Statistical Methods for Neuroimaging Data Analysis and Cognitive Science

by

Yin Song
B.Sc., Northwest A&F University, 2010
M.Sc., University of Alaska Fairbanks, 2013

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Mathematics and Statistics

Statistical Methods for Neuroimaging Data Analysis and Cognitive Science

by

Yin Song
B.Sc., Northwest A&F University, 2010
M.Sc., University of Alaska Fairbanks, 2013

Supervisory Committee

Dr. Farouk Nathoo., Supervisor
(Department of Mathematics and Statistics)

Dr. Laura Cowen., Departmental Member
(Department of Mathematics and Statistics)

Dr. Michael E. J. Masson., Outside Member
(Department of Psychology)

Supervisory Committee

Dr. Farouk Nathoo., Supervisor
(Department of Mathematics and Statistics)

Dr. Laura Cowen., Departmental Member
(Department of Mathematics and Statistics)

Dr. Michael E. J. Masson., Outside Member
(Department of Psychology)

ABSTRACT

This thesis presents research focused on developing statistical methods with emphasis on tools that can be used for the analysis of data in neuroimaging studies and cognitive science. The first contribution addresses the problem of determining the location and dynamics of brain activity when electromagnetic signals are collected using magnetoencephalography (MEG) and electroencephalography (EEG). We formulate a new spatiotemporal model that jointly models MEG and EEG data as a function of unobserved neuronal activation. To fit this model we derive an efficient procedure for simultaneous point estimation and model selection based on the iterated conditional modes algorithm combined with local polynomial smoothing. The methodology is evaluated through extensive simulation studies and an application examining the visual response to scrambled faces.

In the second contribution we develop a Bayesian spatial model for imaging genetics developed for analyses examining the influence of genetics on brain structure as measured by MRI. We extend the recently developed regression model of Greenlaw et al. (*Bioinformatics*, 2017) to accommodate more realistic correlation structures typically seen in structural brain imaging data. We allow for spatial correlation in the imaging phenotypes obtained from neighbouring regions in the same hemisphere of

the brain and we also allow for correlation in the same phenotypes obtained from different hemispheres (left/right) of the brain. This correlation structure is incorporated through the use of a bivariate conditional autoregressive spatial model. Both Markov chain Monte Carlo (MCMC) and variational Bayes approaches are developed to approximate the posterior distribution and Bayesian false discovery rate (FDR) procedures are developed to select SNPs using the posterior distribution while accounting for multiplicity. The methodology is evaluated through an analysis of MRI and genetic data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and we show that the new spatial model exhibits improved performance on real data when compared to the non-spatial model of Greenlaw et al. (2017).

In the third and final contribution we develop and investigate tools for the analysis of binary data arising from repeated measures designs. We propose a Bayesian approach for the mixed-effects analysis of accuracy studies using mixed binomial regression models and we investigate techniques for model selection.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my supervisor Dr. Farouk Nathoo for top-notch research guidance, illuminating chats, buying coffee, providing funding, sharing cookies and frequent advice over the years at University of Victoria.

Many thanks to my committee members Dr. Michael E. J. Masson and Dr. Marc Fredette for helpful comments and insights. I'm beholden to Amy Almeida, Kelly Choo, Carol Anne Sargent, all fellow graduate students and my friends for providing all the help and support.

Finally, I owe a tremendous amount to my beloved parents for endless love, support and encouragement.

DEDICATION

To my grandma (1933.10.20 - 2019.02.14),
my parents,
my friends,
this ugly and yet beautiful world.
Thank you for everything.

Chapter 1

Introduction

1.1 Background

The general focus of this thesis lies with the development of statistical methods for applications in neuroimaging and cognitive neuroscience. Within this broad setting we focus specifically on three problems, each of which is fairly distinct, and each providing significant challenges for the development and evaluation of statistical methodology and associated software. Throughout the thesis we develop Bayesian approaches (Gelman et al. 2014) to statistical inference problems. The Bayesian paradigm summarizes all sources of uncertainty through probability distributions on models, parameters, latent variables and data. Inference is centered on the posterior distribution, which is the distribution of the unknown quantities conditional on the observed data. A key ingredient is the prior distribution, a probability distribution or a σ-finite non-negative measure over the parameter space summarizing the information and uncertainty associated with the parameter before observing the data. The movement from prior distribution to posterior distribution in light of observing data and under an assumed model for the observations is governed by Bayes rule (Gelman et al. 2014). For high-dimensional settings where the number of parameters can be large relative to the number of observations, the Bayesian paradigm incorporates regularization naturally through the prior distribution. This notion is found useful in Chapters 2 and 3 both of which involve high-dimensional regression problems where priors are used to either encourage spatial smoothness (in the case of reconstructing brain images) and encourage group sparsity (in the case of relating a genetic marker to a set of brain structure measurements). In Chapter 4 we consider model selection

