings algorithm (Hastings, 1970). Based on some initial value in the Markov chain it uses a proposal distribution $q(\nu' \mid \nu)$ to generate a proposed next step in the Markov chain ν' and this proposal is then accepted with probability $\alpha(\nu', \nu)$ which is defined by

$$\alpha(\nu', \nu) = \min\left(1, \frac{\tilde{p}(\nu')q(\nu \mid \nu')}{\tilde{p}(\nu)q(\nu' \mid \nu)}\right), \tag{3.26}$$

where $\tilde{p}(\nu) = Zp(\nu)$ for some constant Z emphasizing that it is sufficient to evaluate the distribution up to a constant. If the proposal ν' is not accepted ν is used as the next element in the Markov chain. This gives the following transition operator:

$$T(\nu' \mid \nu) = \overbrace{q(\nu' \mid \nu)\alpha(\nu', \nu)}^{\text{Acceptance}} + \delta(\nu', \nu)\overbrace{\left[1 - \int q(\tilde{\nu} \mid \nu)\alpha(\tilde{\nu}, \nu)d\tilde{\nu}\right]}^{\text{Rejection}}. \tag{3.27}$$

Note, that if $\nu' = \nu$ we have either proposed the previous state or we have rejected a state and the probability staying at ν will therefore have a contribution from both the case where we accept and reject a proposal. It is easily seen that this transition density satisfies the detailed balance criterion. For $\nu' \neq \nu$ the transition probability is $T(\nu' \mid \nu) = q(\nu' \mid \nu)\alpha(\nu', \nu)$ and multiplying with $Zp(\nu)q(\nu' \mid \nu)$ on both sides of Equation (3.26) yields

$$\begin{aligned} p(\nu)q(\nu' \mid \nu)\alpha(\nu', \nu) &= \min(p(\nu)q(\nu' \mid \nu), p(\nu')q(\nu \mid \nu')) \\ &= \alpha(\nu, \nu')p(\nu')q(\nu \mid \nu') \end{aligned} \tag{3.28}$$

and the constructed transition kernel thus satisfies Equation (3.25). If $\nu' = \nu$ then a proposal was rejected and the detailed balance principle is clearly satisfied as well.

The Metropolis-Hastings algorithm is a generalization of the Metropolis sampling method Metropolis et al. (1953) that uses a symmetric proposal distribution simplifying the expression in Equation (3.26). Metropolis-Hastings does not necessarily produce correlated samples which can be observed by simply using the desired distribution as the proposal distribution independent on the previous sample. In practice, however, the proposal distributions are selected for convenience, for instance using a Gaussian distribution centered on the previous point. If the variance of the proposal distribution is set too high, the proposals will be far from the previous proposal and rarely accepted while if it is too low most proposals will be accepted but the Markov chain will for a finite number of samples be unable to satisfactorily explore the desired distribution.

This tradeoff between the level of dependency between samples and the ability of the sampler to explore the distribution is always apparent when sampling in high dimensional spaces and the Gibbs sampler makes even more correlated samples by only consider the change of a single variable z_i of $\boldsymbol{z}$.

Gibbs Sampling

Gibbs sampling (Geman and Geman, 1984) can be viewed as a special case of MH where we are guaranteed to accept the proposals. Instead of generating a new sample for the entire vector ν, which might be high dimensional, in Gibbs sampling we iterate the dimensions of ν. Let $\nu = \{\nu_1, \dots, \nu_M\}$ for some M and in each step of the Gibbs sampler we generate a new proposal from $p(\nu_i \mid \boldsymbol{\nu}_{\setminus i})$. Correctness for the Gibbs sampler is most easily shown by showing that the correct distribution is invariant and that it is irreducible and ergodic. It is easily seen that the distribution $p(\nu)$ is invariant to each of the Gibbs steps individually and thus to the whole chain since we are sampling from the correct conditional distributions. Furthermore, if none of the conditional distributions are anywhere zero it is also irreducible and ergodic.

To see that Gibbs is a special case of Metropolis-Hastings, let the proposal ν^* be such that $\nu_i^* = \nu_i'$ and $\nu_{\setminus i}^* = \nu_{\setminus i}$. Then the Metropolis-Hastings acceptance probability is

$$\alpha(\nu^*, \nu) = \min\left(1, \frac{p(\nu^*)q(\nu \mid \nu^*))}{p(\nu)q(\nu^* \mid \nu))}\right) = \min\left(1, \frac{p(\nu_i' \mid \nu_{\setminus i})p(\nu_{\setminus i})p(\nu_i \mid \nu_{\setminus i}^*)}{p(\nu_i \mid \nu_{\setminus i})p(\nu_{\setminus i})p(\nu_i' \mid \nu_{\setminus i})}\right) = 1, \tag{3.29}$$

and the proposals are thus always accepted. Note that the Markov chain generated by Gibbs sampling is highly correlated and the chain should be thinned to acquired uncorrelated samples. If it is not possible to draw samples from $p(z_i \mid \boldsymbol{z}_{\setminus i}, \mathcal{D})$ directly it is often useful to embed Metropolis-Hasting sampling in Gibbs sampling for some of the variables in the model.

Now we return to the Bayesian mixture models as specified by the generative process of Equations (3.17)-(3.19), where we, in the notation, ignore the dependencies on the hyperparameters. The posterior distribution for the i'th index in the latent variable $\boldsymbol{z}$ is given by

$$p(z_i \mid \boldsymbol{z}_{\setminus i}, \boldsymbol{x}_{1:N}) = \frac{p(\boldsymbol{x}_{1:N} \mid z_i, \boldsymbol{z}_{\setminus i}) p(z_i \mid \boldsymbol{z}_{\setminus i})}{\sum_{z_i} p(\boldsymbol{x}_{1:N} \mid z_i, \boldsymbol{z}_{\setminus i}) p(z_i \mid \boldsymbol{z}_{\setminus i})} \tag{3.30}$$

which is a categorical distribution over the possible components that z_i can be assigned to. In Equations (3.13) and (3.14) we derived the expression for $p(z_i \mid \boldsymbol{z}_{\setminus i})$ and the probability of assigning element i to component k is therefore given by:

$$P(\boldsymbol{z}_i = k \mid \boldsymbol{z}_{\setminus i}, \boldsymbol{x}_{1:N}) \propto \begin{cases} \frac{n_k}{\alpha+N-1} p(\boldsymbol{x}_{\mathcal{Z}_k \cup i} \mid \boldsymbol{z}_{\setminus i}, z_i) & \text{if } k \text{ populated} \\ \frac{\alpha}{\alpha+N-1} p(\boldsymbol{x}_i) & \text{if } k \text{ empty} \end{cases} \tag{3.31}$$

where the proportionality is the denominator in Equation (3.30) and shared for both cases. $p(\boldsymbol{x}_{\mathcal{Z}_k \cup i} \mid \boldsymbol{z}_{\setminus i}, z_i)$ is marginalized contribution to the joint distribution for component k, i.e.

$$p(\boldsymbol{x}_{\mathcal{Z}_k \cup i} \mid \boldsymbol{z}_{\setminus i}, z_i) = \int p(\boldsymbol{x}_{\mathcal{Z}_k \cup i} \mid \boldsymbol{\theta}_k, \boldsymbol{z}_{\setminus i}, z_i) p(\boldsymbol{\theta}_k) d\boldsymbol{\theta}_k, \tag{3.32}$$

and for a new component

$$p(\boldsymbol{x}_i) = \int p(\boldsymbol{x}_i \mid \boldsymbol{\theta}_{K+1}) p(\boldsymbol{\theta}_{K+1}) d\boldsymbol{\theta}_{K+1}. \tag{3.33}$$

Since the proportionality constant is shared for all possibilities this can be efficiently calculated for conjugate mixture models. For non-conjugate mixture models we need to propose new parameters for the new cluster, for example using the algorithm proposed by (Neal, 2000, Algorithm 8).

Sequentially-Allocated Merge-Split Sampler

The problems of correlated samples are especially detrimental when sampling the latent assignment variables. It is frequently the case that a cluster is supposed to be split into two clusters but every step towards splitting the cluster decreases the joint distribution even though it it is favorable in the joint distribution for the cluster to be divided. Split-merge and its derivative, sequentially-allocated merge-split, are methods for jumping directly to the configuration where the cluster split in two.

The Gibbs procedure for sampling the clustering parameter can become trapped in local modes due to the incremental nature of the Gibbs sampler (Celeux et al., 2000; Albers et al., 2013). For instance, to escape a mode where one component is wrongly divided in two components with similar parameters, the Gibbs sampler must transition through a number of states of low probability. The split-merge algorithm was introduced by Jain and Neal (2004) to allow more flexible sampling in Dirichlet process mixtures addressing this problem. Later, Dahl (2005) proposed the sequentially-allocated merge-split (SAMS) algorithm that slightly changes the split-merge algorithm and shows that it, in many cases, is an improvement over the split-merge algorithm. Both methods are implementations of the Metropolis-Hastings sampling method and uses the evaluation of the categorical distribution in the Gibbs sampler to generate the proposal for the latent variables in the Dirichlet process mixture models. In this thesis, we employ both methods and here follows a thorough description of the SAMS algorithm and a discussion of the changes between split-merge and SAMS as well as our proposal for changing the order in the procedure to minimize the time spent evaluating the transition probability of proposals that will be rejected regardless.

For the following again consider a nonparametric mixture model as in Equations (3.17)-(3.19) with N observations divided into K occupied clusters. The SAMS algorithm is specified in the following procedure:

1. Sample indices i and j uniformly at random from $\{1, \dots, N\}$.
2. Generate proposal state

   Propose split if $z_i = z_j$**, denote** $S = \{n \mid z_n = z_i, n \neq z_i, n \neq z_j\}$

   Compute the split proposal state by the following:

   (a) Remove indices i and j from S and form singletons $S^i = \{i\}$ and $S^j = \{j\}$ and let $z_j = K + 1$.

   (b) For n in S in random order, assign n according to the categorical distribution

   $$p(z_n \in \{i, j\} \mid \boldsymbol{x}_{S^k}, S^i \cup S^j). \tag{3.34}$$

   and add n to the associated set, S^i or S^j. Note, that this is equivalent to the distribution in Gibbs sampling and is computed similarly.

   Propose merge if $z_i \neq z_j$**, denote** $S = \{n \mid z_n = z_i$ **or** $z_j, n \neq z_i, n \neq z_j\}$

   For the merge proposal state assign all elements of z_i and z_j to z_i and remove component z_j.

3. Compute the acceptance probability α:

$$\alpha(\boldsymbol{z}^*, \boldsymbol{z}) = \min\left(1, \frac{p(\boldsymbol{z}^* \mid \mathcal{D})q(\boldsymbol{z}, \boldsymbol{z}^*)}{p(\boldsymbol{z} \mid \mathcal{D})q(\boldsymbol{z}^*, \boldsymbol{z})}\right), \tag{3.35}$$

where the transition probabilities $q(\boldsymbol{z}, \boldsymbol{z}^*)$ and $q(\boldsymbol{z}^*, \boldsymbol{z})$ must be computed as follows:

$\boldsymbol{z}^*$ **split** For the split proposal $q(\boldsymbol{z}, \boldsymbol{z}^*)$ is transition probability of the merge. Since this can only happen in one way $q(\boldsymbol{z}, \boldsymbol{z}^*) = 1$. On the other hand, there are several possible split configurations and the transition probability to $\boldsymbol{z}^*$ is the product of the probabilities in Equation (3.34) from the selected assignments.

$\boldsymbol{z}^*$ **merge** For the merge, by similar arguments, $q(\boldsymbol{z}^*, \boldsymbol{z}) = 1$ and the transition probability $q(\boldsymbol{z}, \boldsymbol{z}^*)$ is the probability of transitioning to precisely $\boldsymbol{z}$ out of all the possible splits. Therefore $q(\boldsymbol{z}, \boldsymbol{z}^*)$ is the product of the probabilities given by Equation (3.34) where the i'th observation is assigned to the component z_i, ie. the component it was assigned to before the SAMS procedure.

The split-merge algorithm is quite similar to the SAMS algorithm. The main change is that instead of assigning the elements of S sequentially to generate a proposal state for the split case, the elements are assigned the two components at random followed by a number of Gibbs sweeps that are restricted to only sample the elements of S. These restricted Gibbs sweeps refines the proposal state from a socalled launch state to make it more likely and therefore increase the acceptance ratio. This also entails minor changes to the transition probabilities and it must now be computed as the probability of transitioning from the launch state to the proposed state.

We propose a small but crucial change to the order in which the quantities in both the SAMS and split-merge procedures are computed for case where the merge of two components is considered. The acceptance probability α of Equation (3.35) is bounded by the ratio of the posterior distributions

$$\alpha(\boldsymbol{z}^*, \boldsymbol{z}) = \min\left(1, \frac{p(\boldsymbol{z}^* \mid \mathcal{D})}{p(\boldsymbol{z} \mid \mathcal{D})} \frac{q(\boldsymbol{z}, \boldsymbol{z}^*)}{q(\boldsymbol{z}^*, \boldsymbol{z})}\right) \geq \min\left(1, \frac{p(\boldsymbol{z}^* \mid \mathcal{D})}{p(\boldsymbol{z} \mid \mathcal{D})}\right) \tag{3.36}$$

Therefore, without having to evaluate the transition probability, it is possible to reject the merge proposal with probability $\min\left(1, \frac{p(\boldsymbol{z}^*|\mathcal{D})}{p(\boldsymbol{z}|\mathcal{D})}\right)$. In case we cannot reject the merge proposal based on this initial test, continue to calculate the transition probabilities and accept the merge proposal with probability $\min\left(1, \frac{q(\boldsymbol{z},\boldsymbol{z}^*)}{q(\boldsymbol{z}^*,\boldsymbol{z})}\right)$ such that the total probability of acceptance is the same as in Equation (3.35).

For practical problems, this small change has a considerable impact on the time the samplers spends in evaluating the merge proposals. It is difficult to say something general on this since it is highly dependent on the distribution on the size of clusters and on the likelihood of accepting a merge of two clusters at random. Instead we compared the impact on performance on whole-brain fMRI data in paper A. In general, the SAMS or split-merge samplers are very important for the mixing properties of the sampler and it is but once a level of convergence has been reached the impact of these proposals lessens.

Since the both the split-merge and SAMS sampler are manifestations of the Metropolis-Hastings algorithm, the correctness for the SAMS algorithm is therefore quite trivial, but the split-merge algorithm it is the transition from the random launch state to the state with two clusters that is the transition probability. The proof that this transition probability equals the transition probability from the initial state to the state with two clusters can be found in Jain and Neal (2004).

Change of variables

One way to handle constraints for hyperparameters is using a change of variables for the distributions. In general, if $p(x)$ is a probability distribution over some space X it is possible to compute the distribution of $Y = g(X)$. If $g\colon X \to Y$ is a monotonic function, then the distribution of Y, $p_Y(y)$ is given by

$$p_Y(y) = \left|\frac{d}{dy}(g^{-1}(y))\right| p_X(g^{-1}(y)) \tag{3.37}$$

In the case of sampling of hyperparameters we want to draw samples around some $y_0 \in Y$ but it is much easier for us to draw samples around $g^{-1}(y_0) \in X$. Using change of variables, we can do exactly that. Two frequently used transformations in this thesis are the transformations $g\colon \mathbb{R} \to \mathbb{R}_+$ using $g(x) = \exp(x)$ and $g\colon \mathbb{R} \to (0,1)$ using a sigmoid function such as $g(x) = 1/(1 + \exp(-x))$.

3.4.3 Verifying correctness of sampler

Sampling can be surprisingly resistant to minor errors in the implementation and still recover the true clustering on synthetic data. Since the sampling procedure is often computationally demanding all steps in the sampling procedure must be optimized from the evaluation of the joint distribution to the computation of the categorical distribution in the Gibbs sweeps of Equation (3.30), the evaluation of the transition in the SAMS sampler of Equation (3.35), and the evaluation of the Metropolis-Hastings proposals for the hyperparameters of

Equation (3.26). Ensuring that these equations are correctly computed can be done by unit testing them as suggested by Grosse and Duvenaud (2014).

Using the fact that

$$\frac{p(x' \mid z)}{p(x \mid z)} = \frac{p(x', z)}{p(x, z)} \tag{3.38}$$

it is possible to verify that each of the optimized evaluations of the posterior distributions can be verified against the ratio of the joint distribution.

Furthermore, it is possible to jointly test the samplers on small problems where it is possible to analytically evaluate the evidence by brute force summing over all possible configurations as in Equation (3.22). An example of this is provided in Figure 3.2.

3.5 Models

With either the Chinese restaurant process or the Dirichlet-multinomial distributions as prior for the clustering $p(\boldsymbol{z})$ we need to further specify a likelihood $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ and a prior on the parameters $p(\boldsymbol{\theta})$ to complete the specification of a Bayesian mixture model. We here further present the marginal likelihood for N observations in a component as using this makes calculating the posteriors in the Gibbs sampler for a mixture model primarily a question of book keeping. The focus of this thesis is mixture models in the context of whole-brain clustering and therefore we must factor in the domain knowledge:

1. The problem is large scale and inference in the clustering models must scale linearly in both the number of subjects, temporal dimension, number of clusters, and number of voxels. A model that does not satisfy this criterion will be intractable for the sampling based inference.

2. The BOLD response as measured by the MR scanner is time-series data and we would expect the signal to be temporally smooth in accordance with the temporal dynamics of the BOLD response.

3. The fMRI data is often heavily preprocessed using spatial smoothing and normalization of the data using z-scoring.

The simplest, and probably most widely used, example is the Gaussian mixture model which can be phrased in three levels of complexity in the prior for the

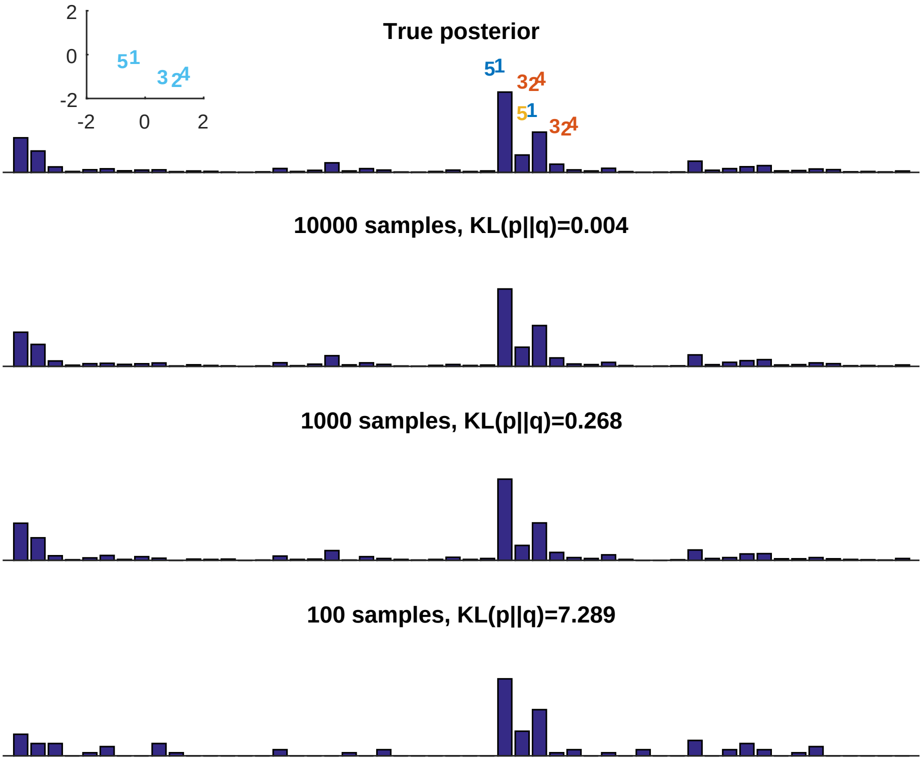


Figure 3.2: Comparison between the inferred posterior distribution by sampling and the exact posterior distribution by manually evaluating Bell(5)=52 possible configurations for $N = 5$ observations. Each bar represents the posterior probability of that particular $\boldsymbol{z}$ parameter. Note how the Kullback-Leibler divergence diminishes as more samples are used to approximate the posterior distribution. In the top left corner is the generated data two dimensional data and the two most likely partitions are illustrated by color-coding the cluster label for each observation above the true posterior distribution.

parameters; with the covariance matrix as either a full rank, diagonal, or diagonal with identical elements (Gilles and Gérard, 1995; Rasmussen, 1999). For an illustration of the three types of Gaussians, see Figure 3.3. We include the Gaussian distributions based mixture models as a benchmark against traditional mixture models.

Normal Inverse Wishart distribution

The multivariate Gaussian or normal distribution is a distribution over $\mathbb{R}^{\mathrm{D}}$ and is given by

$$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{3.39}$$

The conjugate prior for the parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the Normal inverse Wishart distribution ($\mathcal{NIW}$):

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \gamma, \nu) = \mathcal{N}(\mu \mid \boldsymbol{\mu}_0, \gamma\boldsymbol{\Sigma})\mathcal{IW}(\boldsymbol{\Sigma} \mid \boldsymbol{\Sigma}_0, \nu), \tag{3.40}$$

where $\mathcal{IW}$ is the inverse Wishart distribution with probability density function

$$\mathcal{IW}(\boldsymbol{\Sigma} \mid \boldsymbol{\Sigma}_0, \nu) = \frac{|\boldsymbol{\Sigma}_0|^{\nu/2}}{2^{\frac{\nu D}{2}}\Gamma_D(\frac{\nu}{2})}|\Sigma|^{-\frac{\nu+D+1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}^{-1})\right), \tag{3.41}$$

where Γ_D is the multivariate Gamma function, $\boldsymbol{\Sigma}_0$ is a positive definite matrix called the scale matrix, and $\nu_0 > D-1$ is the degrees of freedom parameter.

Using conjugacy, it is possible to analytically marginalize the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for the marginal likelihood:

$$\begin{aligned} &p(\boldsymbol{x}_{1:N} \mid \boldsymbol{\Sigma}_0, \nu) = \\ &\frac{|\boldsymbol{\Sigma}_0|^{\nu/2}\Gamma_p(\frac{\nu+p}{2})(\gamma/(N+\gamma))^{D/2}2^{(N+\nu)D/2}}{(2\pi)^{\frac{ND}{2}}2^{\nu D/2}\Gamma_D(\nu/2)|\bar{\boldsymbol{\Sigma}}+\boldsymbol{\Sigma}_0+\boldsymbol{\mu}_0\boldsymbol{\mu}_0^\top-\frac{1}{N+\gamma}(\bar{\boldsymbol{x}}+\gamma\boldsymbol{\mu}_0)(\bar{\boldsymbol{x}}+\gamma\boldsymbol{\mu}_0)^\top|^{\frac{\nu+D}{2}}}, \end{aligned} \tag{3.42}$$

where $\bar{\boldsymbol{x}} = \sum_{i=1}^N \boldsymbol{x}_i$ and $\bar{\boldsymbol{\Sigma}} = \sum_{i=1} \boldsymbol{x}_n\boldsymbol{x}_n^\top$. Note that this calculation requires the inversion of a D by D matrix which in general has computational complexity $\mathcal{O}(D^3)$. When an observation is added to a component, for instance in a Gibbs sampling step, the sufficient statistics likelihood needs to be updated and thus requires recalculating the inversion of the matrix. Using the Cholesky factorization this can be done in $\mathcal{O}(D^2)$ but this complexity still renders the mixture model based on the $\mathcal{NIW}$ intractable for problems of the scale of whole-brain parcellation. Furthermore, there are $D(D+1)/2$ parameters in the covariance matrix that are to be determined for each cluster. Even though these parameters are somewhat regularized by the use of the $\mathcal{IW}$ prior it might still be an issue for high dimensional data with few observations per cluster.

The generative model for the mixture of Gaussians with normal inverse Wishart prior, in this thesis denoted GMM, is summarized in the following generative process:

Full covariance Gaussian Mixture model (GMM)

Prior	$p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \mathcal{NIW}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \gamma, \nu)$
Likelihood	$p(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

Normal Inverse Gamma distribution

For a univariate Gaussian distribution, a conjugate prior is given by the normal inverse gamma distribution and this construction can also be used as a prior for a multivariate Gaussian distribution with diagonal covariance matrix:

$$p(\boldsymbol{\mu}, \boldsymbol{\sigma}^2 \mid \boldsymbol{\mu}_0, \gamma, a, b) = \mathcal{N}(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, \gamma \cdot \operatorname{diag}(\boldsymbol{\sigma}^2)) \prod_{d=1}^{D} \mathcal{IG}(\sigma_d^2 \mid a, b), \tag{3.43}$$

where $\mathcal{IG}$ is the inverse gamma distribution given by

$$\mathcal{IG}(\sigma^2 \mid a, b) = \frac{b^a}{\Gamma(a)} (\sigma^2)^{-(a+1)} \exp(-\frac{b}{\sigma^2}), \tag{3.44}$$

where Γ is the univariate Gamma function. This is clearly not as flexible a prior as the $\mathcal{NIW}$ prior but it is no longer necessary to compute the inversion of a full rank matrix for the marginal likelihood. The marginal likelihood is then given by

$$p(\boldsymbol{x}_{1:N} \mid \boldsymbol{\mu}_0, \gamma, a, b) = \tag{3.45}$$
$$\prod_{d=1}^{D} \frac{(\gamma/(N+\gamma))^{1/2} a^b \Gamma(N/2+b)}{\Gamma(b)(2\pi)^{N/2} \left[\frac{1}{2}(\bar{\sigma}_d^2 + 2\boldsymbol{\gamma}) + \frac{1}{2}\gamma\boldsymbol{\mu}_{0,d}^2 - \frac{1}{2(N+\gamma)}(\bar{\boldsymbol{x}}_d + \gamma\boldsymbol{\mu}_{0,d})^2\right]^{(N/2+b)}},$$

where $\bar{\sigma}_d^2$ is the d'th entry in $\bar{\boldsymbol{\sigma}}^2 = \sum_{i=1}^N \boldsymbol{x}_i^2$ and $\bar{\boldsymbol{x}} = \sum_{i=1}^N \boldsymbol{x}_i$. Mixture models based on this generative process has been used to cluster the striatum with a distance dependent version of CRP prior by Janssen et al. (2015). In this thesis, we denote the generative process GMMd:

Diagonal matrix covariance Gaussian Mixture model (GMMd)

Prior	$p(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2) = \mathcal{NIG}(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2 \mid \boldsymbol{\mu}_0, \gamma, a, b)$
Likelihood	$p(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

A further simplification can be made, if we assume that all the elements of the diagonal matrix $\text{diag}(\boldsymbol{\sigma}^2)$ are identical in which case the conjugate prior is

$$p(\boldsymbol{\mu}, \sigma^2 \mid \boldsymbol{\mu}_0, \gamma, a, b) = \mathcal{N}(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, \gamma\sigma^2\boldsymbol{I})\mathcal{IG}(\sigma^2 \mid a, b), \tag{3.46}$$

and with $\bar{\sigma}^2 = \sum_{i=1}^{N} = ||\boldsymbol{x}||^2$ this gives the following marginal likelihood:

$$p(\boldsymbol{x}_{1:N} \mid \boldsymbol{\mu}_0, \gamma, a, b) = \tag{3.47}$$

$$\frac{(\gamma/(N+\gamma))^{D/2} a^b \Gamma(ND/2+b)}{\Gamma(b)(2\pi)^{ND/2}[\frac{1}{2}(\bar{\sigma}^2 + 2\boldsymbol{\gamma}) + \frac{1}{2}\gamma||\boldsymbol{\mu}_0||^2 - \frac{1}{2(N+\gamma)}||\bar{\boldsymbol{x}} + \gamma\boldsymbol{\mu}_0||^2]^{(ND/2+b)}}.$$

We refer to the mixture model based on this construction as the GMMs model and it is summarized in the following generative process:

Spherical covariance Gaussian Mixture model (GMMs)

Prior	$p(\boldsymbol{\mu}_k, \sigma_k^2) = \mathcal{NIG}(\boldsymbol{\mu}_k, \sigma_k^2 \mid \boldsymbol{\mu}_0, \gamma, a, b)$
Likelihood	$p(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \sigma_k^2) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \sigma_k^2\boldsymbol{I})$

The three levels of complexity in three presented Gaussian mixture models is summarized in Figure 3.3 where the respective covariance structures are visualized from data generated from either a spherical, diagonal, or full covariance multivariate Gaussian distribution.

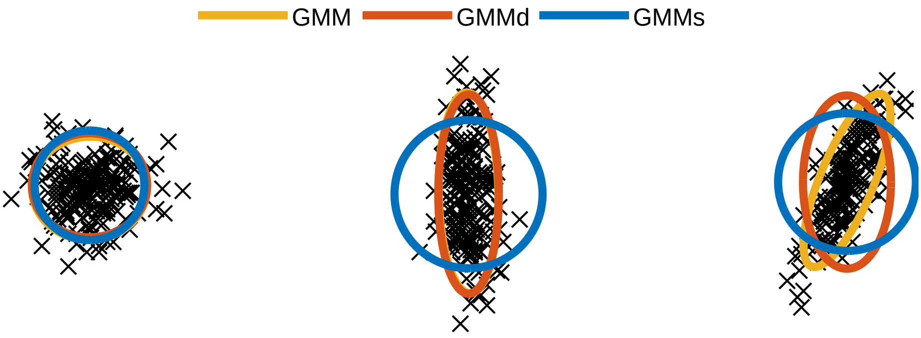


Figure 3.3: Illustration of the difference in flexibility of the three versions of the Gaussian mixture model.

Normal-Gaussian Process distribution

A *Gaussian process* is a collection of random variables any finite subset of which are jointly Gaussianly distributed and is specified by a mean function $m(x)$ and covariance function $\ker(x, x')$ (Rasmussen and Williams, 2006). A Gaussian process can be interpreted as a nonparametric prior over functions such that given a finite sampling of the domain, the function values are distributed according to the multivariate Gaussian distribution given by $\mathcal{N}(m(\boldsymbol{x}), \boldsymbol{\Sigma})$ where element $\{i, j\}$ of the covariance matrix is given by $\ker(\boldsymbol{x}_i, \boldsymbol{x}_j)$. The kernel function thus specifies how two points in the domain covariate and given the points in the domain, the Gaussian process is basically just a multivariate Gaussian distribution. Since we subtract the mean from the data before modeling we fix the Gaussian process with zero mean. We further introduce parameters for modeling the scaling of the noise and signal independently for each element. We denote the mixture model with a Gaussian process prior as the GMMGP and it is specified by the following generative process:

Gaussian mixture model with GP prior(GMMGP)

Prior	$p(\boldsymbol{\mu}_k) = \mathrm{GP}(\boldsymbol{\mu}_k, \mid \mathbf{0}, \boldsymbol{\Sigma})$
Likelihood	$p(\boldsymbol{x}_i \mid \boldsymbol{\mu}_k, \tau_k) = \mathcal{N}(\boldsymbol{x} \mid w_i \boldsymbol{\mu}_k, \sigma_i^2 \boldsymbol{I})$

In the model, we have further introduced two parameters for modeling the noise and scaling of signal independently for each observation. This enables the model to group together observations regardless of the magnitude of signal and noise and the model should therefore be able to learn a more compact representation of the data.