and hypothesis testing for repeated measures designs with focus on studies of cognitive processes typically considered in studies of memory and language. The Bayesian approach in this context is motivated by a general trend towards the use of Bayesian methods for the analysis of data in cognitive science (see e.g., Wagenmakers, 2007; Masson, 2011; Rouder et al., 2012, Nathoo and Masson, 2016, Nathoo, Kilshaw and Masson, 2018).

In Chapter 2 we consider the analysis of data collected using Magnetoencephalography (MEG) and electroencephalography (EEG). MEG can be used to obtain measurements of the magnetic fields around the scalp that are associated with electrical neural activity within the brain. Similarly, EEG can be used to obtain measures of the time-varying electric field at different locations of the scalp. These neuroimaging modalities have been used extensively because of their excellent temporal resolution and non-invasive nature. Unfortunately, the spatial resolution associated with MEG and EEG is restricted with observations taken around the scalp. In addition, an observation made at a particular scalp location may comprise sources of neural activity from many locations of the brain. Localizing the neural activity generators of MEG and/or EEG signals is thus a problem of interest in moving from scalp data to studying neural activity at the level of the brain. This inverse problem, known as the neuro-electromagnetic inverse problem, involves inverting Maxwell's equations (Sarvas, 1987) in order to estimate the underlying sources of electrical neural activity within the brain generating scalp level electromagnetic fields. The associated forward problem involves predicting observations from a given configuration of neural activity and involves solving Maxwell's equations. Figure 1.1 presents an illustration of forward and inverse problems with MEG/EEG data.

The neuro-electromagnetic inverse problem is commonly considered to be one of the most challenging problems in neuroimaging data analysis. It is an ill-posed inverse problem in the sense that the solution is not unique. A considerable amount of research has been produced on developing practical methods for the MEG/EEG inverse problem. This includes methods involving regularization such as the L_1 penalty (Matsuura and Okabe, 1995) and L_2 penalty (Pascual-Marqui, Michel, and Lehmann, 1994). Bayesian approaches have also been proposed where the constraints enter as priors and the objective of model inversion is to estimate the conditional or posterior probability of the model parameters (see, e.g., Wipf and Nagarajan, 2009; Friston et al., 2008; Henson et al., 2009b; Henson et al., 2010; Long et al., 2011; Nathoo et al., 2014; Calvetti et al., 2015; Vivaldi and Sorrentino, 2016; Lim et al., 2017; Sorrentino

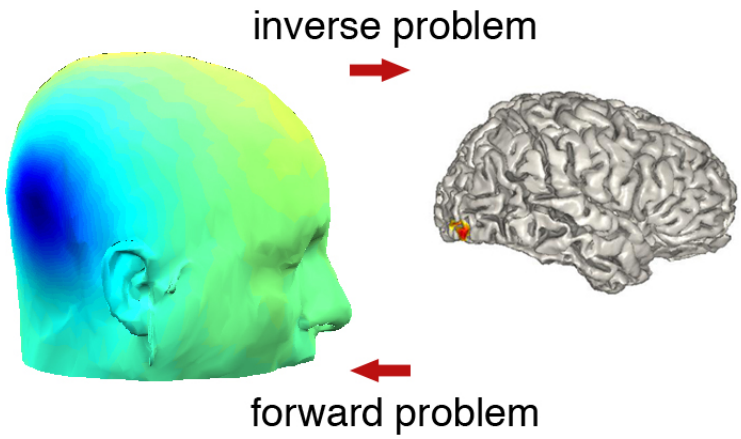


Figure 1.1: Illustration of the forward and inverse problems. The forward problem can be described as the problem of predicting the electric potential or magnetic field vector that would be externally measured at the sensors given the neural source activity inside the brain (the red region shown on the right). The inverse problem can be described as the problem of estimating the current density or activity values of the source that underlies the measured electric potential or magnetic field on the scalp (blue region shown on the left). This figure is taken from Ramírez (2008).