For our purpose the Gaussian process prior serves the purpose of an informed prior with which we can focus on the part of the signal that has the expected temporal dynamics. This is possible using the squared exponential covariance function which, for two observations $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ that are separated by τ in the input domain, is given by

$$\ker_{\mathrm{SE}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma_f^2 \exp\left(\frac{||\tau||^2}{2l^2}\right) \tag{3.48}$$

where l is the length-scale and σ_f^2 is the signal variance hyperparameters. Using this construction, we can use a full rank covariance matrix that is specified by only these two parameters. This construction does, however, break the possibility of analytically marginalizing the observation noise parameter. The marginal likelihood is given by

$$p(\boldsymbol{x}_{1:N}|\boldsymbol{\sigma}^2, \boldsymbol{w}, \boldsymbol{\Sigma}) = \int p(\boldsymbol{x}_{1:N}|\boldsymbol{\mu}, \boldsymbol{w}, \boldsymbol{\sigma}^2) p(\boldsymbol{\mu} \mid \boldsymbol{\Sigma}) d\boldsymbol{\mu} \tag{3.49}$$

$$= \sqrt{|\boldsymbol{\Sigma}|/|\boldsymbol{S}|} \exp\left\{\frac{1}{2}\bar{\boldsymbol{x}}^\top \boldsymbol{S}^{-1}\bar{\boldsymbol{x}}\right\} \prod_i^N (2\pi\sigma_i^2)^{-D/2} \exp\left\{-\frac{1}{2\sigma_i^2}\boldsymbol{x}_i^\top \boldsymbol{x}_i\right\},$$

where

$$\bar{\boldsymbol{x}} = \sum_{i=1}^N \frac{w_i}{\sigma_i^2}\boldsymbol{x}_i, \qquad \text{and} \qquad \boldsymbol{S} = \left(\boldsymbol{\Sigma}^{-1} + \sum_{\boldsymbol{z}(i)=k} \frac{w_i^2}{\sigma_i^2}\boldsymbol{I}\right). \tag{3.50}$$

The inversion of the full rank covariance matrix $\boldsymbol{S}$ can be done in linear time in D using the spectral decomposition of $\boldsymbol{\Sigma} = \boldsymbol{V}^\top \boldsymbol{D}\boldsymbol{V}$ as proposed in paper A.

We fix the length-scale of the squared exponential covariance function to that of the canonical hemodynamic response function such that the clustering procedure focuses on the part of the signal that is caused by the slowly varying level of oxygenation in the cerebral blood flow. Note that this has the further effect of regularizing the size of the clusters since the informed prior will penalize small clusters more strongly than the regular Gaussian mixture models.

We illustrate the effect of using a strong prior in the clustering model in Figure 3.4 where we present data with two different sources of temporal dynamics; observations contains a contribution from either the red and blue lines in the left plot and share a contribution from the more slowly varying black line. The iGMMGP sampler with a fixed length-scale at 1 distinguishes between the observations from the two nuisance sources but with a length-scale at 10 all observations are clustered together as seen in the middle and right panel of Figure 3.4.

This framework for clustering sequential data is very general and prior knowledge on the dependency in the sequence can be encoded by using different kinds of kernel functions. Popular covariance matrices include the Matérn covariance and periodic covariance functions.

A mixture model with a Gaussian process prior on the mean parameter was proposed by Lázaro-Gredilla et al. (2012) to model overlapping mixtures of Gaussian processes. The model was applied to the problem of tracking objects, and specifically for identifying ground-to-air missile trajectories. Later Ross and Dy (2013) augmented the model with constraints implemented by a Markov random field with sampling based inference to identify meaningful subtypes of lung disease in a longitudinal study.

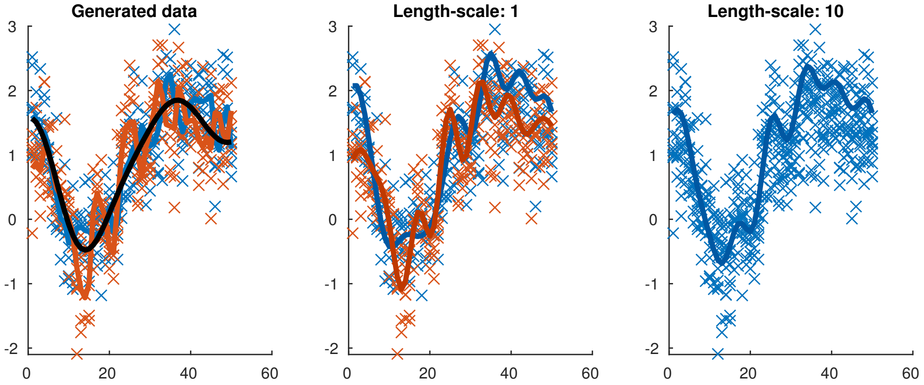


Figure 3.4: This figure illustrate the effect of using a Gaussian Process prior in clustering. In the left panel, we present the data generated; the blue crosses are generated iid. around the blue line, and the red cross around the red line. Both the blue and red line are generated by drawing from a Gaussian process with a squared exponential covariance with length-scale 1 and the black line as mean. The black line is drawn from a zero mean Gaussian Process with a squared exponential covariance and length-scale 10. In the middle panel is shown the clustering results from the infinite GMMGP model with a fixed length-scale at 1 and the model identifies two clusters. In the right panel the length-scale is fixed at 10 and identifies only one cluster.

Von-Mises Fisher distribution

A frequently used preprocessing step when working with fMRI data is normalizing the data using z-scoring such that the time-series for each voxel gets zero mean and standard deviation 1. This projects the data into the hypersphere with radius $\sqrt{D-1}$ and should therefore be modeled using directional statistics (Mardia and Jupp, 2009).

The von-Mises Fisher distribution is a distribution on the unit length hypersphere and is similar to the spherical Gaussian distribution. The von-Mises Fisher distribution is often parameterized by a mean vector $\boldsymbol{\mu} \in \mathcal{S}^D$ and a concentration parameter $\tau \in \mathrm{I\!R}_+$ and is given by

$$\mathrm{vMF}(\boldsymbol{x} \mid \tau, \boldsymbol{\mu}) = C_D(\tau) \exp(\tau \boldsymbol{\mu}^\top \boldsymbol{x}), \tag{3.51}$$

where $C_D(\tau) = \frac{\tau^{D/2-1}}{(2\pi)^{D/2}\mathcal{I}_{D/2-1}(\tau)}$ with $\mathcal{I}_\nu(x)$ the Bessel function of first kind of order ν and argument x. A mixture model based on von-Mises Fisher distributions was introduced in Banerjee et al. (2005) and recent variational inference approaches has been presented by Taghia et al. (2014) and Gopal and Yang (2014).

The von-Mises Fisher distribution is part of the exponential family and by Eq. (3.6) the conjugate prior is given by

$$p(\tau, \boldsymbol{\mu} \mid a, b, \boldsymbol{\mu}_0) \propto C_D(\tau)^a \exp\left\{\tau b \boldsymbol{\mu}^\top \boldsymbol{\mu}_0\right\}. \tag{3.52}$$

Note that the normalization constant is not available in closed form (Nunez-Antonio and Gutiérrez-Pena, 2005) and furthermore this construction does not allow the concentration parameter, τ, and the mean direction, $\boldsymbol{\mu}$, to be modeled independently. Therefore, it is beneficial to use the following prior that still allows the parameter $\boldsymbol{\mu}$ to be marginalized analytically

$$\begin{aligned} p(\tau, \boldsymbol{\mu} \mid \tau_0, \boldsymbol{\mu}_0, a, b) &\propto \mathrm{vMF}(\boldsymbol{\mu} \mid \tau_0, \boldsymbol{\mu}_0) \frac{C_D(\tau)^a}{C_D(b\tau)} \\ &= \mathrm{vMF}(\boldsymbol{\mu} \mid \tau_0, \boldsymbol{\mu}_0) f(\tau \mid a, b). \end{aligned} \tag{3.53}$$

Since the normalization constant of the conjugate prior for the vMF distribution is not available in closed form there is little computational difference between using the more flexible prior of Equation (3.53). Using this prior the marginal likelihood is given by

$$p(\boldsymbol{x}_{1:N} \mid \tau_0, \boldsymbol{\mu}_0, a, b) = \int \frac{C_D(\tau_0) C_D(\tau)^N}{C_D(\tau \| \boldsymbol{\mu}_0 + \sum_{i=1}^N \boldsymbol{x}_i \|)} p(\tau) d\tau, \tag{3.54}$$

where $p(\tau) \propto \frac{C_D(\tau)^a}{C_D(b\tau)}$ is the prior for τ. This integral is one dimensional and can easily be evaluated numerically, for instance using integration by MCMC. The generative process for the mixture of von-Mises Fisher distributions, denoted vMFmm, is summarized by the following generative process

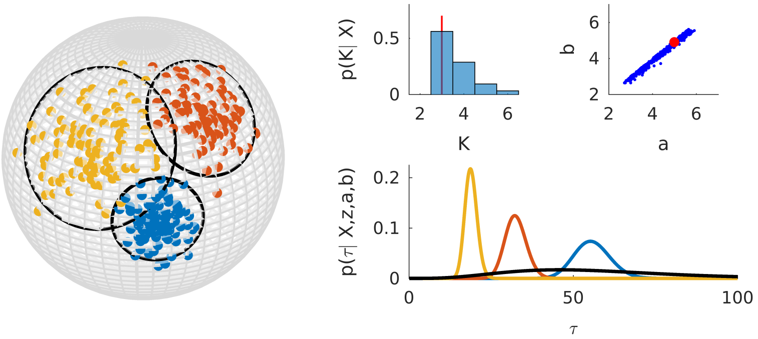


Figure 3.5: Illustration of the von-Mises Fisher model. On the left is the clustering of the highest likelihood sample along with 95% credibility regions denoted by the black circles around each cluster. On the right are the posterior marginal distribution for the number of clusters, the samples for the hyperparameters a and b with the highest likelihood sample marked by the red dot, and the posterior distributions for τ for each of the three clusters and the prior for τ in black for the highest likelihood sample.

von-Mises Fisher Mixture model (vMFmm)

Prior	$p(\boldsymbol{\mu}_k, \tau_k) = \mathrm{vMF}(\boldsymbol{\mu}_k, \mid \boldsymbol{\mu}_0, \tau_0) f(\tau_k \mid a, b)$
Likelihood	$p(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \tau_k) = \mathrm{vMF}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \tau_k)$

Mixture models based on the von-Mises Fisher distribution has been previously used in neuroimaging. Using normalized brain maps of Z-statistics for a number of different tasks was modeled by a mixture of von-Mises Fisher distributions by Lashkari et al. (2010); Lashkari and Golland (2009). Yeo et al. (2011) computed the functional connectivity of each voxel to 1000 ROIs spread across the brain and modeled the normalized FC vectors using a vMFmm dividing the brain into 14-34 parcels. Finally, Ryali et al. (2013) modeled the normalized voxel time series for several regions of the brain using mixture of vMF distributions and restricting the clustering to contiguous clusters using a Markov random field.

3.5.1 Effect of hyperparameter sampling

In Bayesian modeling, there is often a distinction between parameters and hyperparameters. Parameters control the distribution of the observed variables while hyperparameters control the distribution of the parameters. The hyperparameters in the mixture models used in this thesis are the parameters for the prior distributions, such as the set $\{\alpha, \boldsymbol{\mu}_0, \tau_0, a, b\}$ for the vMFmm. It is crucial in order to be able to infer the correct clustering that these hyperparameters learned as illustrated by Albers et al. (2016) for the infinite relational model.

In Figure 3.6 we illustrate the effect of sampling the hyperparameters in the vMFmm applied to the dataset generated for Figure 3.5. From the list of sampled hyperparameters, we select two sets of hyperparameters where the mode for the prior distribution of the concentration parameter is (1) too high, and (2) too low. As expected, the set (1) overestimates the number of clusters while the set (2) has a higher peak at the correct number of clusters compared to when the hyperparameters are learned. The inference chain with sampled hyperparameters are, with respect the adjusted mutual information with the true clustering, better than both of the selected sets of parameters.

As previously mentioned, Miller and Harrison (2013) argues that Dirichlet process mixtures overestimates the number of clusters. While, this can be partly addressed by inferring the hyperparameters, most importantly, the number of clusters in generated samples is not a very robust statistic of the data. In most samples from large models, some observations will be in singleton clusters, but these singletons might be eliminated if the parameters are optimized in the inference chain, i.e. running several iterations of the MCMC procedure where the optimal proposals are accepted.

3.5.2 Model summary

The presented models are summarized in Table 3.1. Note that for satisfyingly sample the distributions for whole-brain fMRI the sampling chains would require an exponential number of samples. Instead, we often use the highest likelihood sample from the inference chain and from this sample optimize the parameters to a local maximum. This means that the sampling procedure might better be compared with stochastic optimization. In Table 3.1 we include a summary of the implemented models with number of parameters that are estimated and time per MCMC iteration for a synthetically generated dataset.

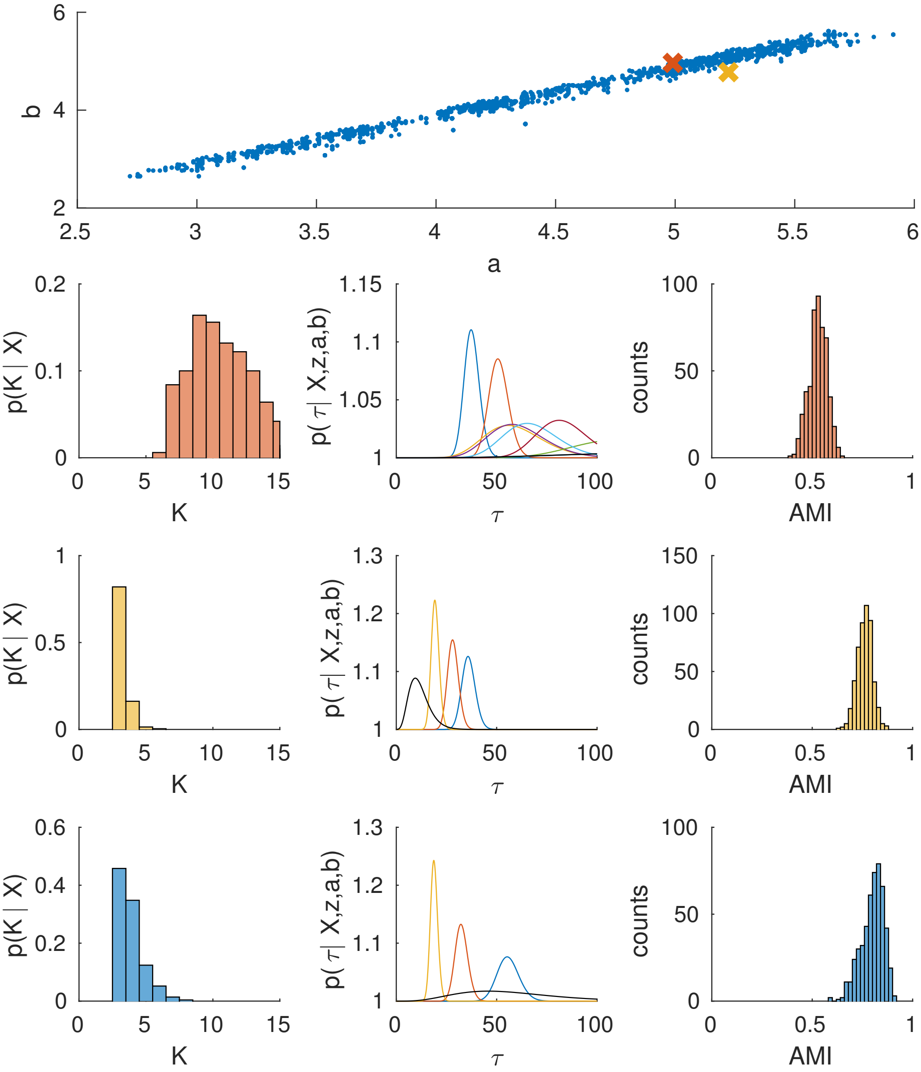


Figure 3.6: Illustration of the effect of sampling the hyperparameters. In the top panel we present the sampled sets of parameters (a,b) for the prior on the concentration parameter along with the two selected sets marked by the colored crosses. The second and third panel shows the distribution of the number of clusters and adjusted mutual information for hyperparameters for the set of parameters corresponding to the color of the histograms. In the bottom panel are the equivalent graphs for the inference chain where all hyperparameters are sampled.

Model	Inf. model	Complexity	Time/iter	# parameters
GMM	iGMM	$\mathcal{O}(SKND^2)$	-	$\mathcal{O}(SKD^2)$
GMMd	iGMMd	$\mathcal{O}(SKND)$	1min 18s	$\mathcal{O}(SKD)$
GMMs	iGMMs	$\mathcal{O}(SKND)$	26s	$\mathcal{O}(SKD)$
GMMGP	iGMMGP	$\mathcal{O}(SKND)$	1min 19s	$\mathcal{O}(SKND)$
vMFmm	iVMFmm	$\mathcal{O}(SKND)$	1min 3s	$\mathcal{O}(SKD)$

Table 3.1: Table summarizing the considered models, note that the models using the nonparametric CRP prior are preceded with an *i*. The time per MCMC iteration is for the nonparametric models and is reported for a 2.1 GHz core i7 laptop processor running `Matlab 2015b` on a synthetic dataset generated with parameters $N = 45000, K = 500, S = 1, D = 120$. The non-parametric version of each model was used and the models was initialized from the true clustering such that each model iterates over the same number of clusters.

Chapter 4

Implementation of sampling based inference in mixture models

In this chapter we present a tutorial on how to use the software tool for clustering using the mixture models presented in this thesis. The code is written in `Matlab` and both code and examples are available from `www.brainconnectivity.compute.dtu.dk` or `github.com/rasmusroege`.

Mixture model	Description	Inf. class name	Class name
GMMs	Spherical GMM	`igmmsmodel`	`gmmsmodel`
GMMd	axis aligned diagonal GMM	`igmmddmodel`	`gmmddmodel`
VMFmm	infinite von-Mises Fisher MM	`ivmfmodel`	`vmfmodel`
GMMGP	infinite GMM with GP prior	`igmmgpmodel`	`gmmgpmodel`

Table 4.1: The implemented mixture models.

4.1 Usage

This section demonstrates the usage of the mixture modeling framework. First, we generate a small dataset that is used in the clustering with S subjects consisting of N observations of dimension T in K clusters.

```
K=10;    N=100;    T=20;    S=3;
z=kron((1:K)',ones(N/K,1));   % Clustering vector, see Fig. 4.1
muk=randn(T,K,S);             % Generate cluster means
x=repmat({zeros(T,N)},S,1);   % cell array with observations
for s=1:S
  for k=1:K
    % generate observations for cluster k
    x{s}(:,z==k)=bsxfun(@plus,muk(:,k,s),randn(T,sum(z==k),1));
  end;
end
```

We can use the infinite spherical Gaussian Mixture model to cluster the generated dataset by the following code snippet:

```
z_init=randi(K,N,1);
m=igmmsmodel(x,z_init);
infsample(x,m);
```

The first line creates a random clustering configuration that is used for initializing the model. The next line creates a model object that initializes all parameters of the probabilistic model. The model thus represents one sample from the posterior distribution of the parameters of the mixture model given the dataset x. The model is used in the second line by the sampling function, **infsample**, that modifies the parameters of the model. Per default the inference procedure performs 3 iterations and the most recent parameters are stored in the model object. We can visually compare the true and inferred clustering by the following snippet that produces Figure 4.1

```
subplot(1,2,1);
bar(z); xlim([0 100]); title('True clustering');
subplot(1,2,2);
bar(m.par.z); xlim([0 100]); title('Inferred clustering');
```

It is possible to use any of the implemented models in both their finite and infinite implementations. The finite models must be initialized with a number of components, K, in addition to the data and the initial clustering:

```
imd=igmmdmodel(x,z); md=igmmdmodel(x,z,K);
```

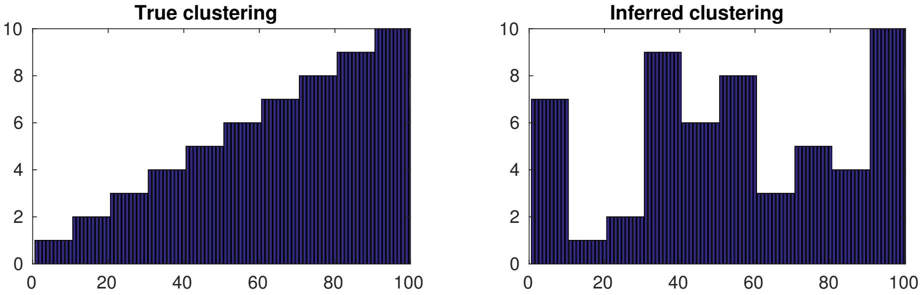


Figure 4.1: Visual comparison of the inferred and true clustering for the generated dataset. The observation indices is on the x-axis and the associated cluster labels are on the y-axis. Note, that the ordering of the cluster labels is random and the inferred clustering and the true clustering are therefore equivalent.

```
ivmfm=ivmfmodel(x,z); vmfm=vmfmodel(x,z,K);
igpmod=igmmgpmodel(x,z); gpmode=gmmgpmodel(x,z,K);
```

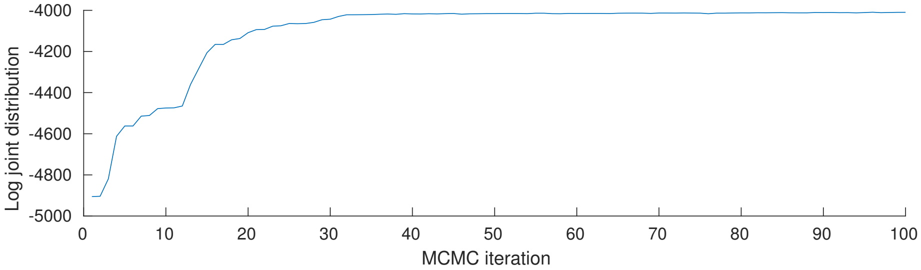


Figure 4.2: A convergence plot with the log joint distribution for the first MCMC 100 iterations.

The two models based on the von Mises-Fisher distribution require the input data, `x`, to be normalized, i.e. for each observation to have unit norm and zero mean. It is possible to pass options to the `infsample` function, and the function call returns a number of variables:

```
o=struct();
o.maxiter=100;
o.verbose=0;
[z,llh,noc,best_sample]=infsample(x,m,o);
```

`z` is the clustering configuration of the highest likelihood sample, `llh` is a vector of the log joint probability for each sample, `noc` is a vector of the number of clusters in each sample, and best_sample is the highest likelihood sample.

4.2 Features

The framework uses sampling based inference and most interesting quantities are therefore analyzed by collecting samples. This includes evaluating the posterior distribution for hyper parameters, marginal distribution for the number of clusters, and approximating the predictive likelihood. To illustrate how these quantities are computed we generate a test and training dataset:

```
K=10;    N=100;    T=20;    S=3;
z=kron((1:K)',ones(N/K,1));
muk_train=randn(T,K,S);
muk_test=randn(T,K,S);
x_train=repmat({zeros(T,N)},S,1);
x_test=repmat({zeros(T,N)},S,1);
for s=1:S
  for k=1:K
    sigk=gamrnd(4,1);
    % generate training dataset
    x_train{s}(:,z==k)=bsxfun(@plus,muk_train(:,k,s),sqrt(sigk)*
        randn(T,sum(z==k),1));
    sigk=gamrnd(4,1);
    % generate dataset for test
    x_test{s}(:,z==k)=bsxfun(@plus,muk_test(:,k,s),sqrt(sigk)*
        randn(T,sum(z==k),1));
  end
end
```

We then run the igmmsmodel discarding the first 50 for burnin and then collecting 50 samples with a thinning factor of 3. We store the samples in a cell arrray:

```
% initialize model
m=igmmsmodel(x,z);
o=struct();
o.maxiter=50;
% burnin
infsample(x_train,m,o);
nsamples=50;
thinning=3;
samples=cell(nsamples,1);
```

```
% inference
o.maxiter=thinning;
for iter=1:nsamples
  infsample(x,m,o);
  samples{iter}=copy(m);
end
```

We then approximate the predictive likelihood using the collected samples:

```
pred_llh=zeros(nsamples,1);
for i=1:nsamples
    tmp=samples{i};
    tmp.calcss(x_test);
    pred_llh(i)=tmp.llh;
end
log(sum(exp(pred_llh-max(pred_llh))))+max(pred_llh)-log(nsamples)
     % -7.2593e+03
```

We have further implemented a number of switches for using either the SAMS or split-merge sampler (for infinite models), using the speedup for the SAMS or split-merge sampler, switching off the printed status for each iteration, for using the last few (10 here) iterations to go to a local optimum, and finally to activate the debugging functionalities:

```
o.UseSequentialAllocation=1;  % SAMS / split-merge (default:1)
o.UseSMspeedup=1;             % (default:1)
o.maxiter=100;                % (default:3)
o.verbose=1;                  % (default:1)
o.optim=10;                   % (default:0)
o.debug=1;                    % (default:0)
```

The GMMGP models supports the options for heteroscedastic modeling of the noise and scaling of the signal. The choice for this option is passed as an additional argument when the object is constructed.

```
gmmgp_choice='hh';            % 'hh','hs','sh','ss'
m=gmmgpmodel(x,z_init,gmmgp_choice)
```

The different choices are described in the following table along with the expression for the likelihood.

`gmmgp_choice`	Heteroscedastic Signal	Heteroscedastic Noise	Likelihood
`'hh'`	Yes	Yes	$\boldsymbol{x}_{i,s} \sim \mathcal{N}(w_{i,s}\boldsymbol{\mu}_{z(i),s}, \sigma_{i,s}^2\boldsymbol{I})$
`'hs'`	Yes	No	$\boldsymbol{x}_{i,s} \sim \mathcal{N}(w_{s}\boldsymbol{\mu}_{z(i),s}, \sigma_{i,s}^2\boldsymbol{I})$
`'sh'`	No	Yes	$\boldsymbol{x}_{i,s} \sim \mathcal{N}(w_{i,s}\boldsymbol{\mu}_{z(i),s}, \sigma_{s}^2\boldsymbol{I})$
`'ss'`	No	No	$\boldsymbol{x}_{i,s} \sim \mathcal{N}(w_{s}\boldsymbol{\mu}_{z(i),s}, \sigma_{s}^2\boldsymbol{I})$

4.3 OOP implementation of mixture models

The inference of the mixture models is created in an object-oriented programming (OOP) framework where the sampling function, `infsample`, uses a mixture model object, such as the `gmmgpmodel`, to perform inference. The implementation of the mixture modeling framework has been developed under considerations for generalizability, code reuse, and easily understandable implementation but weighted against the computational performance. Implementing the inference procedure in a object oriented framework means that any object that implements a specific set of methods can be used for the sampling procedure. The possibility to reuse both the code for inference and debugging allows for fast and easy extension of the framework with additional mixture models as described in § 3.4.3.

The mixture model object implements either the `AbsFiniteModel` or `AbsInfiniteModel` abstract classes such that `infsample` knows whether the model is a finite mixture model or not. The model object must implement functions that allow `infsample` to use the model for Gibbs sampling and possibly split-merge or SAMS sampling. The model objects are responsible for sampling hyper parameters. The general methods that must be implemented are:

```
llh(obj)                        % evaluate the log joint prob.
copy(obj)                       % create a deep copy.
calcss(obj,x)                   % eval. sufficient statistics.
```

Note that the `obj` argument is a reference to the object itself and therefore only specifies that the method has access to the parameters of the object. The `llh` returns the log joint distribution for the current set of parameters, `copy` creates a deep copy of the object, which is necessary since the models extends the `handle Matlab` class. `calcss` computes the sufficient statistics of the model class object. The required methods for Gibbs sampling are

```
remove_observation(obj,x,n)       % remove obs. n
cat=compute_categorical(obj,x,n)  % compute categorical dist.
add_observation(obj,n,k)          % add obs. n to cluster k
remove_empty_clusters(obj)        % remove empty clusters
```

`remove_observation` removes observation `n` from the sufficient statistics.

`compute_categorical` returns the categorical distribution of assigning observation `n` to any of the possible clusters.

`add_observation` adds observation `n` to cluster `k`.

`remove_empty_clusters` removes any empty clusters in the sufficient statistics.

For the infinite models to support split-merge or SAMS sampling, the object needs to further implement the following two methods:

```
m_split=initLaunch(obj,x,z_t,c);  % initialize to launch config.
m_merge=initMerge(obj,x,z_m,c);   % initialize to merge config.
```

where the `initLaunch` method allows the model to efficiently compute the sufficient statistics of the launch state based on the current state of the model and knowledge of the new clustering assignment, `z_t`, and `c=[k1,k2]` where `k1` and `k2` are indices for the two clusters that are to be considered in the split proposal. The `initMerge` should similarly be an efficient method to update the sufficient statistics to the state whith clustering assignment `z_m` by merging the two components in `c`.

The objects themselves are responsible for hyper parameter sampling. The `infsample` request a cell array of function names using the `get_samplers` class method and then iteratively apply each of the methods using the `feval` matlab function.

```
for sampler=1:length(model.get_samplers)
    feval(str2func(model.get_samplers{sampler}),model,X,1);
end
```

For further documentation, we refer to the comments in the source available from `www.brainconnectivity.compute.dtu.dk` or `github.com/rasmusroege`.

Chapter 5

Research contributions

In the included papers, we follow the aim of developing and applying probabilistic models to parcellate whole-brain fMRI. We therefore apply the mixture models and inference procedures from Chapter 3 in the modeling framework described in Chapter 4. The mixture models of primary focus were the Gaussian mixture model with Gaussian Process prior (GMMGP) and the von Mises-Fisher mixture model (vMFmm). Note that all the code that has been used for these papers is available online.