and Piana, 2017; Zakharova et al., 2017). In formulating a solution to the inverse problem, it can be useful to think of the neural dynamics, represented through a high-dimensional parameter, as being generated in a low-dimensional latent space. This notion is useful when dealing with high-dimensional data problems and their associated parameter spaces. In the case of MEG and EEG data this leads to mixture models and along these lines Daunizeau and Friston (2007) proposed a mixture model, termed the meso-statespace model (MSM), which formulates neural activity dynamics in a meso-statespace with different states representing different, possibly interconnected, sources of neural activity. Motivated by this idea, we propose a new model that combines the data from three different modalities, MRI, MEG, and EEG, together while at the same time building in spatial correlation through a discrete spatial process. Both aspects, using the combination of MEG and EEG together and introducing spatial dependence in the allocation of brain locations to meso-states are new ideas proposed in Chapter 2. While useful, these new ideas present significant challenges for Bayesian computation and model selection and we address these challenges by developing a novel algorithm for simultaneous point estimation and model selection for latent Gaussian mixture models. Our algorithm is developed for the specific model proposed in Chapter 2 but it can be applied far more generally to latent

Gaussian mixture models.

In Chapter 3 we consider the problem of relating individual neuroimaging data to genetic markers. This is a regression problem that is often referred to as a big-data-squared problem (see, e.g., Nathoo, Kong and Zhu, 2018) as both the response (an image of the brain) and the covariates (a set of genetic markers) can be high-dimensional while the sample size, that is, the number of subjects over which genetic and neuroimaging data are collected can be relatively low or on the same order.

A number of recent approaches including reduced rank regression, massive univariate approaches, group-sparse multi-task regression and Bayesian methods with lasso-type priors have been proposed for regression analysis in this setting (see e.g., Vounou et al., 2010; Stein et al., 2010; Silver et al., 2011; Inkster et al., 2010; Hibar et al., 2011; Ge et al., 2012; Thompson et al., 2013; Stingo et al., 2013; Zhu et al., 2014; Hibar et al., 2015; Huang et al., 2015; Huang et al., 2017; Lu et al., 2017). Nathoo, Kong and Zhu (2018) provide a comprehensive review of recent statistical approaches for the joint analysis of high-dimensional imaging and genetic data and discuss open problems within this area, one of which is the development of models that allow for spatial correlation in neuroimaging phenotypes. More recently, Greenlaw et al. (2017) developed a hierarchical Bayesian model with regularizing shrinkage priors developed so that the associated posterior mode corresponds to the group-sparse multi-task regression estimator proposed by Wang et al. (2012). The Bayesian approach is developed in order to obtain uncertainty quantification on the regression parameters through Bayesian credible intervals and the posterior distribution. Motivated by the model proposed by Greenlaw et al. (2017), we develop an extension of the model that explicitly accounts for spatial correlation and bilateral correlation structure typically seen in the MRI data. After accounting for spatial dependence in the model, Bayesian computation is a significant challenge. We investigate the use of both sampling based Monte Carlo methods as well as variational Bayes deterministic approximations, and we combine these approaches in order to obtain a practical solution that has the advantages of both. Genuine improvements on real data are also demonstrated when using an explicitly spatial model.

In Chapter 4 we consider model selection and hypothesis testing for repeated measures designs often conducted in the memory and language domain where repeated measures of a binary response are obtained from each subject. We introduce Bayesian approaches and associated software for these designs as part of a move within this area towards the use of practical Bayesian methods as an alternative to

null-hypothesis significance testing (NHST). Wagenmakers (2007) discusses problems with NHST and proposes the use of Bayesian inference with implementation based on the Bayesian information criterion (BIC) as an approximation to the Bayes factor (Kass and Raftery, 1995). To date, most approaches in the memory and language literature have emphasized models with Gaussian assumptions or logistic and probit mixed models based on classical methods. The former, which is still very much in common use, requires averaging the response over individual trials within subjects, where at a given trial, a binary response is observed. This averaging approach has a number of drawbacks and we develop an alternative solution based on Bayesian generalized linear mixed models for binary response variables. The effect of experimental conditions on accuracy is assessed through Bayesian model selection and we consider two such approaches to model selection: (a) the Bayes factor through the Bayesian Information Criterion (Wagenmakers, 2007) approximation and (b) the Watanabe-Akaike Information Criterion (Watanabe, 2010). Simulation studies are used to assess the operating characteristics of these approaches and to demonstrate advantages over the more standard approach that consists of aggregating the accuracy data across trials within each condition and over the contemporary use of logistic and probit mixed models with model selection based on the Akaike Information Criterion (Akaike, 1973). While well known and well studied within the statistical literature, these approaches, particularly when considered within the context of Bayesian model selection have not been considered extensively within the memory and language literature. Bayesian methods can be computationally and mathematically intensive, and this may limit the level of practical application. Hence, we provide a suite of models and associated software for their implementation.

1.2 Acronyms and Definitions

Some acronyms and definitions used in Chapters 2 , 3 and 4 are defined below:

- Chapter 2:
  - MEG: Magnetoencephalography - a functional neuroimaging technique for capturing brain activity through very sensitive magnetometers that can record magnetic fields produced by electrical currents occurring naturally in the brain.

  - EEG: Electroencephalography - an electrophysiological monitoring method to record electrical activity of the brain with the electrodes placed along the scalp.
  - MRI: Magnetic resonance imaging - a form of medical imaging that uses the body's natural magnetic properties when placed in a strong magnetic field to produce images of the internal organs.
  - fMRI: Functional magnetic resonance imaging - measures brain activity by detecting changes associated with blood flow.
  - Voxel - a voxel represents a value on a regular grid in three-dimensional space.
  - ICM: Iterated Conditional Modes - a deterministic algorithm for obtaining a configuration of a local maximum of the joint probability of a multivariate probability distribution. It does this by iteratively maximizing the probability of each variable conditioned on the rest. It can be viewed as a deterministic version of Gibbs sampling.

- Chapter 3:

  - ADNI: The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a longitudinal multicenter study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of Alzheimer's disease (AD). Since its launch more than a decade ago, the landmark public-private partnership has made major contributions to AD research, enabling the sharing of data between researchers around the world.
  - APOE gene: The APOE gene provides instructions for making a protein called apolipoprotein E.
  - SNP: single-nucleotide polymorphism - is a substitution of a single nucleotide that occurs at a specific position in the genome.
  - FDR: False Discovery Rate - The false discovery rate (FDR) is a method of conceptualizing the rate of type I errors in null hypothesis testing when conducting multiple comparisons. FDR-controlling procedures are designed to control the expected proportion of "discoveries" (rejected null hypotheses) that are false (incorrect rejections).

- Chapter 4:

- AIC: Akaike information criterion.
- BIC: Bayesian information criterion.
- WAIC: Watanabe - Akaike information criterion.
- Repeated Measures Design: An experimental design involving a sample of subjects where each subject is measured for a response variable more than once over a sequence of trials. The trials can consist of different experimental manipulations and stimuli.
- Trial: within the context of a repeated measures design, a trial is an individual instance where a measurement of the response variables is made on a subject.
- Item: a specific case comprising the stimulus or part of the stimulus in a given trial of a repeated measures design (e.g. a specific word shown to a subject).

1.3 Contributions

To reiterate, this thesis makes three contributions to the development of statistical methodology for the analysis of neuroimaging data and repeated measures designs in cognitive science. Chapters 2, 3 and 4 of this thesis each correspond to one paper. The final chapter concludes with a discussion of future work. As of this writing, one paper has been published, one has been accepted for publication and a third paper has been invited for revision. This is summarized as follows:

1. Song, Y., Nathoo, F.S., Babul A. (2018). A Potts-Mixture Spatiotemporal Joint Model for Combined MEG and EEG Data. *Canadian Journal of Statistics*, accepted for publication.

2. Song, Y., Ge, S., Cao, J., Wang, L., Nathoo, F.S. A Bayesian Spatial Model for Imaging Genetics (2019). Under revision for *Biometrics*. Revision invited May 16, 2019.

3. Song, Y., Nathoo, F.S., Masson, M.E.J. (2017). A Bayesian approach to the mixed effects analysis of repeated measures accuracy studies. *Journal of Memory and Language*, DOI: 10.1016/j.jml.2017.05.002.

Each contribution in the thesis has an associated software package. These software packages are available for download as follows:

1. R software and examples illustrating our methodology for the mixed effects analysis of repeated measures accuracy studies is available here:
   `https://v2south.github.io/BinBayes/`.

2. R software for fitting the latent Gaussian mixture model for combined analysis (solving the inverse problem) of MEG and EEG data is available here:
   `https://github.com/v2south/PottsMix`.