5.1 Paper A, Unsupervised Segmentation of Task Activated Regions in fMRI

Traditional analysis of task fMRI is done using a general linear model to create a statistic parametric map of the contrasts of interest based on specific knowledge of the timing of the task conditions. This approach requires strong knowledge and assumptions on the BOLD response induced by the task conditions. Spatially smoothing the fMRI data is a standard preprocessing step that increases statistical power but at the cost of spatial resolution. We show that the regions of task activation can be identified unsupervised using the iGMMGP mixture model without smoothing the fMRI data.

We apply the iGMMGP mixture model with heteroscedastic noise on non-smooth fMRI finger tapping task data. We initialize the mixture model with all voxels in the same cluster and infer the required model complexity using Bayesian non-parametrics. We further propose a minor, but crucial, change to the split-merge MCMC sampling procedure, and we demonstrate that it causes a significant speedup on the problem of whole-brain parcellation. The inferred clusterings consists of approximately 100 clusters many of which consist of spatially contiguous regions and show right-left hemisphere symmetry. The clusters in the motor cortex have a component in the cerebellum in accordance with expectations. We use cross-subject correlation of the parcel time series and consistency of the inferred clusters across sampling chains to unsupervised identify regions of task activation. We find that clusters with high cross subject correlation of time-series are also consistently delineated and that the two clusters with the highest cross-subject correlation corresponds well to the regions of task activation from a supervised SPM group analysis. We further find interesting regions that cannot be identified by the SPM analysis because they are not sufficiently correlated with the design matrix.

In summary, the method shows great promise for clustering fMRI time series data and the framework of unsupervised extraction of task activated regions is promising, especially for fMRI experiments where strong knowledge of how the task induces a BOLD response is missing.

5.2 Paper B, Infinite von Mises-Fisher Mixture Modeling of Whole-Brain fMRI Data

Cluster modeling of fMRI time-series data is often performed using Gaussian mixture models or non-probabilistic clustering algorithms on *z-scored* fMRI data. Z-scoring projects the time series with D brain volumes of each voxel to the hyper sphere with radius $\sqrt{D-1}$. Modeling data on a hypersphere is properly done using directional statistics since there is no longer any information in the magnitude of the voxel time series. This fact is most often ignored when clustering fMRI time series.

Paper B analyses the effect of modeling directional data with the von Mises-Fisher distribution in contrast to traditional Gaussian distributions and introduces a von Mises-Fisher based mixture modeling framework that scales to the problem of whole-brain fMRI time series modeling. Both types of mixture models are implemented with two different priors on the clustering; one using the Dirichlet-multinomial distribution and the other using the Chinese restaurant process allowing for the model complexity to be adapted to the data. We employ identical methods for inference and initialization for all the considered clustering models, the GMMs, GMMd and vMFmm. The paper presents a thorough analysis of the effect of the numerical approximation required for the vMF based models to be tractable and initialization strategies for fast convergence. The sampling based inference is compared to an approximation based on variational inference on a previously studies topic modeling dataset and found to be at least on par. The von Mises-Fisher and Gaussian based mixture models are applied to a synthetic dataset generated from a mixture of vMF distributions and to a 29-subject resting state fMRI dataset normalized to the unit hyper sphere.

The inferred clusterings are evaluated based on three measures of reliability of clusterings across groups of subjects; normalized mutual information, adjusted mutual information, and the adjusted rand index. On the synthetic data, we find that both the Gaussian and vMF based finite mixture models show very similar reliability performance. The infinite Gaussian based mixture models do, however, tend to overestimate the number of clusters. On the rs-fMRI dataset we analyze, we ran the finite mixture models with the number of clusters in the range from 50 to 1000 and the vMF based mixture model showed consistently more reliable results for all settings compared to both Gaussian models. While the infinite vMF based mixture model found more clusters compared to the Gaussian mixture model, the inferred clustering solutions were more reliable.

The results from clustering analysis indicate that it is possible to increase the reliability of clustering by modeling the normalized fMRI time-series using directional statistics.

5.3 Paper C, Functional Whole-Brain Parcellation using Bayesian Non-Parametric Modeling

In paper C we explore the strategy of constructing a probabilistic mixture model that is sufficiently expressive to account for the variability in fMRI. We extend the model from paper A to also model the scaling of the signal across the brain and we show that the iGMMGP method finds a compact representation of the data and that the extracted parcellation is more reliable when compared to both nonprobabilistic clustering methods and Gaussian mixture models.

We perform a thorough comparison with several traditional clustering models that have been applied to the problem of whole-brain parcellation including clustering using Ward's algorithm, Normalized cut, and clustering based on region growing. We compare the clustering methods on data generated with spatially dependent noise and scaling of the signal and find that the non-parametric GMMGP model can recover both the number of clusters and the clustering configuration more reliably without being informed of neighbors or the spatial location of the synthetic voxels. On fMRI data, we find that the use of voxel specific parameters improves on the predictive performance while allowing for a more compact representation of the data on both a resting state and a finger tapping task dataset. Comparing the reliability in the inferred clustering across groups of subjects, the proposed method is consistently more reliable compared to both the traditional Gaussian based methods and the non-probabilistic methods.

We divide the task and resting state datasets into test and training datasets by splitting each time series in two, and use the iGMMGP model to independently perform a parcellation on the two splits. On the task fMRI dataset, the cluster based SPM analysis is more reliable and with higher sensitivity infers the regions of task activation compared to a traditional voxel based SPM analysis. On the resting state dataset, we select two regions and again compare the functional connections inferred by a cluster based SPM analysis to a voxel based analysis using the averaged region time series as regressors. Again, we see that the cluster based analysis more reliably identifies regions of functional connectivity.

The introduced probabilistic clustering model is able to more reliably perform whole-brain parcellation compared to both probabilistic and traditional clustering models. Based on a predictive analysis there is support for the use of heteroscedastic modeling of noise and signal scaling. Finally, the increased reliability of identified regions of task activation and the functional networks means that the clustering method is a promising tool for preprocessing fMRI data for more reliable analyses.

Chapter 6

Discussion and Conclusion

The aim of this thesis was to develop efficient probabilistic methods for whole-brain parcellation that incorporate domain knowledge in the statistical distributions used to model the fMRI time series. Using Bayesian non-parametrics, the models can adapt to the complexity of the data. The probabilistic models investigated in this thesis in general outperform the tested non-probabilistic clustering models with respect to reliability of inferred clusterings across groups of subjects.

The proposed framework for unsupervised identification of task activated regions in paper A successfully identified regions of task activation in a very well understood dataset. A natural follow up question is to investigate if this finding generalizes to other and more difficult task datasets such as datasets where subjects are exposed to natural stimuli such as listening to an auditory feature film (Hanke et al., 2014).

The von Mises-Fisher and the Gaussian mixture model with a Gaussian process prior represent two different approaches to clustering fMRI time series data as presented in papers B and C. Both models find reliable parcellations, but there are significant differences between the two models: While the vMFmm models the directional standardized fMRI data, it also optimally infers regions that are homogenous as measured by correlation. Since correlation is frequently used both as a homogeneity measure for evaluating parcellation methods as well as for further analysis of fMRI data, this is a promising feature of the model. The GMMGP model has additional advantages: One effect of the strong prior on the parcel time series is that small clusters are penalized and in combination with the heteroscedastic modeling of signal and noise, the nonparametric Bayesian modeling thus provide a more compact representation of the data.

For most results on whole-brain analysis in the presented paper, the inference chains were limited to approximately 100 MCMC samples. This is absolutely not enough to satisfactorily explore the posterior distribution. Chance is that it is not even enough for sufficient burnin. Both variational inference based methods and expectation maximization escape this problem by approximating the distribution but are highly susceptible of being stuck in local modes of the posterior distribution. MCMC sampling can also get stuck in local modes but can break free given enough samples. The problem of inference in mixture models is thus a weighting of vulnerability to local modes and not reaching the equilibrium distribution for sampling. Ultimately the methods should be evaluated based on how well the inferred clusterings perform in practical settings and we should therefore not let the perfect be the enemy of good.

The same philosophy applies to the use of the Bayesian nonparametrics. Since, we cannot guarantee that the sampling chains have reached the equilibrium distributions we cannot draw firm conclusions on the number of clusters in the data. We do, however, know that the split-merge and SAMS sampling procedures, that the CRP prior enables, greatly improves the rate of convergence (Dahl, 2005; Albers et al., 2013). Therefore, while the number of clusters inferred by the nonparametric models might not a good estimate for the true number of clusters in the data, it will be a better estimate than what would be achieved by pruning a parametric model. This advantage in the sampling procedures could perhaps be transferred to a finite model; if we, for instance, are interested in a 500-cluster brain parcellation we could initialize a nonparametric model with 100 clusters and allow this model to populate the 500 clusters. When these 500 clusters are populated inference could continue with a parametric model, initialized from the nonparametric solution, that could be used for further fine tuning of the clustering configuration.

Bayesian probabilistic models allow for using the predictive likelihood on hold-out data to assess the ability of a model to adequately fit the data without overfitting. We use the predictive likelihood to show that there is support for heteroscedastic modeling of the parameters for noise and scaling of signal in paper C. Optimally, we would like to perform comparisons between the classes of probabilistic models considered in this thesis. The Gaussian process in the iGMMGP model constitutes a strongly informed prior that prevents the model from fitting to high frequency fluctuations in the data. Since these fluctuations might be caused by physiological noise they will also be apparent in the test dataset and the model will thus be handicapped in terms of the predictive likelihood. The von Mises-Fisher based mixture models are only defined for data residing on the unit hypersphere. This is ensured by normalizing the time series data, but also entails that we cannot use the predictive likelihood to compare the vMF and Gaussian based models.

In paper C we validate the inferred clustering against the dice overlap of inferred regions of task activation and functionally connected regions. Many fMRI studies are based on either identification of task activated regions or based on network analysis. It is therefore important to increase the reliability and sensitivity of these analyses and we show that this could potentially be done using a parcel based analysis. Shifting the validation of parcellations from reliability and homogeneity towards validation on the analyses that fMRI data is actually used to facilitate these benefits. Therefore, we believe that it would be interesting to extend the presented method to a formalized framework for comparing and evaluating clusterings.

The fMRI datasets analyzed in papers B and C are preprocessed using spatial smoothing. This reduces the spatial resolution of the data but there are considerable difficulties for the models to handle the level of noise in the fMRI data without spatial smoothing. Recent approaches use the data from several hundred subjects for a whole-brain parcellation of non-smoothed fMRI data (Glasser et al., 2016a). This approach would be feasible for brain parcellation by extending the developed framework to exploit parallel computation. The sampling procedure which constitutes the computational bottleneck trivially parallelize across the number of clusters and subjects.

Several traditional methods of clustering fMRI data are based on hierarchical clustering methods. This is also possible using probabilistic methods such as Bayesian hierarchical clustering (Heller and Ghahramani, 2005) or the Bayesian rose trees algorithm (Blundell et al., 2012) for multifurcating hierarchical clustering. Building on the experience using mixture models to model fMRI data it might be beneficial to apply these algorithms to the problem of whole-brain fMRI parcellation in the future.

The research contributions within the thesis show that the use of validated probabilistic clustering methods are beneficial for modeling fMRI data. These can potentially have wide applications in neuroimaging studies and can serve to improve sensitivity, robustness and interpretability of fMRI based studies.

Appendix A

Unsupervised Segmentation of Task Activated Regions in fMRI

Røge, R. E., Madsen, K. H., Schmidt, M. N., Mørup, M. (2015), ‘Unsupervised Segmentation of Task Activated Regions in fMRI’. 2015 IEEE International Workshop on Machine Learning for Signal Processing.

UNSUPERVISED SEGMENTATION OF TASK ACTIVATED REGIONS IN FMRI

Rasmus E. Røge[1], *Kristoffer H. Madsen*[2], *Mikkel N. Schmidt*[1], *Morten Mørup*[1]

[1]Technical University of Denmark
Lyngby, Denmark
{rasr,mnsc,mmor}@dtu.dk

[2]Danish Research Centre for Magnetic Resonance
Hvidovre, Denmark
stoffer@drcmr.dk

ABSTRACT

Functional Magnetic Resonance Imaging has become a central measuring modality to quantify functional activiation of the brain in both task and rest. Most analysis used to quantify functional activation requires supervised approaches as employed in statistical parametric mapping (SPM) to extract maps of task induced functional activations. This requires strong knowledge and assumptions on the BOLD response as a function of activitation while smoothing in general enhances the statistical power but at the cost of spatial resolution. We propose a fully unsupervised approach for the extraction of task activated functional units in multi-subject fMRI data that exploits that regions of task activation are consistent across subjects and can be more reliably inferred than regions that are not activated. We develop a non-parametric Gaussian mixture model that apriori assumes activations are smooth using a Gaussian Process prior while assuming the segmented functional maps are the same across subjects but having individual time-courses and noise variances. To improve inference we propose an enhanced split-merge procedure. We find that our approach well extracts the induced activity of a finger tapping fMRI paradigm with maps that well corresponds to a supervised group SPM analysis. We further find interesting regions that are not activated time locked to the paradigm. Demonstrating that we in a fully unsupervised manner are able to extract the task-induced activations forms a promising framework for the analysis of task fMRI and resting-state data in general where strong knowledge of how the task induces a BOLD response is missing.

***Index Terms*—** Functional connectivity, Gaussian Mixture Model, fMRI analysis

1. INTRODUCTION

Functional Magnetic Resonance Imaging (fMRI) allows for the identification of task related brain activations by measuring the blood oxygenation level dependent (BOLD) response in each voxel of the brain. The reliable identification of these regions of activations poses a major challenge due to a massive multiple comparisons problem and due to the low level of signal to noise found in fMRI data. Traditionally, fMRI data is spatially smoothened and then analysed using statistical parametric mapping (SPM) [1] in order to identify brain regions that are significantly correlated with the expected activation time course. These univariate voxel specific tests have to be corrected for multiple comparisons as fMRI data is high-dimensional i.e. in the order of 10^5 voxels. This is commonly handled by methods for correcting for the family-wise error based on Gaussian Random Fields [2] or correcting for false discoveries [3].

An alternative procedure is to cluster the fMRI time-series [4, 5] and carry out the statistical analysis at the level of clusters. Commonly adopted approaches have here been based on hierarchical and k-means clustering [4] and the cluster based algorithm (CBA) proposed in [5]. Thereby the number of statistical tests reduces to the number of clusters extracted while spatial smoothing no longer is necessary as the signal to noise ratio (SNR) is improved when considering the time series of each cluster centroid [5]. As opposed to segmenting the brain into clusters, independent component analysis (ICA) [6] is a widely applied approach to identify maps of task [6] and resting state [7] activations typically assuming spatial independence of the extracted components. In these approaches activations have been established using supervised evaluation of the extracted time courses [8].

In this paper we focus on a fully unsupervised approach for functional segmentation of task related activity using clustering based on a non-parametric mixture model tailored to the analysis of multi-subject fMRI data. Being non-parametric our model is able to automatically learn from data the number of clusters. It assumes subjects are normalized into a common space such that the extracted functional units are consistent across subjects with subject specific cluster time-series apriori assumed smooth by imposing a Gaussian process prior. To account for inhomogenous noise and misaligments a separate noise parameter is estimated individually for each subject and voxel. Inference in the model is accomplished by Markov-chain Monte Carlo sampling where we propose a new efficient procedure for split-merge sampling [9] that significantly reduces the computations of the in general most occuring merge operations.

For the unsupervised extraction of task-related clusters we evaluate how correlated the cluster time-series are across subjects as well as the stability of the clusters using evidence accumulation [10, 11] hypothesizing that these consistent actitvations correspond to task induced activity. We show that our unsupervised multi-subject analysis extracts the regions expected to be activated in a finger tapping paradigm and that the maps are similar to those extracted using a standard supervised group SPM analysis. Our method further circumvents smoothing as a necessary preprocessing step.

2. METHODS

2.1. The Infinite Gaussian Mixture model with a Gaussian Process prior (IGMMGP)

Let $\boldsymbol{X}_s$ be the $N \times T$ matrix of the N voxels with a time course with T measurements of subject s. We use $\boldsymbol{z}$ as a vector for the group assignment, such that $\boldsymbol{z}(i) = k$ if the i'th voxel is assigned to the cluster k. We thus assume that the voxels are aligned across subjects and that the clustering is shared. Our model is illustrated as

This work was supported by the Lundbeck Foundation, grant no. R105-9813

978-1-4673-7454-5/15/$31.00 ©2015 IEEE

a directed graphical model in Fig. 1 and is described by the following generative process where we use the Chinese Restaurant Process (CRP) [12, 13] as a prior for partitioning the voxels into clusters.:

$$\boldsymbol{z} \sim CRP(\gamma) \qquad \text{groups,} \qquad (1)$$

$$\boldsymbol{\mu}_{k,s} \sim GP(0, \boldsymbol{\Sigma}) \qquad \text{group time series,} \qquad (2)$$

$$\boldsymbol{x}_{i,s} \sim \mathcal{N}(\boldsymbol{\mu}_{z(i),s}, \sigma^2_{i,s}\boldsymbol{I}) \qquad \text{voxel time series,} \qquad (3)$$

where $\boldsymbol{\Sigma}$ is the covariance structure encoding the imposed temporal dynamics, $\boldsymbol{\mu}_{k,s}$ is the time series of cluster k of subject s, $\sigma^2_{i,s}$ is the variance of voxel i of subject s, and $\boldsymbol{x}_{i,s}$ is the time series of voxel i of subject s. The model we propose is an extension of the Infinite Gaussian Mixture Model [14] in which temporal dynamics are imposed on the mixtures through the Gaussian Process (GP) prior with covariance function $\boldsymbol{\Sigma}$. A Gaussian Process prior has previously been considered in the Infinite Gaussian Mixture Models also imposing Markov Random Field constraints [15]. A benefit of our model is that it includes voxel and subject specific noise that can potentiallty account for misalignment across subjects as well as spatially varying noise levels. In the following we call our model the Infinite Gaussian Mixture model with a Gaussian Process prior (IG-MMGP).

We define by $\{\boldsymbol{X}_s\}$, $\{\boldsymbol{\mu}_s\}$, and $\{\boldsymbol{\sigma}^2_s\}$ the collections of all subjects voxel time series, group time series and voxel variances respectively. With $\boldsymbol{\sigma}^2_s$ we denote the noise variance for all voxels for subject s, and with $\boldsymbol{x}_{\boldsymbol{z}(i)=k,s} = \{\boldsymbol{x}_{i,s} \mid \boldsymbol{z}(i) = k\}$ we denote the time series of all voxels assigned to cluster k of subject s.

According to the generative model the joint distribution of data and parameters can be expressed as

$$p(\boldsymbol{z}, \{\boldsymbol{\mu}_s\}, \{\boldsymbol{X}_s\} \mid \{\boldsymbol{\sigma}^2_s\}, \boldsymbol{\Sigma}, \gamma) \qquad (4)$$
$$= \left[\prod_s p(X_s|\boldsymbol{\mu}_s, \boldsymbol{\sigma}^2_s) p(\boldsymbol{\mu}_s|\boldsymbol{\Sigma})\right] p(\boldsymbol{z}|\gamma).$$

Due to conjugacy it is possible to analytically integrate out the group time series from the joint distribution

$$p(\boldsymbol{z}, \{\boldsymbol{X}_s\} \mid \{\boldsymbol{\sigma}^2_s\}, \boldsymbol{\Sigma}, \gamma) \qquad (5)$$
$$= \int \left[\prod_s p(X_s|\boldsymbol{\mu}_s, \boldsymbol{\sigma}^2_s) p(\boldsymbol{\mu}_s|\boldsymbol{\Sigma}) d\boldsymbol{\mu}_s\right] p(\boldsymbol{z}|\gamma))$$
$$= \frac{\Gamma(\gamma)\gamma^K \prod_{k=1}^K \Gamma(n_k)}{\Gamma(N+\gamma)} \prod_{i,s} (2\pi\sigma^2_{i,s})^{-T/2} \exp\left\{-\frac{1}{2\sigma^2_{i,s}} \boldsymbol{x}^\top_{i,s}\boldsymbol{x}_{i,s}\right\}$$
$$\prod_{k=1}^K (|\boldsymbol{S}_{k,s}|/|\boldsymbol{\Sigma}|)^{-1/2} \exp\left\{\frac{1}{2}\sum_{k=1}^K \bar{\boldsymbol{x}}^\top_{k,s}\boldsymbol{S}^{-1}_{k,s}\bar{\boldsymbol{x}}_{k,s}\right\},$$

where

$$\bar{\boldsymbol{x}}_{k,s} = \sum_{Z(i)=k} \frac{1}{\sigma^2_i}\boldsymbol{x}_{i,s}, \quad \boldsymbol{S}_{k,s} = \left(\boldsymbol{\Sigma}^{-1} + \sum_{\boldsymbol{z}(i)=k} \frac{1}{\sigma^2_{i,s}}\boldsymbol{I}\right),$$

and n_k is the number of voxels assigned to cluster k with N being the total number of voxels.

2.2. Model inference and accellerated merge steps

For inference of the clustering $\boldsymbol{z}$ we use Gibbs sampling with split-merge moves [9]. In each Gibbs move a voxel can be placed in any of the currently occupied K clusters or create a new. This means that the expression $\bar{\boldsymbol{x}}^\top_{k,s}\boldsymbol{S}^{-1}_{k,s}\bar{\boldsymbol{x}}_{k,s}$ has to be evaluated $K+1$ times. Let $\boldsymbol{\Sigma} = \boldsymbol{V}^\top\boldsymbol{D}\boldsymbol{V}$ be the eigendecomposition of $\boldsymbol{\Sigma}$. Then

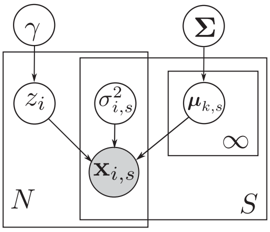


Fig. 1. Directed graphical model representation of the generative process.

$$\bar{\boldsymbol{x}}^\top_{k,s}\boldsymbol{S}^{-1}_{k,s}\bar{\boldsymbol{x}}_{k,s} = \bar{\boldsymbol{x}}^\top_{k,s}\left((\boldsymbol{V}^\top\boldsymbol{D}\boldsymbol{V})^{-1} + \sum_{\boldsymbol{z}(i)=k}\frac{1}{\sigma^2_{i,s}}\boldsymbol{I}\right)^{-1}\bar{\boldsymbol{x}}_{k,s}$$
$$= (\boldsymbol{V}\bar{\boldsymbol{x}}_{k,s})^\top(\boldsymbol{D}^{-1} + \sum_{\boldsymbol{z}(i)=k}\frac{1}{\sigma^2_{i,s}}\boldsymbol{I})^{-1}\boldsymbol{V}\bar{\boldsymbol{x}}_{k,s}.$$

Keeping $\boldsymbol{V}\bar{\boldsymbol{x}}_{k,s}$ and $(\boldsymbol{D}^{-1} + \sum_{\boldsymbol{z}(i)=k}\frac{1}{\sigma^2_{i,s}}\boldsymbol{I})$ in memory reduces the computational complexity of evaluating $\bar{\boldsymbol{x}}^\top_{k,s}\boldsymbol{S}^{-1}_{k,s}\bar{\boldsymbol{x}}_{k,s}$ to $O(T)$ when the projections $V\{\boldsymbol{X}\}$ and eigenvalues $\boldsymbol{D}$ have been computed. This reduces the total time complexity of a Gibbs sweep to $O(SNKT)$.

In the split-merge procedure two nodes, or voxels as denoted in this paper, are randomly sampled. If they are in the same group the group is proposed split and if they are in two different groups these two groups are proposed merged. The split/merge move is accepted according to the Metropolis-Hastings ratio

$$\alpha(\boldsymbol{z}^* \mid \boldsymbol{z}) = \min\left[1, \frac{p(\boldsymbol{z}^*, \{\boldsymbol{X}_s\}|\gamma, \{\boldsymbol{\sigma}^2_s\}, \boldsymbol{\Sigma})q(\boldsymbol{z}|\boldsymbol{z}^*)}{p(\boldsymbol{z}, \{\boldsymbol{X}_s\}|\gamma, \{\boldsymbol{\sigma}^2_s\}, \boldsymbol{\Sigma})q(\boldsymbol{z}^*|\boldsymbol{z})}\right], \qquad (6)$$

where the transition probability $q(\boldsymbol{z}^{split}|\boldsymbol{z}^{merge})$ for splitting a cluster is calculated using Gibbs sampling restricted to the observations influenced by the move. The transition probability is calculated by first sampling a so-called launch state and keeping track of the transition probabilities from this state to the final split configuration. As the merge move is deterministic we have for the transition probability $q(\boldsymbol{z}^{merge}|\boldsymbol{z}^{split}) = 1$, see also [9].

Provided no group contains more than 50 % of the observations there will be more merge than split proposals. As merge moves are deterministic we further have for these moves that $\frac{q(\boldsymbol{z}|\boldsymbol{z}^*)}{q(\boldsymbol{z}^*|\boldsymbol{z})} \leq 1$. We can thus significantly accelerate the evaulation of these proposals if we are able to reject the move by the ratio of the joint probabilities alone, i.e. $\frac{p(\boldsymbol{z}^*, \{\boldsymbol{X}_s\}|\gamma, \{\boldsymbol{\sigma}^2_s\}, \boldsymbol{\Sigma})}{p(\boldsymbol{z}, \{\boldsymbol{X}_s\}|\gamma, \{\boldsymbol{\sigma}^2_s\}, \boldsymbol{\Sigma})}$ thereby circumventing the more computationally demanding restricted Gibbs sweeps. This leads us to propose the following accelerated merge procedure: Before computing the launch state and final configuration using restricted Gibbs sampling compute the preliminary acceptance probability for a merge step

$$\alpha_1(\boldsymbol{z}^* \mid \boldsymbol{z}) = \min\left[1, \frac{p(\boldsymbol{z}^*, \{\boldsymbol{X}_s\}|\gamma, \{\boldsymbol{\sigma}^2_s\}, \boldsymbol{\Sigma})}{p(\boldsymbol{z}, \{\boldsymbol{X}_s\}|\gamma, \{\boldsymbol{\sigma}^2_s\}, \boldsymbol{\Sigma})}\right]. \qquad (7)$$

In case we cannot accept the proposal based on α_1 we will not be able to accept it based on α as $\alpha(\boldsymbol{z}^*|\boldsymbol{z}) \leq \alpha_1(\boldsymbol{z}^*|\boldsymbol{z})$.

In order to infer the hyperparameters γ and $\{\boldsymbol{\sigma}^2\}$ we impose the non-informative and improper prior $p(\theta) \propto \theta^{-1}$. We use Metropolis-Hastings sampling by transforming the variable to the log-domain and use the symmetric normal distribution as proposal density. Using the eigendecomposition of $\boldsymbol{\Sigma}$ and $\boldsymbol{V}\boldsymbol{x}_{k,s}$ the cost of evaluating the joint density ratio of the Metropolis-Hastings ratio is $O(T)$. Therefore a sweep of evaluating proposals of $\{\boldsymbol{\sigma}^2\}$ is $O(SNT)$, i.e. this inference step is less computationally demanding than the Gibbs sweep.

2.3. Unsupervised extraction of consistent clusters

The posterior of the cluster time series given the data and clustering can be calculated using Bayes theorem:

$$\begin{aligned} & p(\boldsymbol{\mu}_{k,s}|\boldsymbol{x}_{\boldsymbol{z}(i)=k,s}, \boldsymbol{\sigma}^2_{\boldsymbol{z}(i)=k,s}, \boldsymbol{\Sigma}) \qquad (8) \\ & = \frac{p(\boldsymbol{x}_{\boldsymbol{z}(i)=k} \mid \boldsymbol{\mu}_{k,s}, \boldsymbol{\sigma}^2_{\boldsymbol{z}(i)=k,s}) p(\boldsymbol{\mu}_{k,s} \mid \boldsymbol{\Sigma})}{\int p(\boldsymbol{x}_{\boldsymbol{z}(i)=k} \mid \boldsymbol{\mu}_{k,s}, \boldsymbol{\sigma}^2_{\boldsymbol{z}(i)=k,s}) p(\boldsymbol{\mu}_{k,s} \mid \boldsymbol{\Sigma}) d\boldsymbol{\mu}_{k,s}} \\ & = \mathcal{N}(\boldsymbol{S}_k^{-1}\bar{\boldsymbol{x}}_k, \boldsymbol{S}_k). \end{aligned}$$

To un-supervised select clusters of relevance for the task we evaluate the consistency of the cluster time-series as well as the consistency of the clustering across separate chains based on the sample with highest value of the joint-distribution $p(\boldsymbol{z}, \{\boldsymbol{X}_s\} \mid \{\boldsymbol{\sigma}^2_s\}, \boldsymbol{\Sigma}, \gamma)$ presently denoted the MAP solution.

We evaluate the consistency of the cluster time-series across subjects by computing the posterior mean, $\boldsymbol{S}^{-1}_{k,s}\bar{\boldsymbol{x}}_{k,s}$, for all clusters and all subjects and rank the clusters according to the mean correlation (over all pairs of subjects), i.e.

$$R(k) = \frac{1}{S(S-1)/2} \sum_{s>s'} \text{correlation}(\boldsymbol{\mu}_{k,s}, \boldsymbol{\mu}_{k,s'}). \qquad (9)$$

To evaluate the consistency of the clusterings we use evidence accumulation [10, 11] in order to quantify how consistent across L separate sampling chains (excluding the chain with the MAP solution) voxels are grouped together according to the following cluster specific consistency score

$$C(k) = \frac{1}{n_k(n_k-1)/2} \sum_{i>j: z_i^{\text{MAP}} = z_j^{\text{MAP}} = k} \frac{1}{L} \sum_l \mathbb{I}(z_i^{(l)} = z_j^{(l)}), \qquad (10)$$

where $\mathbb{I}(a)$ is the indicator function that evaluates to 1 if a is true and 0 otherwise and n_k is the number of voxels in cluster k.

3. RESULTS AND DISCUSSION

To impose smoothness we use as kernel for the covariance of the Gaussian Process, $\boldsymbol{\Sigma}$, that is generated by the following expression:

$$k_{\text{SE}}(\boldsymbol{x}_i(t), \boldsymbol{x}_i(t')) = \exp\left(-\frac{(t-t')^2}{2l^2}\right), \qquad (11)$$

with the characteristic length-scale as the optimal length-scale for modeling the hemodynamic response function as provided by the SPM12 software (SPM12, Wellcome Trust Centre for Neuroimaging, http://www.fil.ion.ucl.ac.uk/spm/software/spm12/). The length-scale was inferred by optimizing the fit of a Gaussian Process with a squared exponential kernel and was $l = 4.6s$ or 1.85 frames (with TR = 2.49). In our analysis we initialized the clustering configuration to have all voxels in the same cluster. The noise parameter $\boldsymbol{\sigma}^2$ was initialized to the variance of the data, i.e. $\sigma_i^2 = \text{var}(\boldsymbol{x}_i)$ and the CRP parameter γ was initialized to 5.