3. An R package for fitting the bivariate conditional autoregressive regression model with group lasso priors is available for download on The Comprehensive R Archive Network (CRAN) from within R. More information on the software package is available here:
   `https://cran.r-project.org/web/packages/bgsmtr/`
   and a manual for the software package is available here:
   `https://cran.r-project.org/web/packages/bgsmtr/bgsmtr.pdf`.

Chapter 2

A Potts-Mixture Spatiotemporal Joint Model for Combined MEG and EEG Data

2.1 Introduction

Magnetoencephalography (MEG) and electroencephalography (EEG) are neuroimaging modalities that have been widely used to study the function of the brain non-invasively using an array of sensors placed on (EEG) or above the scalp (MEG). These sensor arrays can be used to capture the time-varying electromagnetic field that exists around the head as a result of electrical neural activity within the brain. At a given position of the array, each sensor records a scalar-valued time series representing either the electric potential (EEG) or the magnetic field (MEG) at that position. While the magnetic field is a vector field in space, most MEG machines measure only one component of this field, so both EEG and MEG usually measure scalar fields around the scalp. The entire array thus produces a multivariate time series such as the data depicted in Figure 2.1, panel (a), an example of an MEG, and Figure 2.1, panel (c), an example of an EEG. Viewed from a spatial perspective, at a given time point, each array records a spatial field such as that depicted in Figure 2.1, panel (b), which shows the MEG spatial field at a particular time point, and Figure 2.1, panel (d), which shows the EEG spatial field at the same time point.

While both EEG and MEG are generated by neural activity, each modality measures this activity indirectly in a different way, through the associated electric field

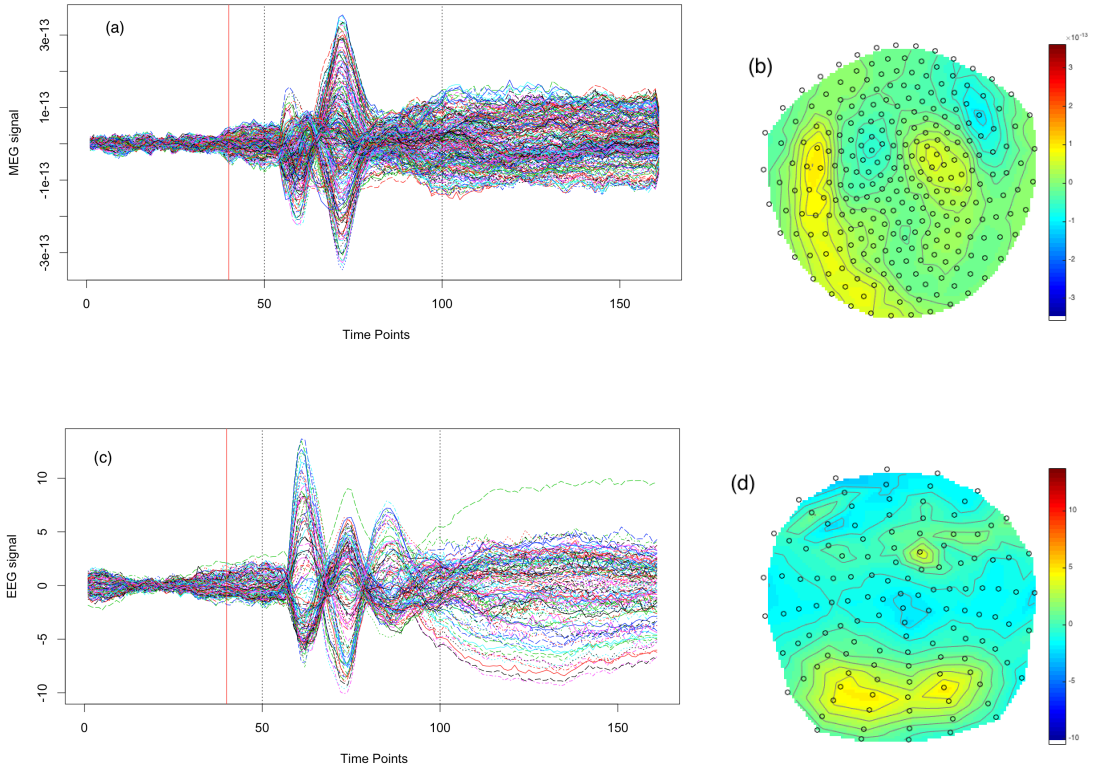


Figure 2.1: The MEG and EEG data considered for one individual subject in the face perception study: panels (a) and (c) show the time series observed at each MEG sensor and EEG sensor, respectively; panels (b) and (d) depict the spatially interpolated values of the MEG data and the EEG data, respectively, each observed at time points$t$ where $t = 80$, roughly 200ms after presentation of the stimulus. In panels (b) and (d) the black circles correspond to the sensor locations after projecting these locations onto a 2-dimensional grid (for presentation). The MEG and EEG data represent averages over 336 and 344 independent and identically distributed trials respectively for one subject.

and magnetic field respectively. In typical studies, EEG and MEG are used separately to study brain activity that is evoked by a particular stimulus, or to study the brain while it is at rest. Our emphasis in this article is the development of methodology for the analysis of such data in the former case when brain activity is evoked with a particular stimulus. The focus is on situations where the MEG and EEG data are collected simultaneously or the data are collected sequentially in a situation where the data mimic a simultaneous recording paradigm. MEG and EEG pick up currents from mostly disjoint, though contiguous sections of the cortex. Therefore, technically the sources of the EEG signal will contribute only modestly to the MEG signal, and vice versa, although if the activity is continuous over large regions of the cortex, the sources will be spatially correlated. Thus a joint spatial model for combined MEG and EEG data should lead to improved source estimation.

While EEG and MEG data can and often are analyzed directly at the sensor level (see, e.g. Ismail et al., 2013), our objective in the analysis of these data is to localize the sources of neural activity within the brain. That is, we want to take the data collected over the sensor arrays and map these data back to the brain. In doing so we want to combine the MEG and EEG data together as both datasets are generated from the neural response of interest. Our proposed methodology is applicable to settings where it is believed that the neural activity is generated by a small number (e.g., 2 to 9) of brain activity sources, known as latent states with each state having its own dynamics. Thus two primary challenges we are faced with when considering this problem are: (i) a combined analysis of MEG and EEG data, and (ii) estimation of low-dimensional latent structure.

In considering our objective of mapping the sensor array data back to the brain we must consider the relationship between the observed data and the unknown neural activity within the brain. This relationship is governed by the theory of electrodynamics, which is described by Maxwell's equations. This theory will thus play a role in our solution to the inverse problem and will be incorporated into an associated statistical model.

Generally, an inverse problem is said to be well-posed if a solution exists, the solution is unique, and if the solution's behavior changes continuously with the initial conditions (Hadamard, 1902). A problem that is not well-posed is said to be ill-posed. Within the setting of electromagnetic data it has been shown theoretically (von Helmholtz, 1853) that the problem of finding the sources of electromagnetic fields outside a volume conductor has an infinite number of solutions.

Outside of the statistical literature this inverse problem has received a great deal of attention in the field of neuroimaging. Many existing solutions are based on regularization, often within the context of penalized likelihood estimation. Methods based on either an L_2 penalty (Pascual-Marqui, Michel, and Lehmann, 1994) or an L_1 penalty (Matsuura and Okabe, 1995) have been proposed, as have more general approaches, such as the solution proposed by Tian and Li (2011), where a group elastic net (Zou and Hastie, 2005) for MEG source reconstruction is developed. Bayesian approaches have also been developed by a number of authors including Friston et al. (2008) and Wipf and Nagarajan (2009). They considered Gaussian scale mixtures incorporating a potentially large number of covariance components representing spatial patterns of neural activity. Henson et al. (2009b) extended this approach for combined MEG and EEG data, while Henson et al. (2010) further developed an approach for combined EEG/MEG and fMRI data. Long et al. (2011) account for the dynamic nature of the problem by using the Kalman filter with implementation on a network of high performance computers, while Calvetti et al. (2015), Vivaldi and Sorrentino (2016), and Sorrentino and Piana (2017) considered Bayesian smoothing approaches.

Zakharova et al. (2017) discussed the inverse problem and developed a solution within the context of noninvasive preoperative methods for the localization of sources which can guide the outcome of brain surgery. Aydin et al. (2017) developed an approach for source localization based on combined MEG and EEG data where the source localization is guided by a detailed and calibrated finite element head model that considers the variation of individual skull conductivities and white matter anisotropy. Lim et al. (2017) developed a method for sparse EEG/MEG source estimation based on the group lasso that has been adapted to take advantage of functionally-defined regions of interest for the definition of physiologically meaningful groups within a functionally-based common space. Belaoucha and Papadopoulo (2017) used spatial information based on diffusion MRI to solve the MEG/EEG inverse problem, while Nathoo et al. (2014) used spatial spike-and-slab priors to solve the EEG/MEG inverse problem while incorporating fMRI data.