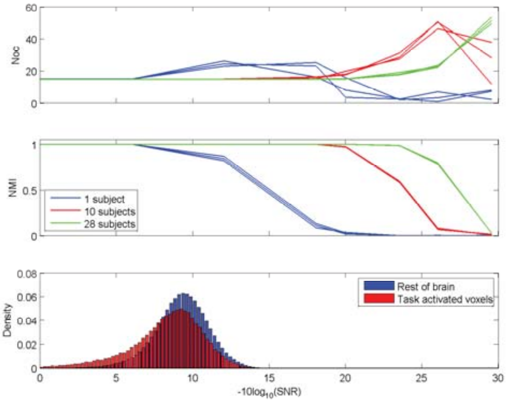


Fig. 2. Number of components extracted as a function of the SNR, NMI using a single subject, and group analysis of 10 and 25 subjects. Below is given the density of the SNR for the real multi-subject fMRI data is given in.

For the inference procedure a full sweep consisted of 1 Gibbs sweep and a number of split-merge moves defined such that the CPU time spent on the split-merge moves matched that of the Gibbs sweep. This means that the number of split-merge moves performed changed dynamically during the model inference. In each split-merge move 3 restricted Gibbs sweeps were performed. Additionally 10 sweeps of Metropolis-Hastings hyper-parameter sampling was performed for each $\sigma^2_{i,s}$ and for γ. On the fMRI dataset consisting of 28 subjects and 48799 voxels each with a time-series of 240 measurements (further details on the data is given below) this entire sampling forming one iteration took approximately one hour to complete.

To test the model on the synthetic data we performed 3 runs for each selection of noise and number of synthetic subjects. On the fMRI dataset 10 runs using the accelerated split-merge procedure and 10 runs with the standard split-merge procedure were performed to illustrate the impact on convergence of the change in the split-merge procedure.

3.1. Synthetic data

In order to investigate the level of noise for which the model can infer the correct clustering, a number of synthetic data sets were generated of varying noise. Furthermore we varied the number of subjects to illustrate how the performance of the model increases with more subjects. For each dataset we generated 15 cluster means for each subject according to Eq. (2) and (11) with a characteristic length-scale of 1.85. For each cluster we generated 400 voxels with the same temporal dimension as the fMRI dataset. This was done idependently for each of the synthetic subjects such that the only thing shared was the clustering configuration. This was done 3 times for each noise level and number of subjects pair in order to verify the

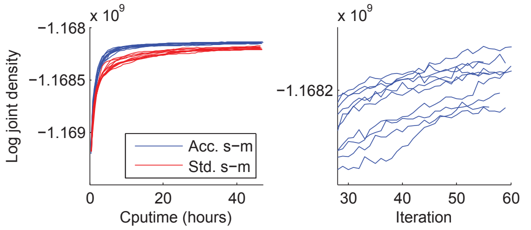


Fig. 3. Progression of the joint distribution given in equation (5) using standard split-merge sampling and using the proposed accelerated merge procedure. On the right are the last 30 iterations of the accelerated chains.

stability of the method.

To quantify the extend in which the clustering matches the ground truth we used the normalized mutual information (NMI) [16] measure as well as the number of clusters inferred as a function of the average SNR of the generated data defined by

$$\mathrm{SNR_{DB}} = 10\log_{10}\frac{\mathrm{P_{signal}}}{\mathrm{P_{noise}}} = 10\log_{10}\frac{\sum_{s,k} n_k \boldsymbol{\mu}_k^\top \boldsymbol{\mu}_k}{\sum_{s,n} T\sigma_{s,n}^2}. \qquad (12)$$

As seen in Fig. 2, the method is able to handle Gaussian noise until a $\mathrm{SNR_{DB}}$ of approximately -5. After that additional subjects are needed in order to be able to infer the correct configuration. In order to get a crude estimate of the SNR in the used fMRI data we used the following estimate: For the mean of the signal we used the 10 repetitions of the task to construct the mean, $\bar{\boldsymbol{\mu}}_v^{(s)} = \frac{1}{10}\sum_{i=1}^{10}\boldsymbol{x}_v^{(s)}$, and then estimates the SNR by:

$$\mathrm{SNR}^2_{v,\mathrm{estimate}} = \frac{\bar{\boldsymbol{\mu}}_v^{(s)}\bar{\boldsymbol{\mu}}_v^{(s)\top}}{\frac{1}{10-1}\sum_{i=1}^{10}(\boldsymbol{x}_v^{(s)} - \bar{\boldsymbol{\mu}}_v^{(s)})(\boldsymbol{x}_v^{(s)} - \bar{\boldsymbol{\mu}}_v^{(s)\top})}. \qquad (13)$$

According to this crude estimate the model should not be able to correctly infer the correct configuration for a single subject. However, this should be possible when using 10 subjects and in our regime using 28 subjects.

3.2. Multi-subject fMRI

To validate our proposed method we used a fMRI finger tapping data set consisting of 28 healthy subjects scanned in a Siemens 3T scanner. The dataset has previously been described in [17, 18]. The finger tapping paradigm consisted of two paced motor conditions each lasting 20 s, first right handed finger tapping followed by left handed finger tapping. Both conditions were paced by a blinking colored circle and were followed by 9.88 s rest. The stimulation cycle was repeated 10 times and 240 scans was acquired in total. Data was preprocessed using a default strategy in the SPM8 software package that comprised the following steps: (1) Rigid body realignment, (2) co-registration, (3) spatial-normalization to the MNI 152 template, (4) re-slicing of images into MNI space at 3 mm isotropic voxels. For the SPM analysis a spatial smoothing was further applied using an isotropic Gaussian filter (6 mm FWHM). Finally a rough grey matter mask (48799 voxels) was applied.

Fig. 3 shows the logarithm of the joint distribution for the 10 different runs. It is clear that the accelerated split-merge procedure using enhanced merge steps significantly improves on the convergence. We also observe that even using this enhanced inference procedure the model has not converged.

In order to show how many clusters are task relevant we sorted the clusters according to the cluster specific mean correlation $R(k)$ computed in Eq. (9). As seen from Fig. 4 it is clear that two clusters show a much higher degree of correlation across subjects whereas 8 clusters show a correlation higher than 0.3. Of these 8 clusters 7 show a consensus score (i.e., $C(k)$) higher than 0.6. These 7 clusters are colored in shades of red and are the clusters of similar color in the consensus score plot in Fig. 4. The cluster that shows a high level of correlation but a low consensus score is colored yellow. The 8 clusters having high correlation are also visualized in figure A)-H) of Fig. 6 and shown in descending order according to their correlation score colored as in Fig 4. According to the consensus score we also selected the 7 clusters with the highest consensus score, shown in shades of green. These 7 clusters are also shown in figure I)-O) of Fig. 6.

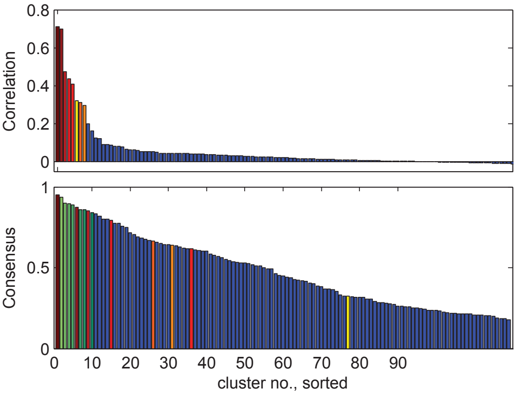


Fig. 4. Average correlation across the estimated cluster specific time-series for each subject and stability as quantified using evidence accumulation. In shades of red is given clusters that have corelation above 0.3. In green the remaining 7 clusters that are in top 10 of the consensus score.

We performed a standard multisubject SPM analysis on the smoothened data with left-right and right-left contrast maps for a comparison with the regions extracted by our method. The activation maps of the SPM analysis are thresholded and compared with the two clusters of average correlation higher than 0.7 in Fig. 5. From the figure it is clear that there is a very high degree of correspondance but also that the two top correlated maps are more localized. The SPM maps are also included in the top of Fig. 6 where it can be seen that the 10 most correlated clusters well correspond to subparcellations of the SPM maps of regions that are task activated whereas the regions with a relative high consensus score but relatively low correlation do not delineate regions that are extracted in the SPM analysis but different cortical regions that robustly group together.

4. CONCLUSION

When analyzing fMRI the data is traditionally smoothened and voxels of brain activation extracted in a supervised manner for instance

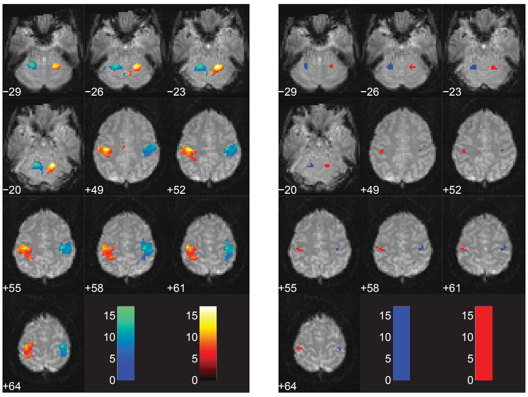


Fig. 5. On the left are SPM heatmaps of activated voxels at selected axial slices. On the right are the inferred clusters of highest cross subject correlation.

using a SPM analysis that identifies voxels with the expected temporal evolution as defined by the imposed HRF convolved with the design matrix. In contrast, our proposed approach is fully unsupervised and automatically groups voxels circumventing the need for smoothing data. It uses the temporal consistency across subjects as well as reliability over separate chains of the sampler in order to infer regions of interests. We find that regions with high temporal consistency well correspond to those derived by a standard SPM analysis whereas regions that are only reliable across chains of the sampler correspond to cortical regions that are neither identified by SPM nor our measure of correlation.

We succesfully demonstrated on a simple finger tapping paradigm that our completely unsupervised approach is able to extract the task-induced activations. This we believe forms a promising framework for the analysis of taskdata in general where there is no good knowledge of how given tasks induce BOLD responses. We also find that there is generally correspondence between regions that correlate across subjects and regions that are robustly identifed by our inference procedure. This indicates that consistency of the clustering by itself can be used to idenfity task relevant regions and can thereby be used to quantify activated regions when information on task is unavailable such as in the analysis of resting state fMRI.

5. REFERENCES

[1] S. D. Forman, J. D. Cohen, M. Fitzgerald, W. F. Eddy, M. A Mintun, and D. C. Noll, “Improved assessment of significant activation in functional magnetic resonance imaging (fmri): use of a cluster-size threshold,” *Magnetic Resonance in medicine*, vol. 33, no. 5, pp. 636–647, 1995.

[2] K. J. Worsley, S. Marrett, P. Neelin, A. C. Vandal, K. J. Friston, and A. C. Evans, “A unified statistical approach for determining significant voxels in images of cerebral activation,” *Human Brain Mapping*, vol. 4, pp. 58–73, 1996.

[3] J. Chumbley, K. Worsley, G. Flandin, and K. Friston, “Topological fdr for neuroimaging,” *Neuroimage*, vol. 49, no. 4, pp. 3057–3064, 2010.

[4] C. Goutte, P. Toft, E. Rostrup, F. Å Nielsen, and L. K. Hansen, “On clustering fmri time series,” *NeuroImage*, vol. 9, no. 3, pp. 298–310, 1999.

[5] R. Heller, D. Stanley, D. Yekutieli, N. Rubin, and Y. Benjamini, “Cluster-based analysis of fmri data,” *NeuroImage*, vol. 33, no. 2, pp. 599–608, 2006.

[6] M. J. McKeown, T. P. Jung, S. Makeig, G. Brown, S. S. Kindermann, T. W. Lee, and T. J. Sejnowski, “Spatially independent activity patterns in functional MRI data during the stroop color-naming task,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 95, pp. 803–810, Feb. 1998.

[7] J. S. Damoiseaux, S. A. R. B. Rombouts, F. Barkhof, P. Scheltens, C. J. Stam, S. M. Smith, and C. F. Beckmann, “Consistent resting-state networks across healthy subjects,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 37, pp. 13848–13853, Sept. 2006.

[8] V. D. Calhoun, T. Adali, V. B. McGinty, J. J. Pekar, T. D. Watson, and G. D. Pearlson, “fmri activation in a visual-perception task: network of areas detected using the general linear model and independent components analysis,” *NeuroImage*, vol. 14, no. 5, pp. 1080–1088, 2001.

[9] S. Jain and R. M. Neal, “A split-merge markov chain monte carlo procedure for the dirichlet process mixture model,” *Journal of Computational and Graphical Statistics*, vol. 13, no. 1, 2004.

[10] A. L. N. Fred and A. K. Jain, “Combining multiple clusterings using evidence accumulation,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 6, pp. 835–850, 2005.

[11] S. Ryali, T. Chen, A. Padmanabhan, W. Cai, and V. Menon, “Development and validation of consensus clustering-based framework for brain segmentation using resting fmri,” *Journal of neuroscience methods*, vol. 240, pp. 128–140, 2015.

[12] D. J. Aldous, *Exchangeability and related topics*, Springer, 1985.

[13] J. Pitman et al., “Combinatorial stochastic processes,” Tech. Rep., Technical Report 621, Dept. Statistics, UC Berkeley, 2002. Lecture notes for St. Flour course, 2002.

[14] C. E. Rasmussen, “The infinite gaussian mixture model.,” in *NIPS*, 1999, vol. 12, pp. 554–560.

[15] J. Ross and J. Dy, “Nonparametric mixture of gaussian processes with constraints,” in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 1346–1354.

[16] L. N. F. Ana and A. K. Jain, “Robust data clustering,” in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*. IEEE, 2003, vol. 2, pp. II–128.

[17] P. M. Rasmussen, L. K. Hansen, K. H. Madsen, N. W. Churchill, and S. C. Strother, “Model sparsity and brain pattern interpretation of classification models in neuroimaging,” *Pattern Recognition*, vol. 45, no. 6, pp. 2085–2100, 2012.

[18] P. M. Rasmussen, T. J. Abrahamsen, K. H. Madsen, and L. K. Hansen, “Nonlinear denoising and analysis of neuroimages with kernel principal component analysis and pre-image estimation,” *NeuroImage*, vol. 60, no. 3, pp. 1807–1818, 2012.

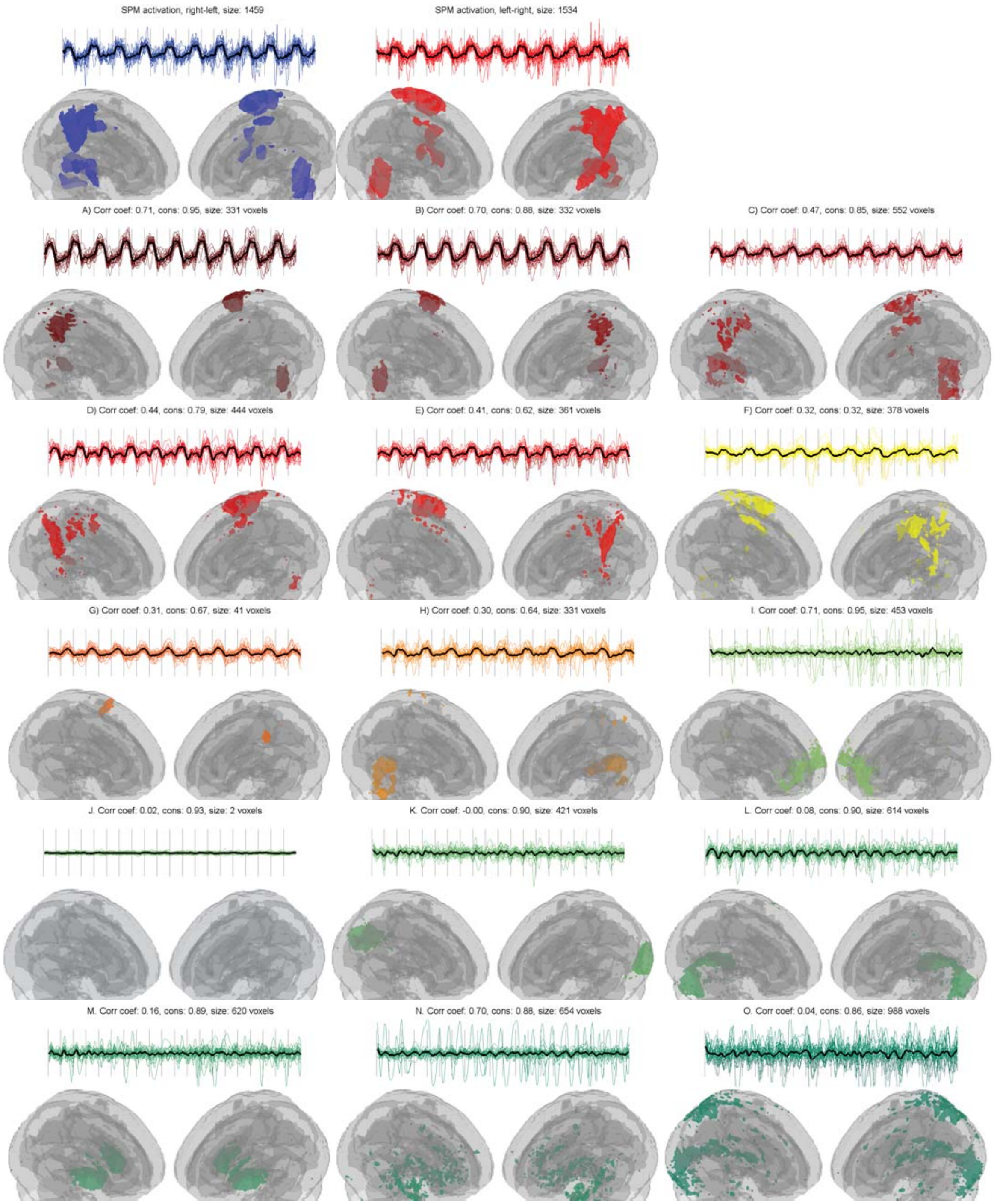

SPM activation, right-left, size: 1459
SPM activation, left-right, size: 1534
A) Corr coef: 0.71, cons: 0.95, size: 331 voxels
B) Corr coef: 0.70, cons: 0.88, size: 332 voxels
C) Corr coef: 0.47, cons: 0.85, size: 552 voxels
D) Corr coef: 0.44, cons: 0.79, size: 444 voxels
E) Corr coef: 0.41, cons: 0.62, size: 361 voxels
F) Corr coef: 0.32, cons: 0.32, size: 378 voxels
G) Corr coef: 0.31, cons: 0.67, size: 41 voxels
H) Corr coef: 0.30, cons: 0.64, size: 331 voxels
I. Corr coef: 0.71, cons: 0.95, size: 453 voxels
J. Corr coef: 0.02, cons: 0.93, size: 2 voxels
K. Corr coef: -0.00, cons: 0.90, size: 421 voxels
L. Corr coef: 0.08, cons: 0.90, size: 614 voxels
M. Corr coef: 0.16, cons: 0.89, size: 620 voxels
N. Corr coef: 0.70, cons: 0.88, size: 654 voxels
O. Corr coef: 0.04, cons: 0.86, size: 988 voxels

Fig. 6. The activation maps for the two conditions using group SPM analysis and the average of the time-series of the activated voxels. In A)-O) are the extraced functional activation maps using our fully unsupervised approach. A)-H) are the clusters of high cross subject correlation, and only F) has a consensus score lower than 0.6. In I)-O) are the reliable maps that are not highly correlated in time across subjects.

Appendix B

Infinite von Mises-Fisher Mixture Model- ing of Whole-Brain fMRI Data

Røge, R. E., Madsen, K. H., Schmidt, M. N., Mørup, M. (2015), 'Non-parametric von-Mishes Fisher based clustering of whole-brain resting state fMRI'. *Submitted 22 Nov. 2016, invited for resubmission within three months 06 Jan. 2017.*

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

kristoffer.madsen@gmail.com

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

Section for Cognitive Systems, DTU Compute, Technical University of Denmark, DK-2800, Kgs. Lyngby, Denmark

Morten Mørup

mmor@dtu.dk

Section for Cognitive Systems, DTU Compute, Technical University of Denmark, DK-2800, Kgs. Lyngby, Denmark

Abstract

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 paper, 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 both models are able to recover the true underlying clustering of spherical data when the number of clusters is known; however, when performing model selection, the two models are not in agreement. Analyzing multi-subject whole brain fMRI data, we find that the vMF mixture model is considerably more reliable than the GMM when comparing solutions across models trained on different groups of subjects, and again we find that the two models disagree on the optimal number of components. The analysis indicates that the fMRI data support more than a thousand clusters, and we confirm this is not a result of over-fitting 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. The developed vMF software can be applied to data clustering on the hypersphere in general, and is available from the authors.

1 Introduction

In many areas of statistical modeling, data is represented only by a direction, thus setting the stage for directional statistics (Mardia and 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 et al., 2010), directions from the earth to stars (Mardia and Jupp, 2009), or locating emergency transmitters (Guttorp and Lockhart, 1988). One of the most frequently used directional distributions is the von Mises-Fisher distribution (vMF) (Fisher, 1953; Mardia and 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 (Nunez-Antonio and Gutiérrez-Pena, 2005) compared to e.g. the Gaussian distribution.

Models based on the von Mises-Fisher distribution have been applied to a wide variety of high-dimensional problems on the unit hypersphere. This includes document topic modeling (Banerjee et al., 2005; Gopal and Yang, 2014) and the modeling of gene expressions data (Banerjee et al., 2005; Taghia et al., 2014). 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 et al., 2012; Hyde and Jesmanowicz, 2012). Z-scoring transforms each voxel time series to have zero mean and unit standard deviation, i.e. voxel time series data consisting of D brain volumes will therefore be projected onto the hypersphere with radius $\sqrt{D-1}$. Since there is no longer any information in the magnitude of the observations, 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, was recently modeled using a mixture model based on the von Mises-Fisher distributions with Markov random field to ensure spatial contiguity (Ryali et al., 2013). The von Mises-Fisher distrubiton has also been frequently used in modeling fMRI task activations (Lashkari et al., 2010; Lashkari and 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 et al., 2012) or only considered fMRI time-series within a small region of interest (Ryali et al., 2013). Furthermore, neither of these studies have 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 von Mises-Fisher distribution might be, as opposed to applying the well-studied and more simple 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 et al., 2015).

In this article, we advance the von Mises-Fisher mixture model to large-scale fMRI clustering. We employ collapsed Markov chain Monte Carlo (MCMC) inference and exploit non-parametric Bayesian modeling for model order quantification. We apply the developed framework to multi-subject whole-brain fMRI segmentation and contrast the performance of the von Mises-Fisher 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 multi-subject fMRI data and contrast the estimated model order based on non-parametric Bayesian modeling to the model order estimated using the predictive distribution based on finite mixtures.

The paper is structured as follows: In §2 we introduce the generative models and inference procedure for our non-parametric vMF mixture model. In §3 we present results regarding the implementation of the vMF models. We apply our model to multi-subject resting state fMRI data and contrast the performance to conventional parametric and non-parametric Gaussian mixture modeling. Finally, in §4 we present our conclusions. In Appendix A we compare our implementation to an existing implementation based on variational inference (Gopal and Yang, 2014).

2 Methods

Clustering using a mixture of von Mises-Fisher 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 von Mises-Fisher 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 non-parametric "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 von Mises-Fisher 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 either using an approximation (Taghia et al., 2014; Gopal and Yang, 2014) or by MCMC sampling (Gopal and 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 contribution we present the Bayesian generative model for clustering directional data based on von Mises-Fisher distributions. Similar to Bangert et al. (2010) we formulate a non-parametric 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 short review of the von Mises-Fisher distribution, we present the von Mises-Fisher mixture model along with the numerical approximations, a description of the inference procedure, and posterior quantities used for the subsequent analyses.

2.1 The von Mises-Fisher distribution

The von Mises-Fisher 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 von Mises-Fisher probability density is given by

$$\text{vMF}(\boldsymbol{x} \mid \boldsymbol{\mu}, \tau) = C_D(\tau) \exp(\tau \boldsymbol{\mu}^\top \boldsymbol{x}), \tag{1}$$

where

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

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 von Mises-Fisher distribution with parameters $\{\boldsymbol{\mu}_0, \tau_0\}$ is in itself a conjugate prior for the mean direction.

For N observations from a von Mises-Fisher distribution with concentration τ, the marginal likelihood for τ is given by

$$p(\boldsymbol{x}_{1:N} \mid \tau) = \int \text{vMF}(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, \tau) \prod_{i=1}^{N} \text{vMF}(\boldsymbol{x}_i \mid \boldsymbol{\mu}, \tau) d\boldsymbol{\mu} = \frac{C_D(\tau)^{N+1}}{C_D\left(\tau \left\|\boldsymbol{\mu}_0 + \sum_{i=1}^{N} \boldsymbol{x}_i\right\|\right)}. \tag{3}$$

Therefore, if we apply a prior given by

$$f(\tau \mid a, b) \propto \frac{C_D(\tau)^a}{C_D(b\tau)}, \tag{4}$$

with parameters a and b where $a > b > 0$, then it corresponds to having seen a observations from a von Mises-Fisher distribution that has the combined length b (cf. Hornik and Grün 2013). The normalization constant for this prior is not available in closed form due to the dependency on the modified Bessel functions. Previous implementations have used either the lognormal or Gamma distribution (Taghia et al., 2014; Gopal and Yang, 2014) as priors, but as shown by Taghia et al. (2014) the gamma distribution very closely resembles the above prior we have chosen for our implementation.

2.2 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 $\frac{\alpha}{K}$. Marginalizing the cluster proportions, the resulting Pólya distribution with parameter $\alpha > 0$ is given by

$$\text{Pólya}(\boldsymbol{z} \mid \alpha) = \frac{\Gamma(\alpha)}{\Gamma(N + \alpha)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})}, \tag{5}$$

where N is the number of observations, K is the number of clusters, and n_k is the number of observation that belong to cluster k. Taking the limit $K \to \infty$ of the Pólya distribution yields the socalled Chinese restaurant process (CRP) (Aldous, 1985; Pitman et al., 2002)

$$\text{CRP}(\boldsymbol{z} \mid \alpha) = \frac{\Gamma(\alpha)\alpha^K}{\Gamma(N+\alpha)} \prod_{k=1}^{K} \Gamma(n_k), \tag{6}$$

where now K denotes the number of non-empty clusters. Utilizing the non-parametric 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 these 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.

2.3 Mixture model specification

Modelig data with a mixture of multiple von Mises-Fisher distributions is the classical misture model and to complete model specification for either the finite or the infinite case we need to include the Pólya distribution or Chinese restaurant process as prior on the clustering. The von Mises-Fisher mixture model is then given by the following generative process:

$$\tau_k \mid a, b \sim f(a,b) \qquad k = 1, \ldots, K \tag{7}$$

$$\boldsymbol{\mu}_k \mid \boldsymbol{\mu}_0, \tau_0 \sim \text{vMF}(\boldsymbol{\mu}_0, \tau_0) \qquad k = 1, \ldots, K \tag{8}$$

$$\boldsymbol{x}_i \mid \boldsymbol{\mu}_{\boldsymbol{z}(i)}, \tau_{\boldsymbol{z}(i)} \sim \text{vMF}(\boldsymbol{\mu}_{\boldsymbol{z}(i)}, \tau_{\boldsymbol{z}(i)}) \qquad i = 1, \ldots, N \tag{9}$$

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 Eq. (4). The joint probability of the generative model is given by

$$p(\boldsymbol{x}_{1:N}, \boldsymbol{\mu}_{1:K}, \tau_{1:K} \mid \boldsymbol{z}, \boldsymbol{\mu}_0, \tau_0, a, b) = \\ \left[\prod_{i=1}^{N} \mathrm{vMF}(\boldsymbol{x}_i \mid \boldsymbol{\mu}_{\boldsymbol{z}(i)}, \tau_{\boldsymbol{z}(i)})\right] \left[\prod_{k=1}^{K} \mathrm{vMF}(\boldsymbol{\mu}_k \mid \boldsymbol{\mu}_0, \tau_0)\right] f(\tau_k \mid a, b). \quad (10)$$

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

$$p(\boldsymbol{x}_{\mathcal{Z}_k}, \boldsymbol{\mu}_k \mid \boldsymbol{z}, \boldsymbol{\mu}_0, \tau_0, \tau_k) = \left[\prod_{i \in \mathcal{Z}_k} \mathrm{vMF}(\boldsymbol{x_i} \mid \boldsymbol{\mu}_k, \tau_k)\right] \mathrm{vMF}(\boldsymbol{\mu}_k \mid \boldsymbol{\mu}_0, \tau_0) \quad (11)$$

$$= C_D(\tau_0) C_D(\tau)^{n_k} \exp(\lambda_k \boldsymbol{m}_k^\top \boldsymbol{\mu}_k), \quad (12)$$

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

$$\lambda_k = \left\| \tau_0 \boldsymbol{\mu}_0 + \tau_k \sum_{i \in \mathcal{Z}_k} \boldsymbol{x}_i \right\|, \qquad \boldsymbol{m}_k = \frac{1}{\lambda_k}\left(\tau_0 \boldsymbol{\mu}_0 + \tau_k \sum_{i \in \mathcal{Z}_k} \boldsymbol{x}_i\right). \quad (13)$$

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

$$p(\boldsymbol{x}_{\mathcal{Z}_k} \mid \boldsymbol{z}, \boldsymbol{\mu}_0, \tau_0, \tau_k) = \int p(\boldsymbol{x}_{\mathcal{Z}_k}, \boldsymbol{\mu}_k \mid \boldsymbol{z}, \boldsymbol{\mu}_0, \tau_0, \tau_k) d\boldsymbol{\mu}_k = \frac{C_D(\tau_0) C_D(\tau_k)^{n_k}}{C_D(\lambda_k)}, \quad (14)$$

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

$$p(\boldsymbol{x}_{\mathcal{Z}_k} \mid \boldsymbol{z}, \boldsymbol{\mu}_0, \tau_0, a, b) = \int p(\boldsymbol{x}_{\mathcal{Z}_k} \mid \boldsymbol{z}, \boldsymbol{\mu}_0, \tau_0, \tau_k) f(\tau_k \mid a, b) d\tau_k \quad (15)$$

$$= C_D(\tau_0) \int \frac{C_D(\tau_k)^{n_k}}{C_D(\lambda_k)} f(\tau_k \mid a, b) d\tau_k. \quad (16)$$

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.