In all of the approaches mentioned above, the unknown neural activity is restricted to the cortical surface, and the solution to the inverse problem is informed by the anatomy of the brain using structural MRI. This has the advantage of using an anatomical constraint that is realistic since it is widely believed that the primary generators of the MEG/EEG signal are restricted to the cortex though it does have the disadvantage of excluding sources deeper in the brain.

While all of the above mentioned techniques can be applied generally to evoked response data, our methodology will focus specifically on settings where it is believed a priori that the neural response is generated by a small set of hidden states so that the continuous-current distribution based approaches will not accurately reflect this prior information on the structure of neural activity. In this case a finite mixture model seems more appropriate. Such a model, known as the mesostate-space model (MSM) was developed by Daunizeau and Friston (2007), based on the following four assumptions, taken directly from Daunizeau and Friston (2007)

1. *bioelectric activity is generated by a set of distributed sources*
2. *the dynamics of these sources can be modelled as random fluctuations about a small number of mesostates*
3. *mesostates evolve in a temporally structured way and are functionally connected (i.e. influence each other)*
4. *the number of mesostates engaged by a cognitive task is small (e.g. between one and a few).*

While it is well suited for the settings considered here, the MSM was developed for either EEG or MEG data only, and it is not directly applicable for combined EEG and MEG data. Equally important, the MSM assigns each location on the cortical surface to one of a small number of mixture components using a simple multinomial labelling process, where the corresponding mixture allocation variables are assumed independent and identically distributed. More realistically, we expect these discrete allocation variables to be spatially correlated across brain locations and it is therefore important to incorporate this prior information into the structure of the mixture model.

Motivated by the issues discussed above, we develop a Bayesian model that builds on the MSM in two ways. First, we formulate the model for combined MEG and EEG data. Second, we relax the assumption of independent mixture allocation variables and instead model these variables using the Potts model (Potts, 1952). The Potts model contains a hyperparameter that controls the degree of spatial correlation and we assign a hyperprior to this parameter that accounts for the phase transition point of the Potts model, so that unrealistic allocations are avoided.

For our new model formulation we propose an approach for simultaneous point estimation and model selection based on the iterated conditional modes (ICM) algorithm (Besag, 1986) combined with a pseudolikelihood approximation to the normalizing constant of the Potts model, and local polynomial smoothing. By model selection, we mean choosing the number of mixture components for the latent Gaussian process and our ICM algorithm results in a very simple and novel estimator that appears to have reasonable properties, while being computationally efficient.

2.2 Spatiotemporal Mixture Model

We let $\boldsymbol{M}(t) = (M_1(t), M_2(t), ..., M_{n_M}(t))'$ and $\boldsymbol{E}(t) = (E_1(t), E_2(t), ..., E_{n_E}(t))'$ denote the MEG and EEG measurements in the unit of tesla and voltage, respectively, at time t, $t = 1, \ldots, T$; where n_M and n_E denote the number of MEG and EEG sensors outside the head. We assume that the sensor locations for the two modalities have been co-registered to the same space through an appropriately defined transformation of coordinates (see, e.g., Penny et al., 2011). Correspondingly, we let $\boldsymbol{X}_M$ and $\boldsymbol{X}_E$ denote $n_M \times P$ and $n_E \times P$ design matrices, respectively, which we also refer to as the forward operators in our model. In this case, P represents a large number of point sources of potential neural activity within the brain corresponding to known locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_P$ covering the cortical surface, and it is typical that the value of P is assumed large enough so that $P >> n_E$ and $P >> n_M$. As cortical neurons with their large dendritic trunks locally oriented in parallel, and pointing perpendicularly to the cortical surface are believed to be the main generators of MEG and EEG. The orientations of the point sources are assumed normal to the cortical surface.