2.4 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 Eq. (16) can be approximated as

$$\int \frac{C_D(\tau_k)^{n_k}}{C_D(\lambda_k)} f(\tau_k \mid a, b) d\tau_k \approx \frac{1}{S} \sum_{s=1}^{S} \frac{C_D(\tau_k^{(s)})^{n_k}}{C_D(\lambda_k^{(s)})}. \tag{17}$$

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 hyper-parameters 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 then 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 and Grün, 2014)

$$\log \mathcal{I}_\nu(x) \approx \sqrt{x^2 + (\nu+1)^2} + (\nu + 1/2) \log \frac{x}{\nu + 1/2 + \sqrt{x^2 + (\nu+1)^2}} + \frac{1}{2} \log x/2 + (\nu + 1/2) \log \frac{2\nu + 3/2}{2(\nu+1)} - \frac{\log 2\pi}{2}. \quad (18)$$

Using this numerical integration, we obtain the following expression for the collapsed joint distribution:

$$p(\boldsymbol{x}_{1:N} \mid \boldsymbol{z}, \boldsymbol{\mu}_0, \tau_0, a, b) = \prod_k \frac{C_D(\tau_0)}{S} \sum_{s=1}^{S} \frac{C_D(\tau_k^{(s)})^{n_k}}{C_D(\lambda_k^{(s)})}. \quad (19)$$

2.5 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 i'th element to the k'th component using Gibbs sampling is (up to proportionality) given by

$$p(\boldsymbol{z}_i = k \mid \boldsymbol{z}_{\setminus i}, \ldots) \propto n_k \frac{\displaystyle\sum_{s=1}^{S} \frac{C_D(\tau_k^{(s)})^{n_k+1}}{C_D\Big(\big\|\tau_0 \boldsymbol{\mu}_0 + \tau_k^{(s)}[\boldsymbol{x}_i + \sum_{j \in \mathcal{Z}_k} \boldsymbol{x}_j]\big\|\Big)}}{\displaystyle\sum_{s=1}^{S} \frac{C_D(\tau_k^{(s)})^{n_k}}{C_D\Big(\big\|\tau_0 \boldsymbol{\mu}_0 + \tau_k^{(s)} \sum_{j \in \mathcal{Z}_k} \boldsymbol{x}_j\big\|\Big)}}, \quad (20)$$

with the convention that observation i has been removed from $\mathcal{Z}_{z_i}$. The posterior for assigning the element to a new cluster is proportional to

$$p(\boldsymbol{z}_i = K+1 \mid \boldsymbol{z}_{\setminus i}, \ldots) \propto \frac{\alpha C_D(\tau_0)}{S} \sum_{s=1}^{S} \frac{C_D(\tau_k^{(s)})}{C_D\big(\|\tau_0 \boldsymbol{\mu}_0 + \tau_k^{(s)} \boldsymbol{x}_i\|\big)}. \quad (21)$$

We further apply the split-merge algorithm (Jain and Neal, 2004) with accelerated merge moves (Røge et al., 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 Eq. (20), 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 proposal distribution. Furthermore, the parameters a and b has 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.

2.6 Multiple dataset 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 Eq. (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 hyper-parameters 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.

2.7 Posterior quantities

We can use Bayes' theorem to obtain the posterior probability for the concentration parameter

$$p(\tau_k \mid \boldsymbol{x}, a, b) = \frac{p(\boldsymbol{x} \mid \tau, a, b)p(\tau \mid a, b)}{\int p(\boldsymbol{x} \mid \tau, a, b)p(\tau \mid a, b)d\tau}. \tag{22}$$

This is proportional to

$$p(\tau_k \mid \boldsymbol{x}, a, b) \propto \frac{C_D(\tau_k)^{n_k} C_D(\tau_k)^a}{C_D(\lambda_k)C_D(b\tau_k)}. \tag{23}$$

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

$$p(\boldsymbol{\mu}_k \mid \boldsymbol{X}, \tau_k, \mu_0, \tau_0) \propto p(\boldsymbol{X} \mid \boldsymbol{\mu}_k, \tau_k)p(\boldsymbol{\mu}_k \mid \tau_0, \mu_0) \propto \exp(\lambda_k \boldsymbol{m}_k^\top \boldsymbol{\mu}_k). \tag{24}$$

Since this is the functional form of a von Mises-Fisher distribution we know the normalization constant and obtain

$$p(\boldsymbol{\mu}_k \mid \boldsymbol{X}, \tau_k, \mu_0, \tau_0) = C_D(\lambda_k)\exp(\lambda_k \boldsymbol{m}_k^\top \boldsymbol{\mu}_k). \tag{25}$$

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 resembles that of the von Mises-Fisher 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

$$\sigma^2_{m,k}|\nu,\gamma \sim \mathrm{IG}(\nu,\gamma) \qquad m=1,\dots,D \quad k=1,\dots,K \tag{26}$$

$$\boldsymbol{\mu}_k|\gamma,\boldsymbol{\sigma}^2_k \sim \mathcal{N}(\boldsymbol{\mu}_0,\frac{1}{\lambda}\mathrm{diag}(\boldsymbol{\sigma}^2_k)) \qquad k=1,\dots,K \tag{27}$$

$$\boldsymbol{x}_i|\boldsymbol{\mu}_{\boldsymbol{z}_i},\boldsymbol{\sigma}^2_k \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{z}_i},\mathrm{diag}(\boldsymbol{\sigma}^2_k)) \qquad i=1,\dots,N, \tag{28}$$

where IG is the inverse gamma distribution and $\mathrm{diag}(\boldsymbol{\sigma}^2_k)$ the diagonal matrix with the elements of $\boldsymbol{\sigma}^2_k$ on the diagonal. The collapsed joint distribution is, in concordance with the procedure for the von Mises-Fisher based model, obtained by marginalizing over the mean $\boldsymbol{\mu}_k$ and noise $\boldsymbol{\sigma}_k^2$ parameters.

$$\begin{aligned} p(\boldsymbol{x}_{1:N}, \boldsymbol{z} \mid \boldsymbol{\theta}) =& \prod_{k=1}^{K} \int \int \prod_{i \in k} p(\boldsymbol{X}, \boldsymbol{z} \mid \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2) p(\boldsymbol{\mu}_k \mid \boldsymbol{\sigma}_k^2) p(\boldsymbol{\sigma}_k^2) d\boldsymbol{\mu}_k d\boldsymbol{\sigma}_k^2 \\ =& \prod_{k=1}^{K} \prod_{m=1}^{D} \frac{(\lambda/[n_k+\lambda])^{1/2} \gamma^{\nu} \Gamma(n_k/2+\nu)}{(2\pi)^{n_k/2} \Gamma(\nu) R_{mk}^{n_k/2+\nu}}, \end{aligned} \tag{29}$$

where

$$R_{mk} = \gamma + \frac{1}{2} \left(\bar{\sigma}_{mk}^2 + \lambda \mu_{0_m}^2 - \frac{(\bar{x}_k + \lambda \mu_{0_m})^2}{n_k + \lambda} \right), \tag{30}$$

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

$$\sigma_k^2 | \nu, \gamma \quad \sim \text{IG}(\nu, \gamma) \qquad k = 1, \ldots, K \tag{31}$$

$$\boldsymbol{\mu}_k | \sigma_k^2, \lambda \quad \sim \mathcal{N}(\boldsymbol{\mu}_0, \frac{\sigma_k^2}{\lambda} \boldsymbol{I}) \qquad k = 1, \ldots, K \tag{32}$$

$$\boldsymbol{x}_i | \boldsymbol{\mu}_k, \sigma_k^2 \sim \mathcal{N}(\boldsymbol{\mu}_{z_i}, \sigma_k^2 \boldsymbol{I}) \qquad i = 1, \ldots, N \tag{33}$$

and the collapsed joint distribution is given by

$$p(\boldsymbol{x}_{1:N}, \boldsymbol{z} \mid \boldsymbol{\theta}) = \prod_{k=1}^{K} \frac{(\lambda/[n_k+\lambda])^{D/2} \gamma^{\nu} \Gamma(Dn_k/2+\nu)}{(2\pi)^{Dn_k/2} \Gamma(\nu) R_k^{Dn_k/2+\nu}}, \tag{34}$$

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

$$R_k = \boldsymbol{\gamma} + \frac{1}{2} \left(\bar{\sigma}_k^2 + \lambda \|\boldsymbol{\mu}_0\|^2 - \frac{\|\bar{\boldsymbol{x}}_k + \lambda \boldsymbol{\mu}_0\|^2}{n_k + \lambda} \right). \tag{35}$$

We apply the same inference procedure as with the von Mises-Fisher mixture model with suitable priors and transformations on the hyperparameters.

2.1 Predictive analysis

In order 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

$$p(\boldsymbol{x}^* \mid \boldsymbol{X}) = \int p(\boldsymbol{x}^* \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{X}) d\boldsymbol{\theta}, \tag{36}$$

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

$$p(\boldsymbol{x}^* \mid \boldsymbol{X}) \approx \frac{1}{M} \sum_{m=1}^{M} p(\boldsymbol{x}^* \mid \boldsymbol{\theta}^{(m)}). \tag{37}$$

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

$$p(\boldsymbol{x}^* \mid \boldsymbol{X}) = \sum_{k=1}^{K} p(\boldsymbol{z}_{\boldsymbol{x}^*} = k \mid \boldsymbol{X}) p(\boldsymbol{x}^* \mid \boldsymbol{X}, \boldsymbol{z}_{\boldsymbol{x}^*} = k), \tag{38}$$

where $p(\boldsymbol{z}_{\boldsymbol{x}^*} = k \mid \boldsymbol{X})$ 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,

$$\begin{aligned} p(\boldsymbol{x}^* \mid \boldsymbol{X}, \boldsymbol{z}_{\boldsymbol{x}^*} = k) &= \int p(\boldsymbol{x}^* \mid \boldsymbol{X}, \boldsymbol{z}_{\boldsymbol{x}^*} = k, \tau_0, \boldsymbol{z}, a, b) p(\tau_0, \boldsymbol{z}, a, b \mid \boldsymbol{X}) d\{\tau_0, \boldsymbol{z}, a, b\} \\ &= \frac{1}{M} \sum_{m=1}^{M} p(\boldsymbol{x}^* \mid \boldsymbol{X}, \boldsymbol{z}_{\boldsymbol{x}^*} = k, \tau_0^{(m)}, \boldsymbol{z}^{(m)}, a^{(m)}, b^{(m)}). \end{aligned} \tag{39}$$

We can compute part of this expression analytically and part with MCMC samples from the posterior of τ_k as given by Eq. (23):

$$\begin{aligned} p(\boldsymbol{x}^* \mid \boldsymbol{X}, \boldsymbol{z}_{\boldsymbol{x}^*} = k, \tau_0, \boldsymbol{z}, a, b) &= \iint p(\boldsymbol{x}^* \mid \boldsymbol{\mu}_k, \tau_k) p(\boldsymbol{\mu}_k \mid \boldsymbol{X}, \boldsymbol{\mu}_0, \tau_0) d\boldsymbol{\mu}_k p(\tau_k \mid \boldsymbol{X}, a, b) d\tau_k \\ &= \int \frac{C_D(\tau_k) C_D(\lambda_k^{(s)})}{C_D(\lambda_k^{*(s)})} p(\tau_k \mid \boldsymbol{X}, a, b) d\tau_k, \qquad (40) \end{aligned}$$

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 will therefore investigate a number of 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 similarity between clusterings

To quantify the reliability of the inferred clusters we will use the following three frequently used measures of similarity between clusterings; Normalized Mutual Information (NMI) (Strehl and Ghosh, 2002), Adjusted Mutual Information (AMI) (Vinh et al., 2010), and the Adjusted Rand index (AR) (Hubert and Arabie, 1985). The NMI and AMI measures have several variants and we have used the following:

$$\mathrm{NMI} = \frac{\mathrm{MI}(\boldsymbol{z}_z, \boldsymbol{z}_2)}{\sqrt{H(\boldsymbol{z}_1)H(\boldsymbol{z}_2)}} \tag{41}$$

and

$$\mathrm{AMI} = \frac{\mathrm{MI}(\boldsymbol{z}_z, \boldsymbol{z}_2) - \mathrm{E}[\mathrm{MI}(\boldsymbol{z}_z, \boldsymbol{z}_2)]}{\max(H(\boldsymbol{z}_1), H(\boldsymbol{z}_2)) - \mathrm{E}[\mathrm{MI}(\boldsymbol{z}_z, \boldsymbol{z}_2)]}, \tag{42}$$

where $\mathrm{MI}(\boldsymbol{z}_z, \boldsymbol{z}_2)$ is the mutual information between clusterings $\boldsymbol{z}_z$ and $\boldsymbol{z}_2$, 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

$$\mathrm{AR} = \frac{\mathrm{RI} - \mathrm{E}[\mathrm{RI}]}{\max(\mathrm{RI}) - \mathrm{E}[\mathrm{RI}]}, \tag{43}$$

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 tends to have 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.

2.5 Implementation

Both the Gaussian and the von Mises-Fisher 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 extendible with additional statistical clustering models. Our code and examples are available at `https://brainconnectivity.compute.dtu.dk`.

Variational inference based von Mises-Fisher clustering models have previously been applied to a variety of document topic modelling datasets (Gopal and 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 dataset (Gopal and Yang, 2014) based on normalized mutual information between the inferred clusters and ground truth. This comparison can be found in Appendix A 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 multi-subject resting state fMRI dataset 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.

To analyze how many samples are needed, we generated a small dataset according to the model, ie. we fixed the clustering to $K = 10$ with 20 elements in each cluster, gener-

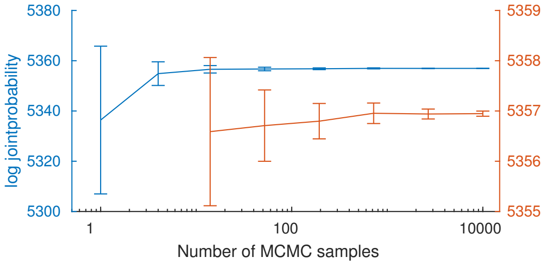
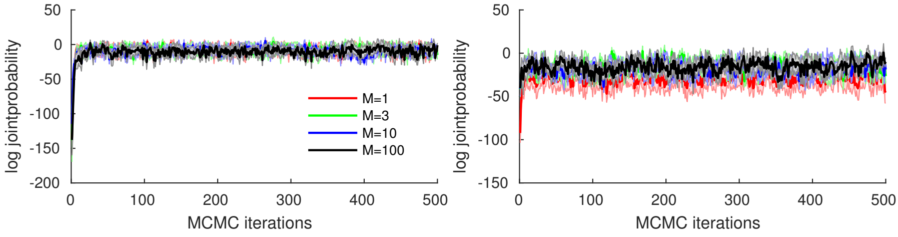


Figure 1: The likelihood of the model as a function of the number of MCMC samples used to approximate the integral. The data used for this comparison is generated according to the VMF model with parameters $T = 50, N = 200, K = 10, \text{avg}(\tau_k) = 30, \text{std}(\tau_k) = 2$. Note that the red graph is just a close-up of the blue.

ated $\tau_k \sim \mathcal{N}(\tau_{\text{avg}}, \tau_{\text{std}})$ 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 $i = 1, \dots, N$. This procedure was used for the generation of each of the synthetic datasets used for the analyses in this section. We then varied the number of samples used to approximate the integral. The results are presented in Fig. 1. With only one sample used to approximate the integral, the standard deviation of the approximated integral is less than 3 pct. 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 datasets and ran the inference procedure with varying number of samples used to approximate the integral. From the results given in Fig. 2a and 2b we observe that on datasets 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 between three and ten samples for the approximation of the integral based on ad hoc tests on each of the datasets.



(a) $\mathrm{std}(\tau_k) = 2$

(b) $\mathrm{std}(\tau_k) = 20$

Figure 2: The effect of different number of samples used in approximating the integral on inference. It is clear that as the clusters differ more in concentration parameter the more samples are needed for sufficient inference. The colored regions are ± the standard deviation over six restarts on different datasets. The data used for this comparison is generated according to the vMF mixture model with parameters $T = 50, N = 200, K = 10, \mathrm{avg}(\tau_k) = 30$.

3.2 Analysis of initialization

To investigate the impact of initialization we compared four different initialization strategies on synthetic data:

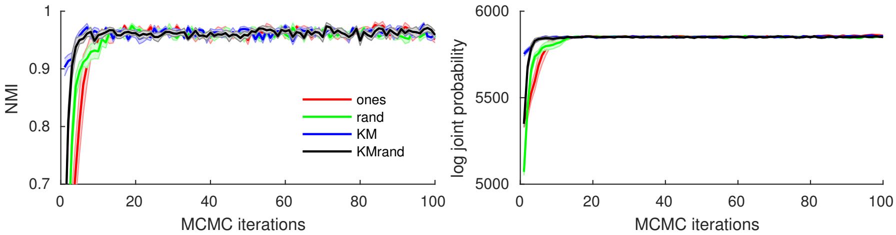
ones Initializing all elements to the same cluster followed by the evaluation of 100 MCMC proposals for each hyperparameter.

rand Initializing each label at random among K clusters followed by the evaluation of 100 MCMC proposals for each hyperparameter.

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 Fig. 3. We achieve must faster convergence with the K-means initialization but we also observe that the other initialization strategies reach similar solutions when the models have converged. For the remainder of the paper we have used the *KMrand* initialization strategy as it avoids initializing to a local minimum.



(a)

(b)

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

3.3 A 3-dimensional example

To illustrate the model we generated a small 3 dimensional dataset comprised of 6 clusters with 40 elements in each cluster. We generated the mean directions from a von Mises-Fisher distribution with mean direction as the 3rd canonical unit vector $\boldsymbol{e}_3$ and $\tau_0 = 0.01$. For each cluster we generated the concentration parameter randomly from $\mathcal{N}(50, 20^2)$.

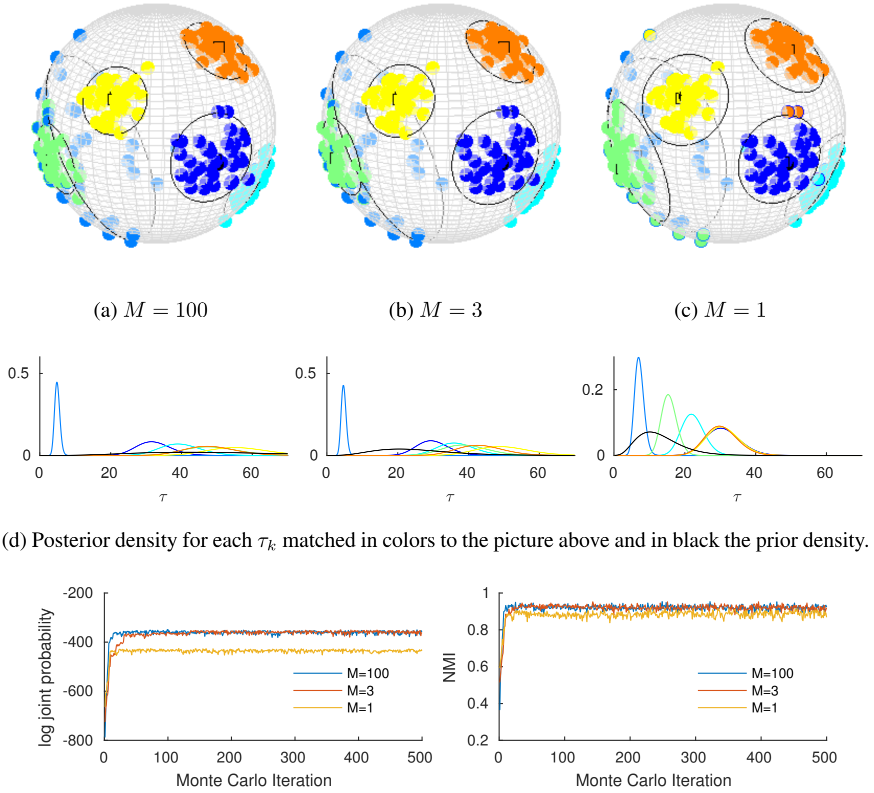
We ran three samplers using either 1, 3, or 100 samples to estimate the integral and stopped the inference chains after 200 Monte-Carlo iterations. For each of the three runs we present the clustering from the highest likelihood sample in Figs. 4a, 4b, and 4c. The circles on the spheres represents the 95 pct. 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 Fig. 4d.

Finally, we present the log joint probability and the NMI for each iteration of the inference chains in Figs. 4e and 4f. It is clear that 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 are 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

To analyse the differences between using a Gaussian and von Mises-Fisher based mixture model we generated several datasets according to the generative model for the mixture of von Mises-Fisher distributions and applied the GMM based on spherical (GMMs) and elliptical covariance (GMMd) as well as the vMF mixture model.

We generated two sets of datasets with either $T = 240$ or $T = 1000$ such that the temporal dimension of the datasets generated matched respectively the resting state fMRI dataset and the CNAE-9 dataset used in Appendix A. Subsequently, we generated $\tau_k \sim \mathcal{N}(\tau_{\text{mean}}, \tau_{\text{std}})$ with $\tau_0 = 0.4 \cdot \text{mean}(\tau_k)$, $\text{std}(\tau_k) = 0.6 \cdot \text{mean}(\tau_k)$ and τ_{mean} in a suitable range such that the signal to noise ratios ranged from easy to hard clustering



(a) $M = 100$

(b) $M = 3$

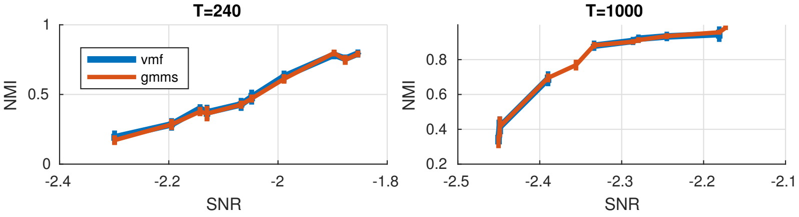
(c) $M = 1$

(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 is presented on the top. The center color denotes the actual clustering while 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.



(a)

(b)

Figure 5: Results on simulated datasets. We see that there is little difference in the ability of the vMF and GMM based models to recover the correct clustering as measured by the normalized mutual information. For each the datasets we have generated 10 clusters of 100 observations.

problems.

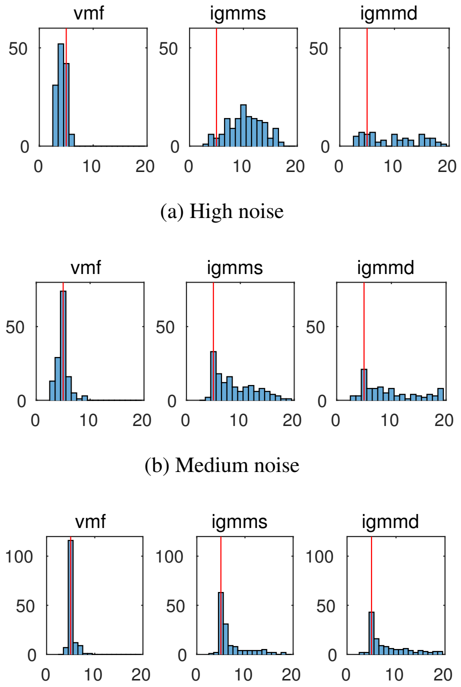
For each dataset generated, we ran 10 repetitions of the GMMs and VMF mixture models and computed the normalized mutual information between the highest likelihood sample and the true clustering configuration. The results are presented in Fig. 5 and show no significant difference in how closely the inferred clusterings resemble the true clustering between the three models even though the data was generated according to the von Mises-Fisher distribution.

Next, we explored how the Gaussian and vMF non-parametric models handled data generated from a mixture of vMFs with parameters $N = 100$, $K = 5$, $T = 30$, $\tau_0 = 30$, $\tau_{\text{std}} = 25$, and $\tau_{\text{mean}} = \{20, 25, \text{ and } 30\}$ for high noise, medium noise, and low noise datasets respectively. For each of the three settings we generated $\tau_k \sim \mathcal{N}(\tau_{\text{mean}}, \tau_{\text{std}})$ and then generated the dataset. 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 Fig. 6. It is clear that the vMF based nonparametric models infer a number of cluster much closer to the truth compared to both the spherical and ellipsical Gaussian mixture models. This emphasizes the importance of modeling data using directional statistics in determining the complexity of a dataset.

3.5 Resting state fMRI analysis

Functional brain connectivity can be assessed by analysing 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 et al., 1995). Rather than specifying seeds, clustering methods extract prominent latent activation profiles and identifies corresponding brain networks (Craddock et al., 2012). These latent class models are useful as they don't rely on a priori specification of seeds and can provide an overview of the functional organization across large high-dimensional datasets. 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 (Strother et al., 2002; Thirion et al., 2014; Churchill et al., 2016). As correlation is formed by the inner product of standardized fMRI time-series thus naturally complying with the von Mises-Fisher distribution assumptions the von Mises-Fisher mixture model is attractive for clustering



(a) High noise

(b) Medium noise

(c) Low noise

Figure 6: Results from non-parametric 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 von Mises-Fisher based non parametric mixture models in general finds solutions closer to the truth both in terms of NMI and the inferred number of clusters.

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 dataset consisting of 30 healthy subjects scanned on a Siemens 3T MRI scanner. The dataset has been previously used in (Andersen et al., 2014). During the functional scans the participants were instructed to keep their eyes closed and to refrain from any voluntary motor or cognitive activity while the 480 brain volumes were scanned over 20 minutes with a repetition time of 2.49s.

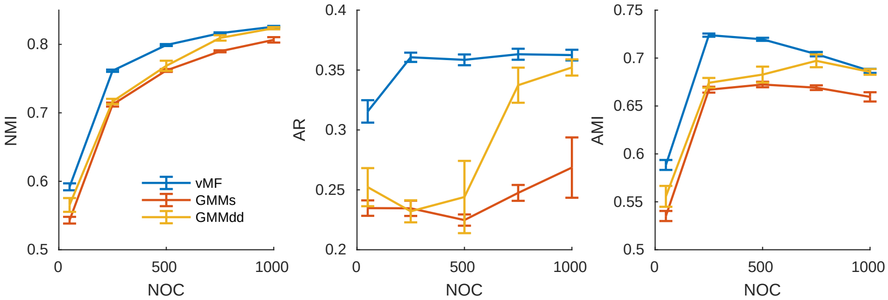
Data was preprocessed using the SPM12 software package (SPM12, Wellcome Trust Centre for Neuroimaging, `http://www.fil.ion.ucl.ac.uk/spm/software/spm12/`) with the following steps: (1) Rigid body realignment, (2) co-registration, (3) spatial normalization to the MNI 152 template, (4) reslicing of images into MNI space at 3 mm isotropic voxels, (5) spatial smoothing was applied with a 6 mm FWHM isotropic Gaussian filter. Finally, a rough grey matter mask consisting of 48799 voxels was applied.

We divide the dataset 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, $K = \{50, 250, 500, 750, 1000\}$ to the time series data using the KMrand initialization strategy. For each model we perform 100 MCMC iterations and repeated the process four times on each of the two datasets for each of the three models and for each of the four settings of K resulting in a total of 96 runs. Note that even though 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 different groups of five subjects: Normalized Mutual Information (NMI), Adjusted Mutual Information (AMI), and the Adjusted Rand index (AR). The results are presented in Fig. 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 50 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 parcellating the Striatum (Janssen et al., 2015). The results here clearly shows 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.

To verify that five subjects is sufficient to get stable clusterings we applied the clusterings to two groups of 15 subjects for the three parametric models with NOC $= 50$ or 250. The mean of the measures of clustering consistency from these clusterings are well within the error bars of the plots in Fig. 7 further emphasizing the results.

We then applied the vMF model on the two datasets 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 Fig. 10 the results of this predictive analysis are given and show that the non-parametric models require in the order of a few thousand parcels



(a) (b) (c)

Figure 7: In (a), (b), and (c) are normalized mutual information, rand index, and adjusted mutual information between groups of 5 subjects.

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 non-parametric 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 Fig. 10a whereas boxplots of the number of clusters inferred as well as NMI across the two groups of subjects is presented in Fig. 10b. 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 in the order of a few thousand clusters in order to adequately account for the functional organization of the fMRI data and that the non-parametric models by identifying a large number of clusters are not overfitting to the data.

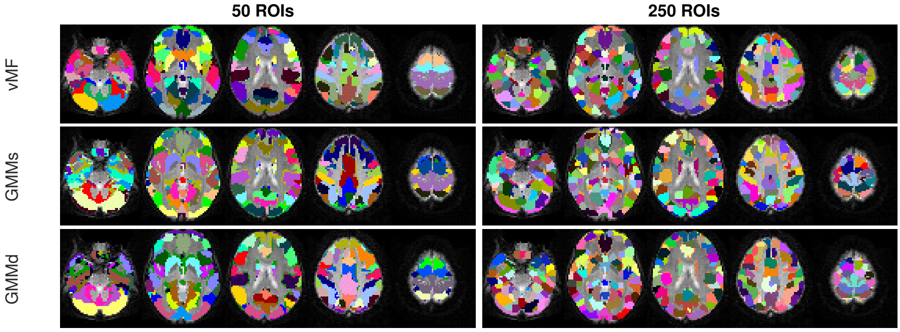


Figure 8: Visual comparison of axial slices of the solutions from the vMF, GMMs, and GMMd clustering methods for $K = 50$ and $K = 250$.

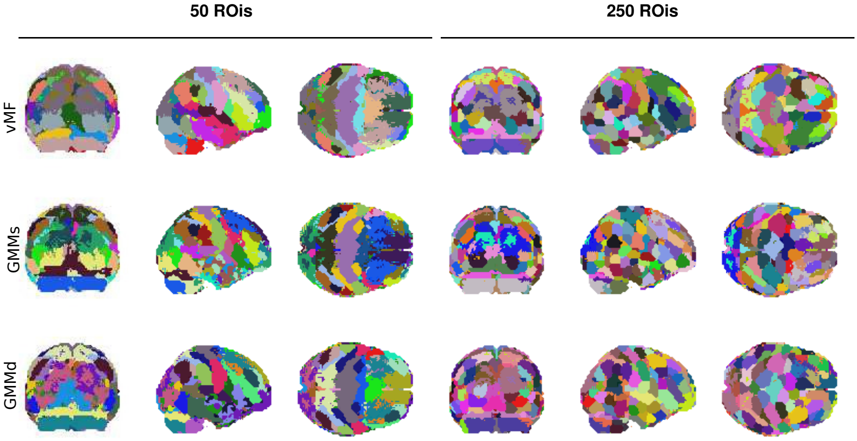
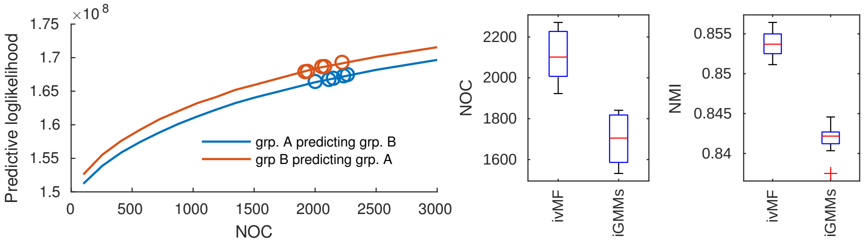


Figure 9: Visual comparison of the surface of the clustering solutions from the vMF, GMMs, and GMMd clustering methods for $K = 50$ and $K = 250$.



(a) (b)

Figure 10: Predictive analysis to determine the number of clusters that is support for in the data is shown in (a) along with the non-parametric results from the vMF model in circles. In (b) you see the comparison between the repetitions of the iVMF, iGMMs, and iGMMd models.

4 Conclusion

In this paper, we presented a thorough comparison of the effect of modeling directional data using von Mises-Fisher based distributions in comparison to assuming the data is Gaussian distributed considering both synthetic data and large scale clustering of resting state whole-brain fMRI time series. We demonstrated that there is a significant improvement in terms of the stability of solutions across groups of subjects when correctly imposing that the data resides on a hyper-sphere over the standard assumption of Gaussian distributed observations. We have further shown that it is computationally feasible to apply sampling based inference on multi subject whole-brain fMRI time series data.

The predictive analysis show that employing Bayesian non-parametrics can be an 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 non-parametric analysis show that the resting state fMRI dataset supports a number of clusters in the order of a few thousands which is in correspondence with recent findings (Thirion et al., 2014). Modelling directional data using the appropriate directional distributions shows great promise and this is an area worth more attention. Therefore, 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 presently considered mixture modeling applications, however, we anticipate the use of the von Mises-Fisher distribution may turn useful in general when modeling standardized fMRI time-series.

Appendix A: Document topic modelling

Document topic modelling is an application where variational inference based von Mises-Fisher models have shown great promise (Gopal and 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 dataset.

The CNAE-9 dataset consists of 1080 documents and each document is a vector of the frequency of occurrence for 857 words, i.e. $N = 1080$ and $T = 857$. The documents are divided into 9 categories, and the true clustering is thus available. Before clustering we perform term frequency - inverse document frequency (tf-idf) on the dataset which is a standard preprocessing step for topic modelling and known to increase performance (Salton and McGill, 1986). 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

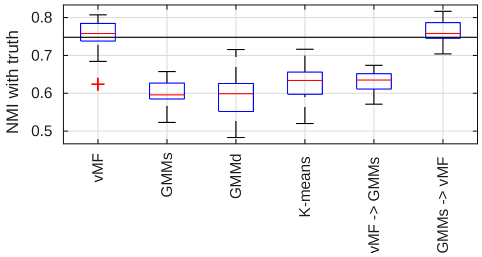


Figure 11: NMI with truth of the methods implemented. The solid black line is the results of averaging 10 repetitions of the VI implementation of the mixture of vMF model reported by Gopal et al. (Gopal and Yang, 2014).

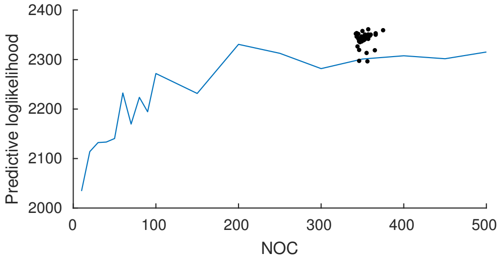


Figure 12: The predictive likelihood on the 10 pct. hold out data. The line is the result of running the vMF models with a fixed number of clusters averaged over 36 repetitions and the black dots are the result of 36 repetitions of the non-parametric vMF models.

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 afterwards. 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 et al. (Gopal and Yang, 2014) in Fig. 11.

We use the same initialization procedure for the non-parametric von Mises-Fisher model and observe that it converges to around 300 clusters. In order to validate that the data has 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 pct. of the data and computed the predictive likelihood on the hold out set. These results can be seen in Fig 12. We see that the non-parametric implementations of the vMF model are able to use the more advanced inference steps in split-merge to increase the predictive performance and that the inferred number of clusters are in a regime also supported by the predictive likelihood on hold-out test data.

References

D. J. Aldous. *Exchangeability and related topics*. Springer, 1985.

K. W. Andersen, K. H. Madsen, H. R. Siebner, M. N. Schmidt, M. Mørup, and L. K. Hansen. Non-parametric bayesian graph models reveal community structure in resting state fmri. *NeuroImage*, 100:301–315, 2014.

A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(Sep): 1345–1382, 2005.

M. Bangert, P. Hennig, and U. Oelfke. Using an infinite von mises-fisher mixture model to cluster treatment beam directions in external radiation therapy. In *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on*, pages 746–751. IEEE, 2010.

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

B. Biswal, M. Mennes, X.-N. Zuo, S. Gohel, C. Kelly, S. M. Smith, C. F. Beckmann, J. S. Adelstein, R. L. Buckner, S. Colcombe, et al. Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107(10): 4734–4739, 2010.

N. W. Churchill, K. Madsen, and M. Mørup. The Functional Segregation and Integration Model: Mixture Model Representations of Consistent and Variable Group-Level Connectivity in fMRI. *Neural Computation*, pages 1–41, aug 2016. ISSN 0899-7667.

R. C. Craddock, G. A. James, P. E. Holtzheimer, X. P. Hu, and H. S. Mayberg. A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human brain mapping*, 33(8):1914–28, aug 2012. ISSN 1097-0193.

R. Fisher. Dispersion on a sphere. In *Proceedings of the Royal Society of London A:*

Mathematical, Physical and Engineering Sciences, volume 217, pages 295–305. The Royal Society, 1953.

S. Gopal and Y. Yang. Von mises-fisher clustering models. In *ICML*, pages 154–162, 2014.

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

K. Hornik and B. Grün. On conjugate families and jeffreys priors for von misesfisher distributions. *Journal of Statistical Planning and Inference*, 145(5):992–999, May 2013.

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

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

J. S. Hyde and A. Jesmanowicz. Cross-correlation: an fMRI signal-processing strategy. *NeuroImage*, 62(2):848–51, aug 2012. ISSN 1095-9572.

S. Jain and R. M. Neal. A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1), 2004.

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

D. Lashkari and P. Golland. Exploratory fmri analysis without spatial normalization. In *International Conference on Information Processing in Medical Imaging*, pages 398–410. Springer, 2009.

D. Lashkari, E. Vul, N. Kanwisher, and P. Golland. Discovering structure in the space of fmri selectivity profiles. *Neuroimage*, 50(3):1085–1098, 2010.

K. V. Mardia and S. A. M. El-Atoum. Bayesian inference for the von mises-fisher distribution. *Biometrika*, 63(1):203–206, 1976.

K.V. Mardia and P. E. Jupp. *Directional statistics*, volume 494. John Wiley & Sons, 2009.

G. Nunez-Antonio and E. Gutiérrez-Pena. A bayesian analysis of directional data using the von mises–fisher distribution. *Communications in Statistics Simulation and Computation®*, 34(4):989–999, 2005.

J. Pitman et al. Combinatorial stochastic processes. Technical report, Technical Report 621, Dept. Statistics, UC Berkeley, 2002. Lecture notes for St. Flour course, 2002.

C. E. Rasmussen. The infinite gaussian mixture model. In *NIPS*, volume 12, pages 554–560, 1999.

R. Røge, K. H. Madsen, M. N. Schmidt, and M. Mørup. Unsupervised segmentation of task activated regions in fmri. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2015.

S. Ryali, T. Chen, K. Supekar, and V. Menon. 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, jan 2013. ISSN 10538119.

G. Salton and M. J. McGill. Introduction to modern information retrieval. 1986.

A. Strehl and J. Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617, 2002.

S. C. Strother, J. Anderson, L. K. Hansen, U. Kjems, R. Kustra, J. Sidtis, S. Frutiger, S. Muley, S. LaConte, and D. Rottenberg. The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. *NeuroImage*, 15 (4):747–71, apr 2002. ISSN 1053-8119.

J. Taghia, Zhanyu M., and A. Leijon. Bayesian estimation of the von-mises fisher mixture model with variational inference. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(9):1701–1715, Sept 2014. ISSN 0162-8828. doi: 10.1109/TPAMI.2014.2306426.

B. Thirion, G. Varoquaux, E Dohmatob, and J Poline. Which fmri clustering gives good brain parcellations? *Frontiers in neuroscience*, 8:167, 2014.

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

E. Vul, D. Lashkari, P. Hsieh, P. Golland, and N. Kanwisher. Data-driven functional clustering reveals dominance of face, place, and body selectivity in the ventral visual pathway. *Journal of neurophysiology*, 108(8):2306–2322, 2012.

B.T. T. Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zöllei, J. R. Polimeni, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 106(3):1125–1165, 2011.

Appendix C

Functional Whole-Brain Parcellation using Bayesian Non-Parametric Modeling

Røge, R. E., Schmidt, M. N., Churchill, N. W., Madsen, K. H., Mørup, M. (2015), ‘Functional Whole-Brain Parcellation using Bayesian Non-Parametric Modeling’ *Manuscript to be submitted*

Functional Whole-Brain Parcellation using Bayesian Non-Parametric Modeling

Rasmus Erbou Røge[a,*], Kristoffer Hougaard Madsen[b,a], Nathan W. Churchill[a,c], Mikkel N. Schmidt[a], Morten Mørup[a]

[a]*Section for Cognitive Systems, DTU Compute, Technical University of Denmark, Denmark*
[b]*DRCMR, Centre for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital Hvidovre, Denmark*
[c]*Neuroscience Reasearch Program, St. Michael's Hospital, Toronto, Ontario, Canada*

Abstract

Analysis of human brain data has been driven by advances in functional magnetic resonance imaging (fMRI), but challenges with poor signal to noise ratio, multiple comparison issues, and misalignment as well as variability between subjects can affect the reliability of the results. By parcellating and aggregating the fMRI data at a suitable spatial resolution, the signal to noise ratio can be increased, yet existing parcellation methods do not well account for the mentioned challenges, and require specifying the number of clusters in advance. We propose a non-parametric probabilistic model for whole-brain parcellation, that targets the signal caused by slow changes in cerebral blood flow and incorporates spatio-subject signal and noise heteroscedasticity. The model learns the number of clusters from data and utilizes efficient sampling based inference, allowing it to handle most typical group fMRI datasets. We compare the model to traditional parcellation approaches including Gaussian mixture models and k-means clustering on both synthetic data and two fMRI datasets. On synthetic data we demonstrate that it better recovers the true clustering, achieves higher correlation with the generating time series, and better copes with subject variability and spatial inconstancy. On fMRI data we find support for the heteroscedastic signal and noise model, and the reliability of the inferred parcellations is on par with the simpler Gaussian models. Furthermore, parcellated SPM analyses of regions of task activation and resting state networks show increased reliability in a test-retest framework compared to traditional voxel based analysis. The proposed fMRI parcellation method is able to find a compact and denoised representation of the data at a suitable spatial resolution, and we believe the method has utility as a general starting point for fMRI based analyses at the representation of the identified salient functional units.

[*] Corresponding Author

Email addresses: rasr@dtu.dk (Rasmus Erbou Røge), stoffer@drcmr.dk (Kristoffer Hougaard Madsen), nchurchill.research@gmail.com (Nathan W. Churchill), mnsc@dtu.dk (Mikkel N. Schmidt), mmor@dtu.dk (Morten Mørup)

1. Introduction

Functional Magnetic Resonance Imaging (fMRI) utilizes the Blood Oxygenation Level Dependent (BOLD) signal to measure the level of activation for each voxel of the brain over time and is a prominent method for the non invasive study of the function of the human brain. The voxel wise analysis of the brain has provided highly detailed information about the function of the brain using analysis tools such as Statistical Parametric Mapping (SPM) (Friston et al., 1994). There are, however, significant challenges associated with analysis using the voxel wise data due to the high level of noise apparent at voxel level as well as variability in the level of noise and signal across the brain caused by artifacts such as receive coil inhomogeneity (Wiggins et al., 2009; Kaza et al., 2011) or misalignment across subjects in group studies (Thirion et al., 2007; Mikl et al., 2008). Of primary interest in fMRI is the temporally smooth part of the signal that is caused by the slowly varying changes in the BOLD signal, however fMRI data is strongly contaminated by several noise sources of physiological or technical origin (Friston et al., 1996; Glover et al., 2000; Lund et al., 2006). Another challenge is the dimensionality of the problem since the univariate voxel wise tests applied in SPM analyses needs to be corrected for multiple comparisons across thousands of voxels imposing assumptions that are problematic(Eklund et al., 2016) whereas more advanced methods such as dynamic causal modeling (Friston et al., 2003) and other dynamic modeling approaches to fMRI (Zalesky et al., 2014) are typically not interested in or unable to work with data of the dimensionality of fMRI at the single voxel scale. Popular approaches to reduce the dimensionality includes the use of principal or independent components (McKeown et al., 1997; Calhoun et al., 2001) or using a number of regions of interest (ROI) that are either chosen from an atlas derived from anatomic analyses (Talairach and Tournoux, 1988), from multiple modalities (Glasser et al., 2016) or computed using data-driven parcellation methods (Goutte et al., 1999). The data-driven parcellation techniques have the advantage of increased interpretability compared to factorization methods and a better description of the latent prominent structure in the data compared to a priori selection of ROIs from a prespecified atlas.

There are a variety of clustering methods that have been utilized to define data-driven parcellations of fMRI voxels on a group of subjects. Popular methods include spectral clustering methods (Thirion et al., 2006; Craddock et al., 2012; Shen et al., 2013), region growing methods (Blumensath et al., 2012), hierarchical agglomerative clustering (Thirion et al., 2014) and probabilistic parcellation methods of voxel time series (Ryali et al., 2013; Janssen et al., 2015), resting state network (Mørup et al., 2010; Andersen et al., 2014; Baldassano et al., 2015), or activation maps (Lashkari and Golland, 2009; Yeo et al., 2011) as well as mixture modeling approaches imposing consistent connectivity profiles while admitting subject specific signal and noise modulation (Churchill et al., 2016).

In this paper we introduce a probabilistic method for parcellating the brain into functional units, which exploits that the underlying BOLD signal is smooth and takes the variability in signal and noise across voxels into account. The model is formulated as a non-parametric Bayesian mixture, which allows it to flexibly adapt to the complexity of the data and thereby automatically identify a suitable number of clusters to account for the salient functional structure in the fMRI data.

Our approach is inspired by the extensive literature exploring properties of the temporal dynamics of the BOLD signal where models have been proposed using Finite Impulse Response (FIR) filters or Linear Time Invariant (LTI) systems on top of a physiological model such as the Balloon model (Goutte et al., 2000; Buxton et al., 1998) to account for the signal dynamics of fMRI. Recently, models employing Gaussian Process (GP) priors (Josef et al., 2016) including our own preliminary work Røge et al. (2015) have been used to model the hemodynamics in fMRI allowing the models to focus on changes on the temporal scale defined by the length scale parameter of the GP and still deviate from the prior if the data is not in agreement. In the present work, we use a GP as prior for the parcel time series with a length-scale that is fixed to optimally fit the canonical hemodynamic response function. This means that changes in the data at this temporal scale is assigned more importance compared to more rapidly varying noise and that the posterior mean for each parcel will therefore potentially better reflect the changes in the cerebral blood flow.

We further build on previous work showing the utility of modeling the changes in noise and signal across the brain (Churchill et al., 2016; Hinrich et al., 2016). By imposing voxel specific signal and noise parameters, the model is able to group together voxels that differ in signal amplitude while reducing the influence of noisy observations. For group analyses, these parameters modulating the signal and noise are specified individually for each subject. Thereby, subjects that are in disagreement with the group parcellation, for instance caused by misalignments, can be down-weighed and their contribution to defining the parcellation reduced.

The combined effect of using the Gaussian Process as prior for the group time series and the per voxel parameters for modeling the scaling of the noise and signal are that the representation of the data can be more compact: Clusters are invariant to signal magnitude, and the posterior time series of each parcel can be robustly estimated when facing model misspecification and variability in signal and noise both within and across subjects.

We perform a thorough comparison of the presented model with two versions of the Gaussian mixture model (GMM) as well as with several non-probabilistic clustering methods: K-means (Goutte et al., 1999), Ward's clustering algorithm (Ward Jr, 1963; Thirion et al., 2014), hierarchical clustering based on region growing (Blumensath et al., 2012), and the normalized cut divisive clustering algorithm (Shi and Malik, 2000; Craddock et al., 2012). Using synthetic data, we assess the ability of the models to infer the true clustering under different levels of noise and model misspecification. We subsequently apply the method to real task- and resting-state fMRI datasets. We use the predictive likelihood to determine whether the model with per voxel parameters for noise and signal are necessary to characterize the data. We further examine the reliability of the inferred clusterings across different groups of subjects or across test/retest scanning sessions, and compare with the traditional GMM models and the non-probabilistic clustering algorithms. Finally, we verify the utility of the method using the inferred posterior cluster time series for parcellation based SPM analysis, comparing the Dice index on a temporal split-half analysis to that of single voxel SPM analysis. Previous studies have focused on the single subject reliability of the inferred regions of task activation (Gorgolewski et al., 2013) and parcellation based SPM analyses have previously shown promising results in a framework working with a number of randomized parcellations (Da Mota et al., 2014) and more recently Glasser et al. (2016) also argued for parcellation based analyses of task fMRI . We perform a similar SPM analysis to test the effect from using a parcellation based analysis on the reliability of the correlation networks of two regions traditionally associated with resting state networks (Fox et al., 2005).

This paper presents the nonparametric Gaussian mixture model with a Gaussian process prior (GMMGP) that is designed to model the artifacts encountered in fMRI data: The parcel time series model is targeted at the part of the fMRI signal that is caused by slow changes in cerebral blood flow, down-weighing the influence of voxels with high noise or low correspondence with the parcel time series. The model is further invariant to the signal magnitude and these features allows for a more compact represenation when the Bayesian nonparametrics adapts to the required complexity of the data. In §2.1 we describe the clustering method, in §2.2 and 2.3 we describe our methods of validation, and in §2.4-2.6 we describe the synthetic dataset and three fMRI datasets. In §3 we argue that traditional GMMs are unable to adequately model several of the artifacts encountered in fMRI data on synthetic data motivating the use of the proposed method. We finally present parcellation based SPM analysis on task-fMRI as well as resting-state-fMRI analysis of correlation with posterior cingulate cortex (PCC) and medial prefrontal cortex (MPF) using the representation induced by the probabilistic parcellation framework.

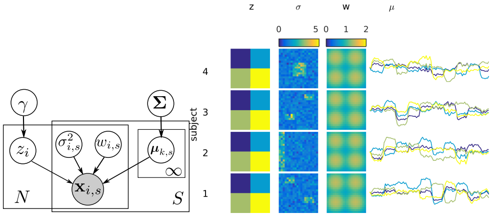
2. Methods

2.1. Probabilistic clustering model

The proposed probabilistic model for clustering fMRI time series data uses a modified Gaussian mixture model that imposes a Gaussian Process prior for the cluster time series previously considered by Ross and Dy (2013); Røge et al. (2015) as well as the ability of both signal and noise to be modulated across voxels (Churchill et al., 2016). The model can be visualized as the graphical model in Fig. 1 and is given by the following generative process:

Gaussian Mixture model with a GP prior (IGMMGP)	
Cluster assignments	$\boldsymbol{z} \sim \mathrm{CRP}(\gamma)$
Cluster mean time series	$\boldsymbol{\mu}_{k,s} \sim \mathrm{GP}(0, \beta_s \boldsymbol{\Sigma}_{\mathrm{SE}})$
Voxel time series	$\boldsymbol{x}_{i,s} \sim \mathcal{N}(w_{i,s}\boldsymbol{\mu}_{z(i),s}, \sigma^2_{i,s}\boldsymbol{I})$

Voxel time series $\boldsymbol{x}_{i,s}$ are thereby modeled by a T dimensional multivariate Gaussian distribution with a voxel (i) and subject specific (s) signal scaling parameter $w_{i,s}$ of the cluster mean time series $\boldsymbol{\mu}_{z(i),s}$ and noise parameter $\sigma^2_{i,s}$. Using the voxel specific parameters for scaling both the signal and noise we ensure that the voxels with a high signal to noise ratio contributes more to the cluster mean time series while the model thereby is able to group voxels together that differs in signal amplitude. The parameter $\boldsymbol{z}$ is a cluster assignment vector such that voxel n is assigned to cluster z_n. Using the Chinese Restaurant Process (CRP) prior for the clustering assignment we get a distribution over any possible clustering with an arbitrary number of clusters. This means that the model can automatically determine an appropriate number of clusters for which there is support in the data (Rasmussen, 1999; Aldous, 1985). In order to better compare with parametric clustering methods and the traditional Gaussian mixture models we also implement a version that employs the multivariate Pólya (compound Dirichlet-Categorical) distribution for the prior, see Appendix A for details.



(a) (b)

Figure 1: In the left panel is the generative model visualized as a directed graphical model and in the right is a visualization of the effect of having a Gaussian Process prior for the cluster time series with covariance $\boldsymbol{\Sigma}_{\mathrm{SE}}$.

The cluster mean time series are generated from a Gaussian Process with a covariance structure that is given by a squared exponential kernel function with a fixed length-scale. This kernel provides a covariance structure that allocates more significance to time series with a temporal smoothness that matches the length-scale of the squared exponential. We fix the length-scale such that the temporal smoothness matches that of the hemodynamic response function which means the model will focus on the part of the signal with this particular temporal dynamic. Since the influence of the prior will decrease as the clusters increase in size, the model will penalize small clusters more heavily in terms of smoothness.

The combined effect of using voxel specific parameters for noise and for scaling the signal and the Gaussian Process as prior for the cluster time sries are that the model is highly specialized in recovering the hemodynamic signal in the time series of each cluster such that the clustering can be better driven by these smooth dynamics than potential confounds having fast fluctuations while providing a compact representation in which the same cluster can be used to represent voxels having different signal and noise amplitudes.

We extend the model to incorporate multi-subject analysis, such that subjects normalized to the same space can share the same clustering while maintaining subject specific parameters. Incorporating signal from multiple subjects in this manner can potentially eliminate the need for long sessions of fMRI measurements to have adequate data to inform the clustering. Furthermore, with the voxel specific parameters for scaling the signal and noise, the model is better able to handle misspecifications for instance caused by misalignment of voxels across subjects, as these misaligned voxels can automatically be down-weighted through the heteroscedastic noise model. Finally, since all the hyper parameters are specific to each subject there is no need to standardize the data as the model will learn the appropriate scale for each subject.

Variation	Likelihood	Description
(S,N)	$\boldsymbol{x}_{i,s} \sim \mathcal{N}(w_{i,s}\boldsymbol{\mu}_{z(i),s}, \sigma_{i,s}^2\boldsymbol{I})$	Per voxel signal scaling and noise parameters.
(-,N)	$\boldsymbol{x}_{i,s} \sim \mathcal{N}(w_s\boldsymbol{\mu}_{z(i),s}, \sigma_{i,s}^2\boldsymbol{I})$	Voxels share signal scaling parameter, per voxel noise parameter.
(S,-)	$\boldsymbol{x}_{i,s} \sim \mathcal{N}(w_{i,s}\boldsymbol{\mu}_{z(i),s}, \sigma_s^2\boldsymbol{I})$	Per voxel signal parameter, voxels share noise parameter scaling parameter.
(-,-)	$\boldsymbol{x}_{i,s} \sim \mathcal{N}(w_s\boldsymbol{\mu}_{z(i),s}, \sigma_s^2\boldsymbol{I})$	Voxels share both signal scaling and noise parameters.

Table 1: Description of the variants of the GMMGP model considered here.

In order to analyze the influence and utility of the voxel specific parameters for scaling and noise we implement the corresponding models also without these parameters resulting in a total of four combinations of model specifications to be considered as described in Table 1. We compare the different model specifications on simulated data based on their ability to both infer the correct clustering configuration, the correct task activations and on their predictive likelihood on a test dataset and finally we compare the different versions on real fMRI data based on their predictive performance.

Combining the equations for the generative model provides an expression for the joint distribution of the parameters and the data. We are interested in the distribution of the parameters given the data but this expression is intractable analytically and instead we apply MCMC sampling techniques to generate samples from the posterior distribution. For each MCMC iteration we perform one

Gibbs sweep where we draw the clustering configuration for each voxel from the posterior conditional distribution of the voxel clustering followed by a number of split-merge proposals (Jain and Neal, 2004) and sample the hyper-parameters, $w_{i,s}, \sigma^2_{i,s}$, and γ, using Metropolis-Hastings proposals. A detailed derivation of the joint distribution and the posterior distributions for each of the parameters can be found in Appendix A. Since we cannot expect the inference method to fully converge within a limited number of MCMC iterations, we accept the highest posterior density sample as the final clustering solution.

The covariance matrix for the prior on the cluster mean time series, $\beta\mathbf{\Sigma}_{\text{SE}}$, i.e. of full rank, and computing the inverse thus requires $\mathcal{O}(T^3)$ time, where T is the dimensionality of the voxel time series or the number of brain volumes. Using the Cholesky factorization of $\mathbf{\Sigma}_{SE}$, it is possible to reduce this complexity to linear time as demonstrated in Røge et al. (2015). This means that the temporal complexity for each Gibbs sweep and the entire MCMC sampling procedure becomes $\mathcal{O}(NTKS)$, with N the number of voxels, K the number of components, and S the number of subjects. We find that implementations with at most this temporal complexity are necessary for sampling based inference in a Bayesian mixture model to be tractable at the scale of whole-brain modeling.

We compare the proposed mixture model with the traditional Bayesian Gaussian mixture model and four non-probabilistic clustering methods; the K-means clustering algorithm (Hartigan and Wong, 1979), Ward's algorithm for hierarchical agglomerative clustering (Ward Jr, 1963; Thirion et al., 2014), clustering based on region growing (Blumensath et al., 2013), and the Normalized cut clustering method (Shi and Malik, 2000; Craddock et al., 2012). Note, that Ward, region growing, and Ncut are all spatially constrained to contiguous clusters. Sampling based inference in a Bayesian mixture of Gaussians was explored in a univariate case by Richardson and Green (1997) and extended to an infinite number of mixtures by Rasmussen (1999). Note that inference in a Gaussian mixture model with full covariance is dependent on inverting a full rank covariance matrix and thus requires at minimum $\mathcal{O}(T^2)$ even when using the Cholesky factorization. Since this is not tractable we restrict the covariance maxtrix to a diagonal matrix, ie. $\boldsymbol{x}_i \sim \mathcal{N}(\mu_{z_i}, \text{diag}(\boldsymbol{\sigma}^2_{z_i}))$ and we refer to the infinite mixture model based on this model specification as the iGMMd model with the CRP prior or GMMd with the multivariate Pólya distribution as prior on the clustering. We furthermore implement a version where the covariance matrix is restricted to a diagonal matrix where the elements are identical, ie. $\boldsymbol{x}_i \sim \mathcal{N}(\mu_{z_i}, \sigma^2_{z_i}\boldsymbol{I})$ for the spherical Gaussian mixture model denoted GMMs or iGMMs.

At the website `https://brainconnectivity.compute.dtu.dk/` we will make the code and examples from this paper publicly available.

2.2. Evaluation of clusterings

We evaluate the performance of the clustering on synthetic data based on three criteria: Reproducibility of the models, the predictive likelihood on hold out data, and the ability to recover the cluster time series. On fMRI data there is no truth available regarding the cluster time series or a true clustering configuration; instead we assess the reproducibility of the clustering on different sets of subjects and we explore the effect of using the time series inferred by the model on both resting state and task fMRI data.

We measure the reproducibility of the model using the Adjusted Mutual Information (AMI) (Vinh et al., 2010) given by

$$\mathrm{AMI} = \frac{\mathrm{MI}(\boldsymbol{z}_1, \boldsymbol{z}_2) - E\{\mathrm{MI}(\boldsymbol{z}_1, \boldsymbol{z}_2)\}}{\max(H(\boldsymbol{z}_1), H(\boldsymbol{z}_2)) - E\{\mathrm{MI}(\boldsymbol{z}_1, \boldsymbol{z}_2)\}}, \tag{1}$$

where $\mathrm{MI}(\boldsymbol{z}_1, \boldsymbol{z}_2)$ is the mutual information between two clusterings $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ and $E\{\mathrm{MI}(\boldsymbol{z}_1, \boldsymbol{z}_2)\}$ is the expected mutual information for two random clusterings with the same number of clusters as $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$. For the analysis on simulated data we have ground truth available and therefore compute the AMI between the true and inferred clusterings. On the fMRI dataset we have no ground truth and instead evaluate the reliability of the clusters by the AMI between inferred clusterings on two different groups of subjects.

The predictive likelihood is the posterior distribution of the test data given the training data and is for our model specification not possible to evaluate analytically. We use a split-half analysis of each subject such that we are given a new dataset where we can assume to know the signal scaling and noise parameters. In this case we can approximate the predictive likelihood by

$$p(\boldsymbol{X}^* \mid \boldsymbol{X}) \approx \frac{1}{M} \sum_{m=1}^{M} p(\boldsymbol{X}^* \mid \boldsymbol{\theta}^{(m)}), \tag{2}$$

where $\boldsymbol{\theta}^{(m)}$ for $m = 1, \ldots, M$ are M samples from the posterior distribution for the parameters $w_{i,s}, \sigma^2_{i,s}$, γ, and $\boldsymbol{z}$.

The third metric of how well the modeling recovers truth is how closely the inferred parcel mean time series resemble that of the true parcels. In the case for synthetic data we can directly assess this by the maximum correlation between the posterior parcel time series with the task design time series used to generate the data but with real fMRI data we naturally have no such information. Instead we compare the results of SPM analysis on the parcellated data with a traditional voxel based SPM analysis.

The posterior distribution of the parcel time series can be computed analytically (details can be found in Appendix A),

$$p(\boldsymbol{\mu}_{k,s} \mid \boldsymbol{X}, \boldsymbol{\theta}) \sim \mathcal{N}\left(\left[\boldsymbol{\Sigma}_{\mathrm{SE}}^{-1} + \sum_{i \in \mathcal{Z}_k} \frac{w^2_{i,s}}{\sigma^2_{i,s}} \boldsymbol{I}\right]^{-1} \bar{\boldsymbol{x}}_k, \left[\boldsymbol{\Sigma}_{\mathrm{SE}}^{-1} + \sum_{i \in \mathcal{Z}_k} \frac{w^2_{i,s}}{\sigma^2_{i,s}} \mathbf{I}\right]^{-1}\right) \tag{3}$$

where $\boldsymbol{\theta} = \{\gamma, \{w\}, \{\sigma^2\}, \boldsymbol{z}\}$, $\mathcal{Z}_k = \{i \in 1, \ldots, N : z_i = k\}$ is the index set of observations in cluster k, and $\bar{\boldsymbol{x}}_k = \sum_{i \in \mathcal{Z}_k} \frac{w_{i,s}}{\sigma^2_{i,s}} \boldsymbol{x}_{i,s}$. The posterior mean is therefore $\left[\boldsymbol{\Sigma}_{\mathrm{SE}}^{-1} + \sum_{i \in \mathcal{Z}_k} \frac{w^2_{i,s}}{\sigma^2_{i,s}} \boldsymbol{I}\right]^{-1} \bar{\boldsymbol{x}}_k$. From this expression we observe that the mean is a sum of the voxel time series in the cluster weighted by $w_{i,s}/\sigma^2_{i,s}$. Voxels with increased noise or voxels that do not match the cluster time series will thus be down-weighted. Furthermore, the sum is smoothened with $(\mathbf{\Sigma}_{\text{SE}}^{-1} + \sum_{i\in\mathcal{Z}_k} \frac{w^2_{i,s}}{\sigma^2_{i,s}}\boldsymbol{I})^{-1}$ which mean that the smoothing kernel $\mathbf{\Sigma}_{\text{SE}}$ is balanced against the sum and large clusters will thus be smoothened less compared to small clusters. Both these effects contribute to recovering more accurate cluster time series.

3. Data

3.1. Simulated data

With the hypothesis that it is beneficial to model the scaling of the signal and the noise across voxels and subjects we examine the effect of data generated with those effects and compare the results with the traditional clustering methods.

We simulated 2D brain slices consisting of four square 10×10 clusters with 100 time points for four subjects for both a training and a test dataset. For each cluster we draw a task activation from $\boldsymbol{\mu}_k \sim \mathcal{N}(0, 5\boldsymbol{I})$, where $\boldsymbol{I}$ is the 100 dimensional identity matrix and convolve the task activation with the Hemodynamic Response Function, $\boldsymbol{h}$, with TR$= 2.49s$ to get the expected task signal, $\tilde{\boldsymbol{\mu}}_k = \boldsymbol{\mu}_k * \boldsymbol{h}$. For each voxel we scale the corresponding task time-course with $w_{i,s}$. Subsequently, we add the same random Gaussian vector, $\boldsymbol{\epsilon}_0 \sim \mathcal{N}(0, \boldsymbol{I})$, to all clusters to introduce a non-smooth artifact into the clusters and draw the time series with white noise $\boldsymbol{\epsilon}_{i.s} \sim \mathcal{N}(0, \boldsymbol{I})$ that is subsequently scaled to have standard deviation λ. Finally we introduce regions with a five-fold increase in the noise denoted by a set of indices for observations A_{noise} and let $\delta_{i,A_{\text{noise}}}$ be 1 if $i \in A_{\text{noise}}$ and zero otherwise. The process for generating vector $\boldsymbol{x}_{i,s}$ can be summarized by the following expression, i.e.,

$$\boldsymbol{x}_{i,s} = w_{i,s}\tilde{\boldsymbol{\mu}}_{z_i} + \boldsymbol{\epsilon}_0 + \lambda(1 + 4\delta_{i,A_{\text{noise}}})\boldsymbol{\epsilon}_{i,s}. \tag{4}$$

The generated dataset is illustrated in the top panel of Fig. 2. In order to validate our assumptions on the model, i.e. that it is beneficial to model changes in signal and noise across the brain, we generate 10 datasets for each level of noise as controlled by the λ parameter ($\lambda = \{1, 3, \ldots, 37, 39\}$) in Eq. 4. For each generated dataset we also generate a test dataset used for predictive comparison of the proposed models.

3.2. fMRI datasets

We apply the model to a finger tapping task and resting state fMRI datasets with the same 30 healthy subjects scanned on a 3T MRI scanner, and of those 30 subjects we use the first 20. The rs-fMRI consists of 480 brain volumes scanned over 20 minutes with a repetition time (TR) of 2.49s and further details can be found in Andersen et al. (2014). The task dataset consists of 240 brain volumes scanned over 10 minutes, also with a TR of 2.49s and has previously been used in Rasmussen et al. (2012b,a); Røge et al. (2015).

The datasets were preprocessed using the SPM12 software package (SPM12, Wellcome Trust Centre for Neuroimaging, `http://www.fil.ion.ucl.ac.uk/spm/soft-ware/spm12/`) with Rigid body realignment, co-registration, spatial normalization to the MNI 152 template, reslicing of images into MNI space at 3 mm isotropic voxels, spatial smoothing with a 5 mm FWHM isotropic Gaussian filter, a high-pass filter (128 Hz), and finally a rough grey matter mask consisting of 48799 voxels for the rs-fMRI and 44820 for the task fMRI was applied.

We split the subjects into two groups using the first 10 for the first group and the 10 following for the second group. The time series were then split into a training and testing set to consisting of 240 time points for the rs-fMRI and 120 time points for the task fMRI dataset. This means we have 2 datasets for training and 2 for testing for both the rs-fMRI and for the task fMRI datasets.

For the task dataset the stimulation cycle of 20 s right handed finger tapping, 10 s rest, 20 s left handed finger tapping, and 10 s rest was repeated 10 times. The finger tapping tasks were assisted by a visual cue.

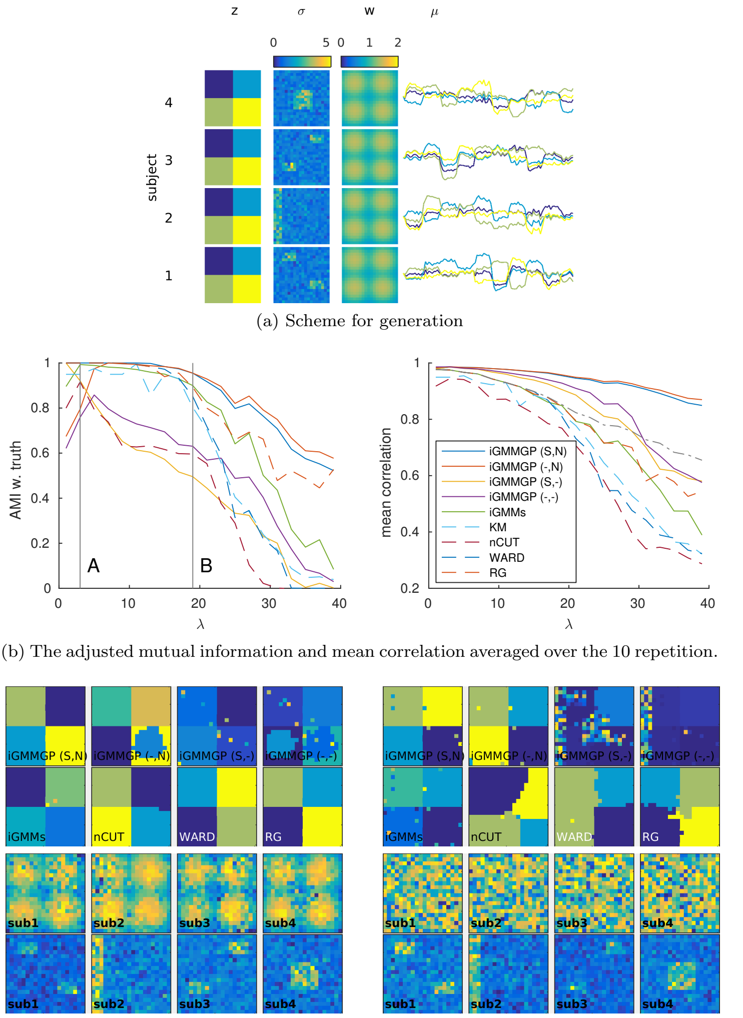
4. Results and discussion

In theory the procedure for initizalizing the parameters of the model is of little importance since, given enough samples, the Markov chain will approach the posterior distribution. In practice, given the size of the problem, we are limited to only a few hundred samples and it is very important that the state of the model for the initialization is supported by the posterior distribution. Therefore, we initialize the clustering configuration of the model using the K-means clustering algorithm and then evaluate 100 Metropolis-Hastings proposals for the other parameters of the models. In order to not initialize the model to a local minimum we then set the clustering configuration to a random clustering before starting the actual inference procedure.

The scale of the problem makes it impossible to generate a number of samples from the posterior distribution sufficient to reach the stationary distribution. However, this does not prevent the sampling procedure from reaching good solutions and we find that the sampler rather quickly converges to a solution. On synthetic data this typically happens within a few iterations following the described initialization procedure, and for the fMRI datasets we find that around 20 MCMC iterations provides good solutions. For the experiments in this contribution we allowed the inference chains to run for 100 MCMC iterations and select the highest likelihood sample to use for further analysis. and after 100 MCMC iterations to be sufficient. On a 2.4 GHz core i7 processor the duration for each MCMC iteration consiting of one Gibbs sweep, evaluating as many split-merge proposals as there are cluster, and Metropolis-Hastings proposals for hyperparameters was approximately 1 hour on a fMRI dataset with 10 subjects, 240 brain volumes and 44820 voxels.

4.1. Simulated analysis

For each generated dataset we perform model inference with the four versions of the GMMGP models for both the parametric and nonparametric versions.



(a) Scheme for generation

(b) The adjusted mutual information and mean correlation averaged over the 10 repetition.

(c) Clustering for all models and signal and noise parameters for iGMMGP(S,N) at $\lambda = 3$ as marked by **A**.

(d) Clustering for all models and signal and noise parameters for iGMMGP(S,N) at $\lambda = 19$ as marked by **B**.

Figure 2: Comparison of different nonparametric clustering models on the *simple* synthetic dataset. The clusterings presented are from the first repetition at the level of noise designated by the vertical line in the top left plot. The grey dashed line in the top right plot denotes the correlation using the average voxel time courses of the correct clustering. The bottom line is the inferred signal and noise parameters for the iGMMGP(S,N) model averaged over 10 samples.

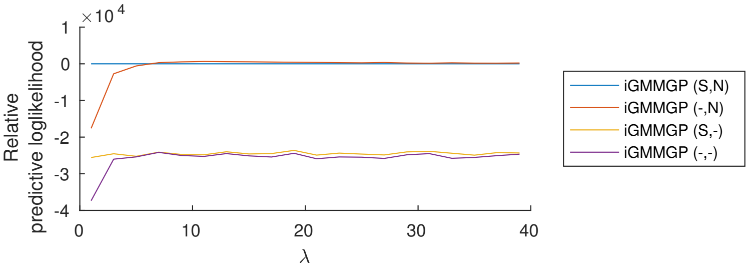


Figure 3: Predictive results on synthetic analysis. The predictive logjlikelihood is reported relative to the iGMMGP (S,N) model.

We allow each model to perform 100 MCMC iterations and select the sample attaining the highest value of the joint distribution. We evaluate the performance based on AMI with ground truth, predictive posterior probability, and in terms of the mean correlation with the simulated BOLD signal, $\tilde{\boldsymbol{\mu}}_k$. We compare the GMMGP model with the nonprobabilistic clustering models and the traditional Gaussian mixture models (both parametric and nonparametric) with either spherical or elliptical (i.e. diagonal) covariance structure on the data and on a version of the data where the time series of each voxel is normalized to have unit length. In Fig. 2 we highlight the performance of the non-parametric model specifications and the nonprobabilistic clustering models. We only include the best performing of the traditional Gaussian mixture models which is is the version with spherical noise using the normalized dataset. Each line in Fig. 2b is the average AMI and average mean correlation over the 10 repetitions. The AMI for the remaining probabilistic models are given in Appendix B. In Fig. 3 we present the predictive comparison of the 4 versions of the GMMGP models.

From the results we see that the proposed iGMMGP model is able to use the temporal dynamics imposed by the Gaussian Process to better extract the underlying simulated BOLD signal compared to just averaging the true parcels or using conventional clustering. The models with the Gaussian Process prior are thus significantly better at both recovering the true clustering and infers cluster means that are better correlated with the simulated BOLD responses. We further observe that the models without heteroscedastic modeling of noise cannot handle areas of increased noise. In the bottom panel of Fig. 2 we further see that the fully heteroscedastic model is able to correctly identify the regions of increased noise. Inspecting the predictive performance of the four model specifications in Fig. 3 we observe that both models with heteroscedastic modeling of noise are approximately equivalent in predictive performance and we also here observe that the heteroscedastic signal modeling is only beneficial in regimes with high SNR, i.e. low values of λ. While it is somewhat surprising heteroscedastic signal modeling has no benefit facing low SNR we attribute this to the high level of noise making the differences in the heteroscedasticity of the signal off less importance.

4.2. Clustering performance on fMRI datasets

We first establish which of the four variants of the iGMMGP model, i.e., with and without heteroscedastic modeling of signal and noise, best account for the structure in real fMRI time data. For this purpose we use the predictive likelihood and split each subjects data into two halfs such that we use the first half for training and the second half for testing. For each of the datasets we perform 100 full MCMC iterations and select the highest likelihood sample. In Fig. 4 we plot the predictive likelihood as function of the number of clusters, for five repetitions of this inference process. The fully heteroscedastic model (i.e., iGMMGP(S,N)) finds the most compact representation of the data while still yielding the best predictive likelihood on all the considered datasets. This supports our initial hypothesis and confirms the results from Churchill et al. (2016) where similar modeling of the signal and noise was also found to improve on generalizability. For brevity and clarity of results we therefore choose to only continue with the full version of the model, i.e. using the iGMMGP(S,N).

In Fig. 5 we visualize the inferred parameters of the model. In the figure is shown the extracted clustering and the inferred $\{\sigma\}$ and $\{w\}$ parameters for axial brain slices using the highest likelihood sample of the MCMC inference chain based on the rs-fMRI dataset with subjects 1 to 10 . We see from the highlighted areas that the model finds that areas near the rim closest to the skull have high noise magnitudes such that the influence from these regions in defining clusters are reduced. From the plot of the average signal to noise ratio it is clear that averaged over all subjects the model assigns most probability to the center of each cluster.

With the selected version of the iGMMGP model we compare the clustering performance to that of the spherical Gaussian mixture model (GMMs) and the elliptical Gaussian mixture model (GMMd) based on reproducibility of clustering over different groups of subjects. We perform 100 MCMC iterations with both the finite and infinite versions of the GMMs and GMMd models on the fMRI datasets and present the Adjusted Mutual Information between the two groups of subjects in Fig. 6. Note that we cannot directly compare the AMI of clustering solutions inferred by the infinite versions of the models since the inferred number of clusters will vary from around 8-900 for the iGMMGP models to 1500-2300 for the two traditional Gaussian mixture model. Furthermore we cannot compare the models based on predictive likelihood either since the iGMMGP model due to the Gaussian Process (GP) prior is tailored to focus on the temporally smooth part of the signal and will therefore be inferior in modeling high frequency content present in the data.

4.3. Parcellation effect on reliability of task activations

In order to investigate the hypothesis, that the inferred parcel time series more accurately reflect the changes in the BOLD signal due to neural activity

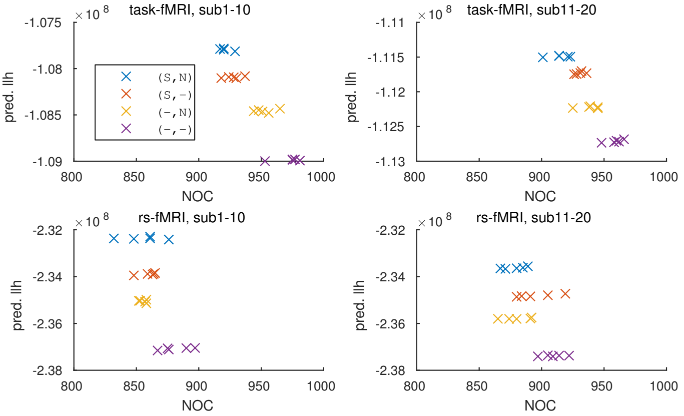


Figure 4: Predictive likelihood as a function of the number of clusters for each of the 5 repetitions on the considered datasets. Note that the considered task and resting state fMRI datasets are divided into subjects 1-10 and 11-20. The most complex model with heteroscedastic modeling of both signal and noise, i.e. iGMMGP(S,N), is marked with blue crosses and it provides both the most compact representation and the best predictive performance.

compared to the voxel time series, we compare the regions of task activation on a temporal split-half on the finger tapping task fMRI dataset. We therefore apply the statistical model to get parcellations on the training and test split of the dataset as previously described, i.e. the data is divided into two groups of 10 subjects and each subject is temporally split into a test and training dataset. Then we apply the model to each of the four datasets to get four independent parcellations, where each parcellation is trained on 10 subjects.

We use SPM12 to perform a GLM (Friston et al., 1994) analyses to estimate the task response by fitting a design matrix with autoregressive AR(1) filtering and the task regressors are extracted on both the original data and on the parcelled data using the estimated posterior cluster mean time courses. The left and right handed finger tapping tasks were modelled with a boxcar regressor convolved with the canonical hemodynamic response function. We then applied the two contrasts Right-Left and Left-Right to get the contrast maps. The thresholded activation maps were computed corrected by family wise error (FWE) with a threshold set to 0.05 for the voxel based SPM analysis. For the parcel based analysis we used Bonferroni correction to account for multiple comparisons as defined by the number of parcels. The tresholded activation maps for the voxel and parcel based analysis for the first subject along with the parcel and voxel time series with the highest Z-score are presented in Fig. 7.

The regions of task activation are considerably larger and the maximal Z-score is considerably higher for the parcel based compared to the voxel based analysis and this is in general the case. Furthermore, the posterior mean parcel

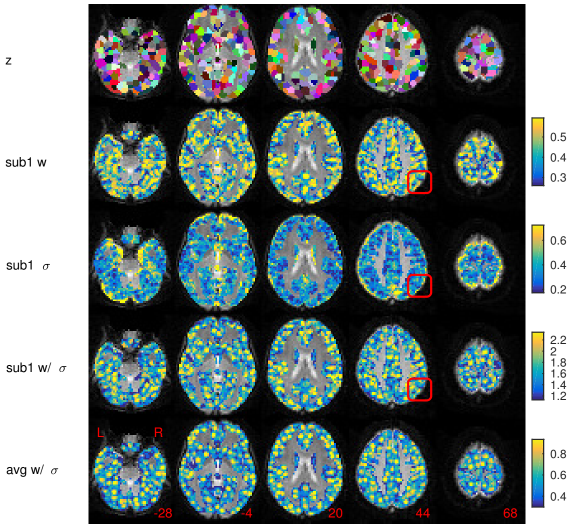


Figure 5: In the top panel is a visualization of the clustering, and in panel 2-4 are respectively the signal, noise and signal to noise ratio for subject 1. Looking at the red rectangle we see that the noise modeling infers high noise along the edge of the grey matter thereby reducing its influence on the parcel mean time series. In the bottom panel are the signal to noise ratio averaged over all subjects and it is clear that on average the center of the clusters are most important in determining the parcel mean time series.

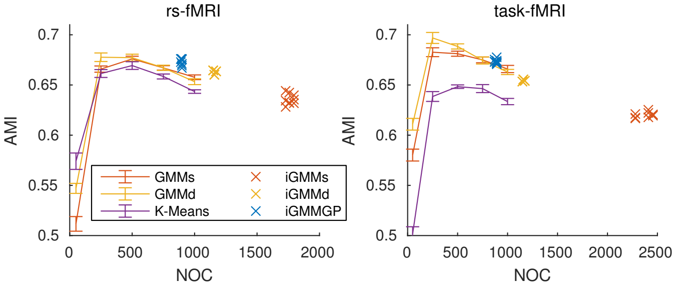


Figure 6: A comparison of the adjusted mutual information between parcellations of subjects 1-10 and 11-20 for the rs-fMRI and task-fMRI datasets inferred by the finite versions of the two Gaussian mixture models with the infinite GMMGP and GMM. Note that it is difficult to compare AMI for clusterings with approximately 900 clusters to one with more than 2000 clusters we have therefore included for comparison also the finite GMM.

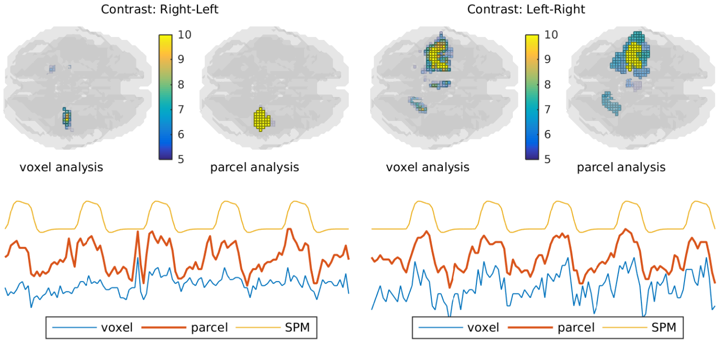


Figure 7: Comparison between the results from voxel and parcel based SPM analysis on the first subject of the Hvidovre task dataset. The time series of the most significant voxel time series are presented in blue and the most significant parcel time series in red along with the design matrix for the corresponding task.

time series is clearly smoothed by the Gaussian Process prior.

We compute the Dice score between activation maps from the training and test temporal split-half and to illustrate the difference in inferred activated regions we plot the activated regions colored by the number of times a voxel or cluster was activated in Fig. 8. Furthermore, we perform two additional SPM analyses where only the brain volumes from the first two task repetitions were included and present these results in the right column of Fig. 8. Note that if there is no activation or no overlapping activation between the training and test activation maps the Dice score is set to zero penalizing the result. This especially affects the results from the SPM analysis including only two repetitions of the task causing the high variance in the reported Dice score.

Using a paired t-test we find that the inferred regions of task activation are significantly more reliable for the Left-Right contrast using all five repetitions ($p< .5$) while the difference is not significant for the Right-Left contrast. Using only two task repetitions, the reliability advantage is significant for both contrasts. This suggests that using the posterior mean time series for SPM analyses increase the reliability and sensitivity to task activation.

The Dice score depends on the level of activation, thus, the Dice score would be 1 for a model where the entire brain is active. This is especially problematic when comparing two methods of inference where one method with higher sensitivity identifies a larger part of the brain as active. In order to account for this phenomenon we also compute the Dice score as a function of the percentage of activated voxels, see Fig. 9.Using two repetitions of the task, the dice curves for the voxel based analyses are dominated by the parcel based analysis. Using all five repetitions of the task the same is the case when more than 0.005 and

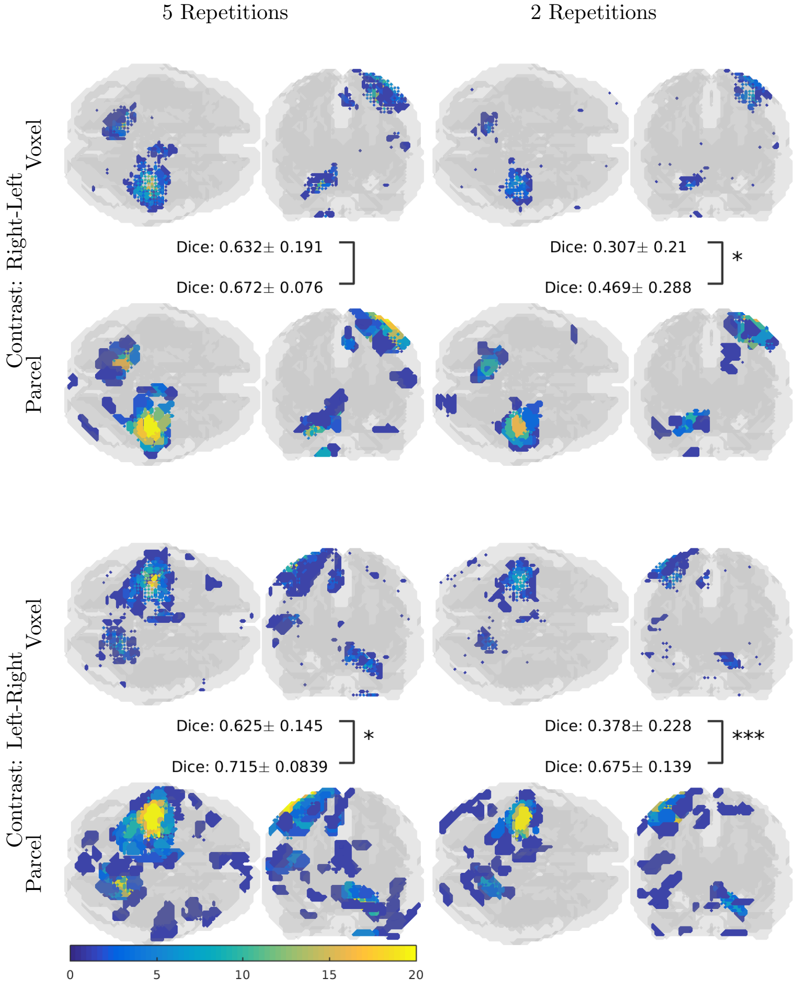


Figure 8: The regions of inferred task activation for the traditional voxel and parcel based SPM analyses colored by the number of times each voxel had significant task activation. The paired t-test show that the parcel based analyses is significantly more reliable for 3 of the 4 comparisons.

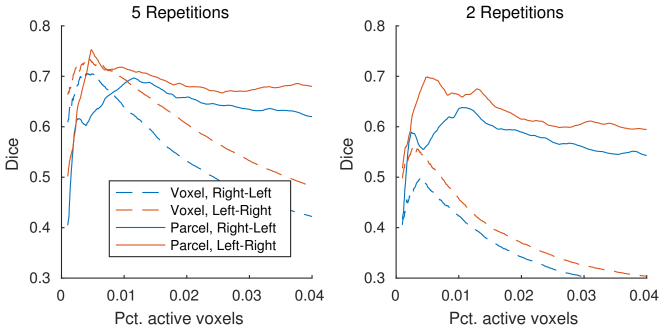


Figure 9: The Dice score as a function of the percent activated voxels of the brain for each of the two contrasts. In the left panel the brain volumes from all 5 repetitions is used while in right panel we use only those of the first two repetitions.

0.01 pct. voxels are active for the two contrasts. The low Dice scores for the parcel based analysis for the small percentages of active voxels is due to the fact that the Dice score is limited by how closely the two parcellations overlap in the active regions.

4.4. Parcellation effect on resting state networks

We finally investigate the effect of a parcellation based analysis on the reliability of functional resting state networks. Two brain regions known for a task negative correlation (Fox et al., 2005) are the posterior cingulate cortex (PCC; -5, -49, 40) and medial prefrontal cortex (MPF; -1, 47, -4) and we use 6 mm spheres around the two region of interest for regressors in both voxel and parcel based SPM analyses.

In order to compute the regressor for the parcel based analysis we reconstruct the voxel time series of the voxels within the sphere such that the time series of each voxel will be the posterior parcel time series weighted with $w_{i,s}$ and then compute the regressor as for the voxel based analysis. The SPM analysis was then performed using the computed regressors for the MPF and PCC with autoregressive AR(1) filtering. For each of the two regions of interest, we perform a 2nd level analysis using the contrast maps from the 1st level SPM analysis for each subject as input. The thresholded 2nd level Z-maps are presented Fig. 10 and from a visual comparison it is difficult to distinguish parcel and voxel based analyses.

Next we compute the Dice score as a function of the number of *active* voxels, similarly to the analysis of task reliability. The Dice score for the parcel based analyses generally dominates the curves for the voxel based analyses, see Fig. 11. Here, we only include the positively correlated voxels for the analysis. Before computing the Dice score we exclude a 12-mm-radius sphere around the seed region from the active region to not bias the curve by the seed region. The point

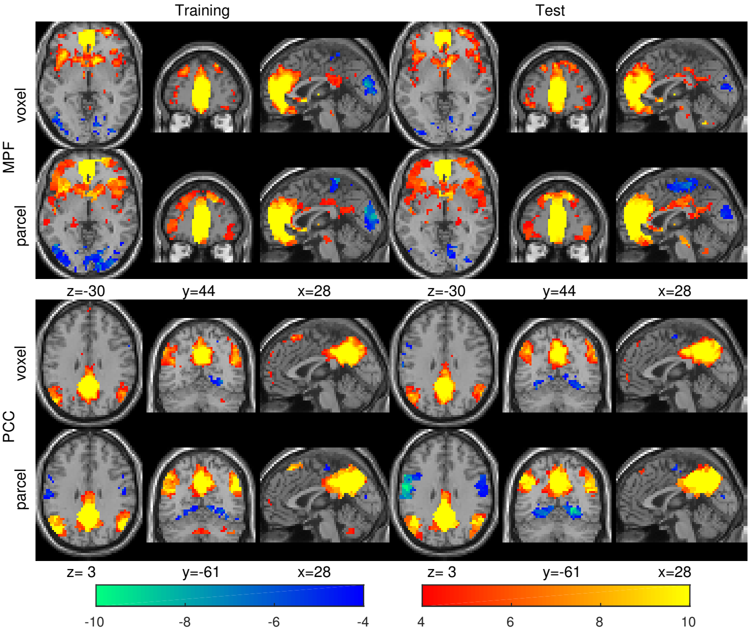


Figure 10: Comparison of the thresholded second level Z-statistic (Z>4, $p< 1.67\cdot10^{-4}$) images from the voxel and parcel time series GLM. For the voxel based analysis we used the mean of the 6mm radii spheres around the medial prefrontal cortex (MPF) and posterior cingulate cortex (PCC) was as regressors and for the parcel based analysis we used the posterior mean time series of the closest parcel.

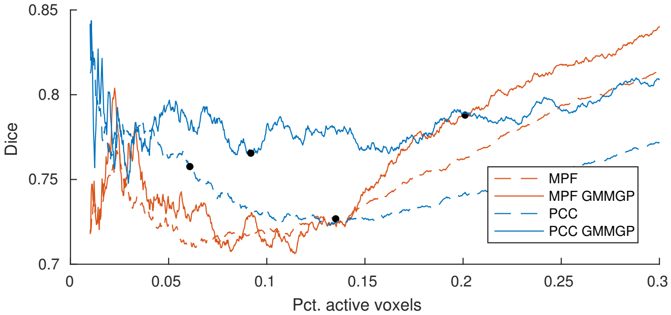


Figure 11: The Dice score as a function of the pct. active voxels in the second level analysis. Note that the parcel based analyses are consistently more reliable except when only a very small number of voxels are active. The black dots on the curve correspond to the level of activation and Dice score for the tresholds used in Fig. 10.

on each curve marked by a black dot shows the threshold corresponding to the threshold selected for the visualization in Fig. 10.

For this illustration we analysed the reliability of connectivity of two task negative regions, PCC and MPF. Both regions show increased reliability of correlations to the rest of the brain when splitting the time series of each voxel in a test and retest dataset. This shows promise in improving the reliability of network analysis using the iGMMGP model representation for deriving functional networks.

5. Conclusion

We have presented the iGMMGP, a probabilistic method for whole-brain parcellation. The model exploits Bayesian non-parametrics to automatically quantify the number of parcels needed to describe the fMRI data thereby not requiring the number of clusters to be specified apriori. The model further uses Gaussian Process priors to incorporate smoothness into the extracted cluster time series in order to focus on the part of the fMRI signal that is caused by slowly varying changes in cerebral blood flow. Finally, voxel and subject specific signal and noise scaling parameters are used to account for changes across space and subjects making the model robust to noise and model misspecification while at the same time providing a more compact representation of the data in terms of parcels constituting salient functional units.

In comparison with traditional clustering methods, the proposed IGMMGP method is more reliable on both synthetic, rs-fMRI and task-fMRI datasets. Compared to traditional SPM analysis of task activation, we show that using the method to denoise the data the model provides a compact representation having increased reliability both when estimating regions of task activation and when estimating functional networks compared to conventional voxel based analyses. This suggests the utility of using the parcellation method as a starting point for future analysis of fMRI data.

The approach can in general identify salient functional units in fMRI data at a group level and is robust to model misspecification as it accounts for subject and voxel specific noise and signal fluctuations. Compared to traditional atlas based approaches the iGMMGP being data driven define parcels in order to optimally account for the distinct functional patterns in the fMRI data. Similar to atlas based approaches the parcel based analysis can in general address the issue facing voxel based analysis of multiple comparisons by substantially reducing the number of regions providing a compact and noise reduced representation that compared to atlas based approaches are tailored to the data. We presently considered a radial basis function kernel with length scale defined by the canonical hemodynamic response function when specifying the covariance. On synthetic data we observed that the GP prior had a significant effect on the ability to extract the underlying generated signals. Notably, the covariance can be further tailored to the assumptions of fMRI and the iGMMGP framework readily generalizes to other covariance specifications that may better account for the assumptions of the fMRI data at hand.

Aldous, D. J., 1985. Exchangeability and related topics. 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.

Baldassano, C., Beck, D. M., Fei-Fei, L., 2015. Parcellating connectivity in spatial maps. PeerJ 3, e784.

Blumensath, T., Behrens, T. E. J., Smith, S. M., 2012. Resting-state fmri single subject cortical parcellation based on region growing. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 188–195.

Blumensath, T., Jbabdi, S., Glasser, M. F., Van Essen, D. C., Ugurbil, K., Behrens, T. E. J., Smith, S. M., 2013. Spatially constrained hierarchical parcellation of the brain with resting-state fmri. Neuroimage 76, 313–324.

Buxton, R. B., Wong, E. C., Frank, L. R., 1998. Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. Magnetic resonance in medicine 39 (6), 855–864.

Calhoun, V. D., Adali, T., Pearlson, G. D., Pekar, J. J., 2001. A method for making group inferences from functional mri data using independent component analysis. Human brain mapping 14 (3), 140–151.

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.

Craddock, R. C., James, G. A., Holtzheimer, P. E., Hu, X. P., Mayberg, H. S., aug 2012. A whole brain fMRI atlas generated via spatially constrained spectral clustering. Human brain mapping 33 (8), 1914–28.

Da Mota, B., Fritsch, V., Varoquaux, G., Banaschewski, T., Barker, G. J., Bokde, A. L. W., Bromberg, U., Conrod, P., Gallinat, J., Garavan, H., et al., 2014. Randomized parcellation based inference. NeuroImage 89, 203–215.

Eklund, A., Nichols, T. E., Knutsson, H., 2016. Cluster failure: Why fmri inferences for spatial extent have inflated false-positive rates. Proceedings of the National Academy of Sciences, 201602413.

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., Raichle, M. E., 2005. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. Proceedings of the National Academy of Sciences of the United States of America 102 (27), 9673–9678.

Friston, K. J., Harrison, L., Penny, W., 2003. Dynamic causal modelling. Neuroimage 19 (4), 1273–1302.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., Frackowiak, R. S. J., 1994. Statistical parametric maps in functional imaging: a general linear approach. Human brain mapping 2 (4), 189–210.

Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S., Turner, R., mar 1996. Movement-related effects in fMRI time-series. Magn Reson Med 35, 346–355.

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., et al., 2016. A multi-modal parcellation of human cerebral cortex. Nature.

Glover, G. H., Li, T. Q., Ress, D., jul 2000. Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. Magn Reson Med 44, 162–7.

Gorgolewski, K. J., Storkey, A., Bastin, M. E., Whittle, I. R., Wardlaw, J. M., Pernet, C. R., 2013. Single subject fmri test–retest reliability metrics and confounding factors. Neuroimage 69, 231–243.

Goutte, C., Nielsen, F. A., Hansen, L. K., 2000. Modeling the hemodynamic response in fmri using smooth fir filters. IEEE transactions on medical imaging 19 (12), 1188–1201.

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

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

Hinrich, J. L., Bardenfleth, S., Røge, R. E., Churchill, N. W., Madsen, K. H., Mørup, M., 2016. Archetypal analysis for modeling multi-subject fmri data. IEEE Journal of Selected Topics in Signal Processing 10 (7).

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).

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.

Josef, W., Villani, M., Eklund, A., 2016. Physiologically movivated gaussian process priors for the hemodynamics in fmri. OHBM 2016, poster 3783. URL `https://ww5.aievolution.com/hbm1601/files/content/abstracts/41188/3783_Wilzn_0627_105211.pdf`

Kaza, E., Klose, U., Lotze, M., 2011. Comparison of a 32-channel with a 12-channel head coil: Are there relevant improvements for functional imaging? Journal of Magnetic Resonance Imaging 34 (1), 173–183.

Lashkari, D., Golland, P., 2009. Exploratory fmri analysis without spatial normalization. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 398–410.

Lund, T. E., Madsen, K. H., Sidaros, K., Luo, W.-L., Nichols, T. E., 2006. Non-white noise in fMRI: does modelling have an impact? NeuroImage 29 (1), 54–66.
URL `http://www.ncbi.nlm.nih.gov/pubmed/16099175`

McKeown, M. J., Makeig, S., Brown, G. G., Jung, T.-P., Kindermann, S. S., Bell, A. J., Sejnowski, T. J., 1997. Analysis of fmri data by blind separation into independent spatial components. Tech. rep., DTIC Document.

Mikl, M., Mareček, R., Hluštík, P., Pavlicová, M., Drastich, A., Chlebus, P., Brázdil, M., Krupa, P., 2008. Effects of spatial smoothing on fmri group inferences. Magnetic resonance imaging 26 (4), 490–503.

Mørup, M., Madsen, K., Dogonowski, A.-M., Siebner, H., Hansen, L. K., 2010. Infinite relational modeling of functional connectivity in resting state fmri. In: Advances in neural information processing systems. pp. 1750–1758.

Rasmussen, C. E., 1999. The infinite gaussian mixture model. In: NIPS. Vol. 12. pp. 554–560.

Rasmussen, P. M., Abrahamsen, T. J., Madsen, K. H., Hansen, L. K., 2012a. Nonlinear denoising and analysis of neuroimages with kernel principal component analysis and pre-image estimation. NeuroImage 60 (3), 1807–1818.

Rasmussen, P. M., Hansen, L. K., Madsen, K. H., Churchill, N. W., Strother, S. C., 2012b. Model sparsity and brain pattern interpretation of classification models in neuroimaging. Pattern Recognition 45 (6), 2085–2100.

Richardson, S., Green, P. J., 1997. On bayesian analysis of mixtures with an unknown number of components (with discussion). Journal of the Royal Statistical Society: series B (statistical methodology) 59 (4), 731–792.

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

Ross, J. C., Dy, J. G., 2013. Nonparametric mixture of gaussian processes with constraints. In: ICML (3). pp. 1346–1354.

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.

Shen, X., Tokoglu, F., Papademetris, X., Constable, R. T., 2013. Groupwise whole-brain parcellation from resting-state fmri data for network node identification. Neuroimage 82, 403–415.

Shi, J., Malik, J., 2000. Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence 22 (8), 888–905.

Talairach, J., Tournoux, P., 1988. Co-planar stereotaxic atlas of the human brain. 3-dimensional proportional system: an approach to cerebral imaging.

Thirion, B., Flandin, G., Pinel, P., Roche, A., Ciuciu, P., Poline, J.-B., 2006. Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fmri datasets. Human brain mapping 27 (8), 678–693.

Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S., Poline, J.-B., 2007. Analysis of a large fmri cohort: Statistical and methodological issues for group analyses. Neuroimage 35 (1), 105–120.

Thirion, B., Varoquaux, G., Dohmatob, E., Poline, J., 2014. Which fmri clustering gives good brain parcellations? Frontiers in neuroscience 8, 167.

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 (Oct), 2837–2854.

Ward Jr, J. H., 1963. Hierarchical grouping to optimize an objective function. Journal of the American statistical association 58 (301), 236–244.

Wiggins, G. C., Polimeni, J. R., Potthast, A., Schmitt, M., Alagappan, V., Wald, L. L., 2009. 96-channel receive-only head coil for 3 tesla: Design optimization and evaluation. Magnetic resonance in medicine 62 (3), 754–762.

Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., et al., 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. Journal of neurophysiology 106 (3), 1125–1165.

Zalesky, A., Fornito, A., Cocchi, L., Gollo, L. L., Breakspear, M., 2014. Time-resolved resting-state brain networks. Proceedings of the National Academy of Sciences 111 (28), 10341–10346.

Appendix A. Inference

The generative model is given by

$$\begin{aligned}
\boldsymbol{z} &\sim CRP(\gamma) && \text{groups,} && \text{(A.1)}\\
\boldsymbol{\mu}_{k,s} &\sim GP(0, \beta\boldsymbol{\Sigma}_{\text{SE}}) && \text{group time series,} && \text{(A.2)}\\
\boldsymbol{x}_{i,s} &\sim \mathcal{N}(w_{i,s}\boldsymbol{\mu}_{z(i),s}, \sigma_{i,s}^2\boldsymbol{I}) && \text{voxel time series,} && \text{(A.3)}
\end{aligned}$$

For a finite model the Chinese Restaurant Process is replaced by the Pólya or marginalized Dirichlet-multinomial distribution, with parameters $\alpha_k = \alpha/K$ for all $k = 1, \ldots, K$ where K is the number of clusters, given by

$$p(\boldsymbol{z} \mid \alpha) = \int p(z_i \mid \boldsymbol{\pi}) p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) d\boldsymbol{\pi} = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \alpha/K)}{\Gamma(\alpha/K)}, \tag{A.4}$$

For the following we use the CRP prior but the calculations for the Pólya distribution is equivalent. Multiplying expressions (A.1)-(A.3) together we get the joint probability

$$\begin{aligned} & p(\boldsymbol{z}, \{\boldsymbol{\mu}_s\}, \{\boldsymbol{X_s}\} \mid \{\boldsymbol{\sigma}_s^2\}, \{\boldsymbol{w}_s\}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \gamma) \\ & \qquad = \left[\prod_s p(X_s | \boldsymbol{\mu}_s, \boldsymbol{w}_s, \boldsymbol{\sigma}_s^2) p(\boldsymbol{\mu}_s | \boldsymbol{\beta}, \boldsymbol{\Sigma})\right] p(\boldsymbol{z}|\gamma). \end{aligned} \tag{A.5}$$

We rearrange and marginalize over the nuisance parameters $\boldsymbol{\mu}_s$ using conjugacy of the Gaussian Process:

$$\begin{aligned} p(\boldsymbol{z}, \{\boldsymbol{X}_s\} \mid \{\boldsymbol{\sigma}_s^2\}, \{\boldsymbol{w}_s\}, \boldsymbol{\Sigma}, \gamma) &= \prod_s \left[\int p(X_s | \boldsymbol{\mu}_s, \boldsymbol{w}_s, \boldsymbol{\sigma}_s^2) p(\boldsymbol{\mu}_s | \boldsymbol{\Sigma}) d\boldsymbol{\mu}_s\right] p(\boldsymbol{z}|\gamma)) \\ &= \frac{\Gamma(\gamma)\gamma^K \prod_{k=1}^K \Gamma(n_k)}{\Gamma(N+\gamma)} \prod_{i,s} (2\pi\sigma_{i,s}^2)^{-T/2} \exp\left\{-\frac{1}{2\sigma_{i,s}^2} \boldsymbol{x}_{i,s}^\top \boldsymbol{x}_{i,s}\right\} \\ & \prod_{k=1}^K (|\boldsymbol{S}_{k,s}|/|\beta_s \boldsymbol{\Sigma}|)^{-1/2} \exp\left\{\frac{1}{2} \sum_{k=1}^K \bar{\boldsymbol{x}}_{k,s}^\top \boldsymbol{S}_{k,s}^{-1} \bar{\boldsymbol{x}}_{k,s}\right\}, \end{aligned} \tag{A.6}$$

where

$$\bar{\boldsymbol{x}}_{k,s} = \sum_{Z(i)=k} \frac{w_{i,s}}{\sigma_{i,s}^2} \boldsymbol{x}_{i,s}, \quad \boldsymbol{S}_{k,s} = \left(\beta_s \boldsymbol{\Sigma}^{-1} + \sum_{\boldsymbol{z}(i)=k} \frac{w_{i,s}^2}{\sigma_{i,s}^2} \boldsymbol{I}\right). \tag{A.7}$$

To get from the marginalized joint distribution to the posterior distribution for $\boldsymbol{z}_i$ we use Bayes theorem. This expression is a categorical distribution over the possible cluster assignments, which is all the populated clusters and one unpopulated cluster. The posterior probability to assign $\boldsymbol{z}_i$ to the populated cluster k is:

$$\begin{aligned} & p(\boldsymbol{z}_i = k \mid \boldsymbol{X}, \boldsymbol{\Sigma}, \{\sigma_n\}, \{w_n\}, \boldsymbol{z}) = \\ & n_k \frac{|\boldsymbol{S}_{k,s} + \frac{w_{i,s}^2}{\sigma_{i,s}^2} \boldsymbol{I}|^{-1/2} \exp\left\{\frac{1}{2}(\bar{\boldsymbol{x}}_{k,s} + \frac{w_{i,s}}{\sigma_{i,s}^2} \boldsymbol{x}_i)^\top \left(\boldsymbol{S}_{k,s} + \frac{w_{i,s}^2}{\sigma_{i,s}^2} \boldsymbol{I}\right)^{-1} (\bar{\boldsymbol{x}}_{k,s} + \frac{w_{i,s}}{\sigma_{i,s}^2} \boldsymbol{x}_i)\right\}}{|\boldsymbol{S}_{k,s}|^{-1/2} \exp\left\{\frac{1}{2} \bar{\boldsymbol{x}}_{k,s}^\top \boldsymbol{S}_{k,s}^{-1} \bar{\boldsymbol{x}}_{k,s}\right\}}, \end{aligned} \tag{A.8}$$

and similarly to assign $\boldsymbol{z}_i$ to an unpopulated cluster we get

$$p(\boldsymbol{z}_i = K+1 \mid \boldsymbol{X}, \boldsymbol{\Sigma}, \{\sigma_n\}, \{w_n\}, \boldsymbol{z}) = \tag{A.9}$$
$$\alpha \frac{|\beta_s \boldsymbol{\Sigma}^{-1} + \frac{w_{i,s}^2}{\sigma_{i,s}^2}\boldsymbol{I}|^{-1/2}}{|\beta_s \boldsymbol{\Sigma}^{-1}|^{-1/2}} \exp\left\{ \frac{1}{2} \frac{w_{i,s}}{\sigma_{i,s}^2} \boldsymbol{x}_i^\top \left(\boldsymbol{\Sigma}^{-1} + \frac{w_{i,s}^2}{\sigma_{i,s}^2} \boldsymbol{I} \right)^{-1} \frac{w_{i,s}}{\sigma_{i,s}^2} \boldsymbol{x}_i \right\}.$$

We employ Metropolis-Hastings sampling for the hyper parameters using Gaussian proposal distributions. The parameters $\sigma_{i,s}$, γ, and β are required to be positive and to achieve to do this while maintaining the symmetric Gaussian proposal distribution we sample in the log-transformed domain.

Appendix A.1. Posterior quantities

The posterior mean time series, $p(\boldsymbol{\mu}_k \mid \boldsymbol{X}, \{\boldsymbol{\sigma}_s^2\}, \{\boldsymbol{w}_s\}, \boldsymbol{\Sigma}, \gamma)$, can similarly be evaluated analytically for a fixed clustering assignment. This is given by

$$p(\boldsymbol{\mu}_{k,s} \,| \boldsymbol{X}, \{\boldsymbol{\sigma}_s^2\}, \{\boldsymbol{w}_s\}, \boldsymbol{\Sigma}, \gamma) \sim \tag{A.10}$$
$$\mathcal{N}((\boldsymbol{\Sigma}_{\mathrm{SE}}^{-1} + \sum_{i \in \mathcal{Z}_k} \frac{w_{i,s}^2}{\sigma_{i,s}^2} \boldsymbol{I})^{-1} \bar{\boldsymbol{x}}_k, (\boldsymbol{\Sigma}_{\mathrm{SE}}^{-1} + \sum_{i \in \mathcal{Z}_k} \frac{w_{i,s}^2}{\sigma_{i,s}^2} \mathbf{I})^{-1}))$$

Appendix B. Additional results

The adjusted mutual information between the true clustering and the clustering inferred by the compared models are presented in Fig. B.12 for the synthetic dataset.

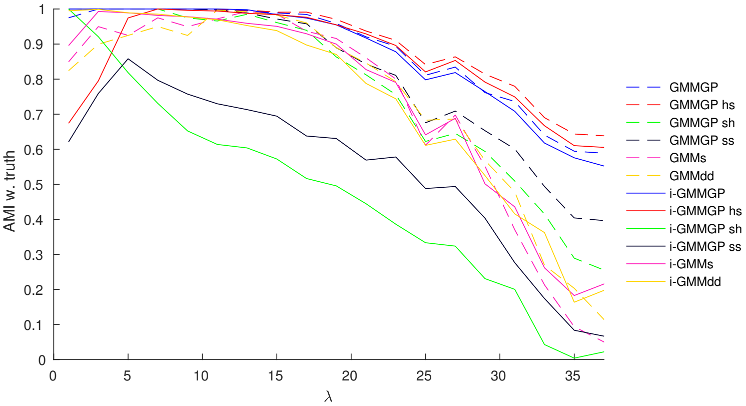


Figure B.12: Synthetic analysis on the *simple* dataset in the top panel and the *hard* dataset in the bottom panel.

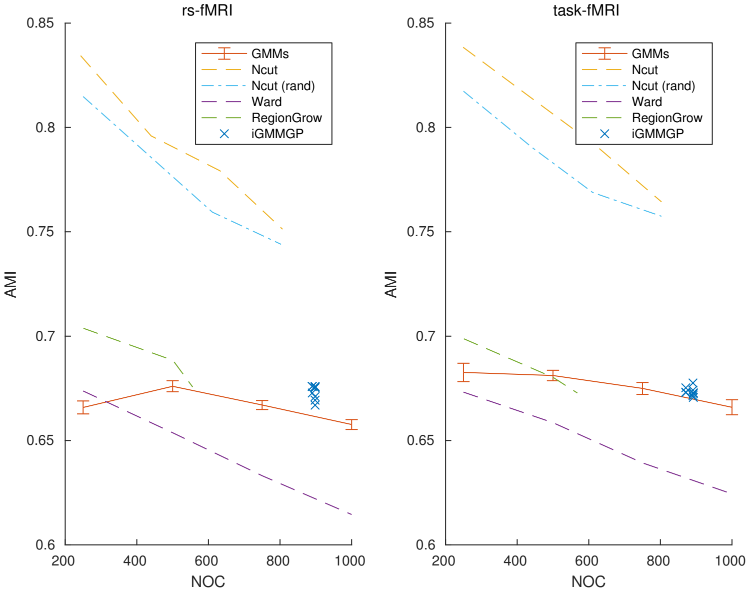


Figure B.13: A comparison of the spatially constrained clustering methods with the GMMs and iGMMGP models. Ncut (rand) is the Ncut algorithm with the similarity matrix randomly permuted to test whether the high AMI is caused by information on the voxel location or by the functional connectivity.

Bibliography

T. Ai, J. N. Morelli, X. Hu, D. Hao, F. L. Goerner, B. Ager, and V. M. Runge. A historical overview of magnetic resonance imaging, focusing on technological innovations. *Investigative radiology*, 47(12):725–741, 2012.

K. J. Albers, A. L. A. Moth, M. N. Schmidt, and M. Mørup. Large scale inference in the infinite relational model: Gibbs sampling is not enough. In *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2013.

K. J. Albers, M. N. Schmidt, et al. The influence of hyper-parameters in the infinite relational model. In *Machine Learning for Signal Processing (MLSP), 2016 IEEE 26th International Workshop on*, pages 1–6. IEEE, 2016.

D. J. Aldous. *Exchangeability and related topics*. Springer, 1985.

K. W. Andersen, K. H. Madsen, H. R. Siebner, M. N. Schmidt, M. Mørup, and L. K. Hansen. Non-parametric bayesian graph models reveal community structure in resting state fmri. *NeuroImage*, 100:301–315, 2014.

D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

C. Baldassano, D. M. Beck, and L. Fei-Fei. Parcellating connectivity in spatial maps. *PeerJ*, 3:e784, 2015.

P. A. Bandettini, E. C. Wong, R. S. Hinks, R. S. Tikofsky, and J. S. Hyde. Time course epi of human brain function during task activation. *Magnetic resonance in medicine*, 25(2):390–397, 1992.

A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(Sep):1345–1382, 2005.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

D. M. Blei and P. I. Frazier. Distance dependent chinese restaurant processes. *Journal of Machine Learning Research*, 12(Aug):2461–2488, 2011.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *arXiv preprint arXiv:1601.00670*, 2016.

T. Blumensath, S. Jbabdi, M. F. Glasser, D. C. Van Essen, K. Ugurbil, T. E. J. Behrens, and S. M. Smith. Spatially constrained hierarchical parcellation of the brain with resting-state fmri. *Neuroimage*, 76:313–324, 2013.

C. Blundell, Y. W. Teh, and K. A. Heller. Bayesian rose trees. *arXiv preprint arXiv:1203.3468*, 2012.

G. Celeux, M. Hurn, and C. P. Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970, 2000.

J. R. Chumbley and K. J. Friston. False discovery rate revisited: Fdr and topological inference using gaussian random fields. *Neuroimage*, 44(1):62–70, 2009.

N. W. Churchill, K. Madsen, and M. Mørup. The functional segregation and integration model: Mixture model representations of consistent and variable group-level connectivity in fmri. *Neural Computation*, 2016.

A. L. Cohen, D. A. Fair, N. U. Dosenbach, F. M. Miezin, D. Dierker, D. C. Van Essen, B. L. Schlaggar, and S. E. Petersen. Defining functional areas in individual human brains using resting functional connectivity mri. *Neuroimage*, 41(1):45–57, 2008.

R. C. Craddock, G. A. James, P. E. Holtzheimer, X. P. Hu, and H. S. Mayberg. A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human brain mapping*, 33(8):1914–28, aug 2012. ISSN 1097-0193. doi: 10.1002/hbm.21333.

D. B. Dahl. Sequentially-allocated merge-split sampler for conjugate and nonconjugate dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 11, 2005.

S. B. Eickhoff, B. Thirion, G. Varoquaux, and D. Bzdok. Connectivity-based parcellation: Critique and implications. *Human brain mapping*, 36(12):4771–4792, 2015.

S. H. Faro and F. B. Mohamed. *Functional MRI: basic principles and clinical applications*. Springer Science & Business Media, 2006.

T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

K. J. Friston, S. Williams, R. Howard, R. S. Frackowiak, and R. Turner. Movement-related effects in fmri time-series. *Magnetic resonance in medicine*, 35(3):346–355, 1996.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

C. Gilles and G. Gérard. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781 – 793, 1995. ISSN 0031-3203. doi: http://dx.doi.org/10.1016/0031-3203(94)00125-6. URL `http://www.sciencedirect.com/science/article/pii/0031320394001256`.

M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature*, 2016a.

M. F. Glasser, S. M. Smith, D. S. Marcus, J. L. Andersson, E. J. Auerbach, T. E. Behrens, T. S. Coalson, M. P. Harms, M. Jenkinson, S. Moeller, et al. The human connectome project's neuroimaging approach. *Nature Neuroscience*, 19(9):1175–1187, 2016b.

S. Gopal and Y. Yang. Von mises-fisher clustering models. In *ICML*, pages 154–162, 2014.

E. M. Gordon, T. O. Laumann, J. F. Adeyemo, B. and. Huckins, W. M. Kelley, and S. E. Petersen. Generation and evaluation of a cortical area parcellation from resting-state correlations. *Cerebral cortex*, page bhu239, 2014.

K. J. Gorgolewski, A. Storkey, M. E. Bastin, I. R. Whittle, J. M. Wardlaw, and C. R. Pernet. Single subject fmri test–retest reliability metrics and confounding factors. *Neuroimage*, 69:231–243, 2013.

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

R. B. Grosse and D. K. Duvenaud. Testing mcmc code. *arXiv preprint arXiv:1412.5218*, 2014.

M. Hanke, F. J. Baumgartner, P. Ibe, F. R. Kaule, S. Pollmann, O. Speck, W. Zinke, and J. Stadler. A high-resolution 7-tesla fmri dataset from complex natural stimulation with an audio movie. *Scientific data*, 1, 2014.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

K. A. Heller and Z. Ghahramani. Bayesian hierarchical clustering. In *Proceedings of the 22nd international conference on Machine learning*, pages 297–304. ACM, 2005.

S. Jain and R. M. Neal. A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1), 2004.

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

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

K. K. Kwong, J. W. Belliveau, D. A. Chesler, I. E. Goldberg, R. M. Weisskoff, B. P. Poncelet, D. N. Kennedy, B. E. Hoppel, M. S. Cohen, and R. Turner. Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences*, 89(12): 5675–5679, 1992.

D. Lashkari and P. Golland. Exploratory fmri analysis without spatial normalization. In *International Conference on Information Processing in Medical Imaging*, pages 398–410. Springer, 2009.

D. Lashkari, E. Vul, N. Kanwisher, and P. Golland. Discovering structure in the space of fmri selectivity profiles. *Neuroimage*, 50(3):1085–1098, 2010.

M. Lázaro-Gredilla, S. Van Vaerenbergh, and N. D. Lawrence. Overlapping mixtures of gaussian processes for the data association problem. *Pattern Recognition*, 45(4):1386–1395, 2012.

T. E. Lund, K. H. Madsen, K. Sidaros, W.-L. Luo, and T. E. Nichols. Non-white noise in fmri: does modelling have an impact? *Neuroimage*, 29(1): 54–66, 2006.

K. Mardia and P. E. Jupp. *Directional statistics*, volume 494. John Wiley & Sons, 2009.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

J. W. Miller and M. T. Harrison. A simple example of dirichlet process mixture inconsistency for the number of components. In *Advances in neural information processing systems*, pages 199–206, 2013.

M. Mørup, K. Madsen, A.-M. Dogonowski, H. Siebner, and L. K. Hansen. Infinite relational modeling of functional connectivity in resting state fmri. In *Advances in neural information processing systems*, pages 1750–1758, 2010.

R. M. Neal. Probabilistic inference using markov chain monte carlo methods. 1993.

R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

T. E. Nichols and A. P. Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1):1–25, 2002.

G. Nunez-Antonio and E. Gutiérrez-Pena. A bayesian analysis of directional data using the von mises–fisher distribution. *Communications in Statistics—Simulation and Computation®*, 34(4):989–999, 2005.

S. Ogawa, T.-M. Lee, A. S. Nayak, and P. Glynn. Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magnetic resonance in medicine*, 14(1):68–78, 1990.

S. Ogawa, D. W. Tank, R. Menon, J. M. Ellermann, S. G. Kim, H. Merkle, and K. Ugurbil. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, 89(13):5951–5955, 1992.

A. X. Patel, P. Kundu, M. Rubinov, P. S. Jones, P. E. Vértes, K. D. Ersche, J. Suckling, and E. T. Bullmore. A wavelet method for modeling and despiking motion artifacts from resting-state fmri time series. *Neuroimage*, 95:287–304, 2014.

C. E. Rasmussen. The infinite gaussian mixture model. In *NIPS*, volume 12, pages 554–560, 1999.

C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning.* the MIT Press, 2006.

P. M. Rasmussen, L. K. Hansen, K. H. Madsen, N. W. Churchill, and S. C. Strother. Model sparsity and brain pattern interpretation of classification models in neuroimaging. *Pattern Recognition*, 45(6):2085–2100, 2012.

C. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation.* Springer Science & Business Media, 2007.

J. C. Ross and J. G. Dy. Nonparametric mixture of gaussian processes with constraints. In *ICML (3)*, pages 1346–1354, 2013.

S. Ryali, T. Chen, K. Supekar, and V. Menon. 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, 2013. doi: 10.1016/j.neuroimage.2012.09.067.

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

S. C. Strother, J. Anderson, L. K. Hansen, U. Kjems, R. Kustra, J. Sidtis, S. Frutiger, S. Muley, S. LaConte, and D. Rottenberg. The quantitative evaluation of functional neuroimaging experiments: the npairs data analysis framework. *NeuroImage*, 15(4):747–771, 2002.

J. Taghia, Z. M., and A. Leijon. Bayesian estimation of the von-mises fisher mixture model with variational inference. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(9):1701–1715, Sept 2014. ISSN 0162-8828. doi: 10.1109/TPAMI.2014.2306426.

B. Thirion, G. Varoquaux, E. Dohmatob, and J. Poline. Which fmri clustering gives good brain parcellations? *Frontiers in neuroscience*, 8:167, 2014.

G. Tononi, O. Sporns, and G. M. Edelman. A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proceedings of the National Academy of Sciences*, 91(11):5033–5037, 1994.

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

N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1073–1080. ACM, 2009.

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

J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

G. S. Wig, T. O. Laumann, and S. E. Petersen. An approach for parcellating human cortical areas using resting-state correlations. *Neuroimage*, 93:276–291, 2014.

K. J. Worsley, S. Marrett, P. Neelin, A. C. Vandal, K. J. Friston, A. C. Evans, et al. A unified statistical approach for determining significant signals in images of cerebral activation. *Human brain mapping*, 4(1):58–73, 1996.

B. T. Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zöllei, J. R. Polimeni, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 106(3):1125–1165, 2011.