The computation of the forward operators is based on a solution to Maxwell's equations under the quasi-static approximation (Sarvas, 1987). A detailed theoretical treatment of the forward problem can be found in Penny et al. (2011). Within the current context it is sufficient to note that $\boldsymbol{X}_{M_{ij}}$ represents the noise free MEG measurement that would be observed at the i^{th} MEG sensor given a unit current source at location $\boldsymbol{s}_j$ in the brain, where $i = 1, ..., n_M$ and $j = 1, ..., P$. The elements of $\boldsymbol{X}_E$ have a similar interpretation for EEG. Under the quasi-static approximation to Maxwell's equations, the MEG/EEG measurement at a given time point, denoted as $\boldsymbol{M}/\boldsymbol{E}$, is related to the unknown neural activity at the same time point $\boldsymbol{S} =$

$(S_1, \ldots, S_P)'$ through a linear relationship (Sarvas, 1987)

$$\boldsymbol{M} = \boldsymbol{X}_M \boldsymbol{S}, \ \boldsymbol{E} = \boldsymbol{X}_E \boldsymbol{S}. \tag{2.1}$$

We assume that the data have been transformed as described in Section A.1 of Appendix A to accommodate different scaling and measurement units across the different sensor-types. Then, we specify a model that is applicable to situations where the number of states is reasonably low or the primary activity can be approximated by a low dimensional process. A voxel is a 3D pixel created by MRI scanning software to represent the brain. We further assume that the P cortical locations are embedded in a 3D regular grid composed of N_v voxels. This assumption allows us to better formulate the hyper-prior for the Potts model and also facilitates a more efficient computational implementation as described below. Given this grid of voxels, we define the mapping $v : \{1, ..., P\} \rightarrow \{1, ...N_v\}$ such that $v(j)$ is the index of the voxel containing the j^{th} cortical location. Beginning with equation (2.1) we add the temporal dimension and incorporate Gaussian measurement error leading to the following specification

$$\begin{aligned}
\boldsymbol{M}(t) &= \boldsymbol{X}_M \boldsymbol{S}(t) + \boldsymbol{\epsilon}_M(t), \quad \boldsymbol{\epsilon}_M(t)|\sigma_M^2 \overset{iid}{\sim} MVN(\boldsymbol{0}, \sigma_M^2 \boldsymbol{H}_M), \ t = 1, \ldots, T \\
\boldsymbol{E}(t) &= \boldsymbol{X}_E \boldsymbol{S}(t) + \boldsymbol{\epsilon}_E(t), \quad \boldsymbol{\epsilon}_E(t)|\sigma_E^2 \overset{iid}{\sim} MVN(\boldsymbol{0}, \sigma_E^2 \boldsymbol{H}_E), \ t = 1, \ldots, T,
\end{aligned}$$

where $\boldsymbol{H}_M$ and $\boldsymbol{H}_E$ are known $n_M \times n_M$ and $n_E \times n_E$ matrices which can be obtained from auxiliary data providing information on the covariance structure of EEG and MEG sensor noise, or we can simply set $\boldsymbol{H}_M = \boldsymbol{I}_M$ and $\boldsymbol{H}_E = \boldsymbol{I}_E$ where $\boldsymbol{I}_M$ and $\boldsymbol{I}_M$ are identity matrix. The latter is assumed hereafter.

At the second level of the model we assume that the unknown neural activity arises from a Gaussian mixture model

$$S_j(t)|\boldsymbol{\mu}(t), \boldsymbol{\alpha}, \boldsymbol{Z} \overset{ind}{\sim} \prod_{l=1}^{K} N(\mu_l(t), \alpha_l)^{Z_{v(j)l}}, \tag{2.2}$$

$j = 1, \ldots, P, \ t = 1, \ldots, T$; where $\boldsymbol{Z} = (Z_1', Z_2', ..., Z_{N_v}')'$ is a labelling process defined over the 3D regular grid of voxels such that for each $v(j) \in \{1, ..., N_v\}$, $\boldsymbol{Z}_{v(j)}' = (Z_{v(j)1}, Z_{v(j)2}, ..., Z_{v(j)K})$ with $Z_{v(j)l} \in \{0, 1\}$, where $Z_{v(j)l}$ denotes the labelling process for j^{th} vertices within voxel $v(j)$ to the l^{th} mixture component and $\sum_{l=1}^{K} Z_{v(j)l} = 1$; $\boldsymbol{\mu}(t) = (\mu_1(t), \mu_2(t), \mu_K(t))' = (\mu_1(t), \boldsymbol{\mu}^A(t)')'$, where $\boldsymbol{\mu}^A(t) = (\mu_2(t), ..., \mu_K(t))'$

denotes the mean of the 'active' states and $\mu_1(t) = 0$ for all t, so that the first component corresponds to an 'inactive' state. The variability of the l^{th} mixture component about its mean $\mu_l(t)$ is represented by $\alpha_l, l = 1, \ldots, K$.

The j^{th} location on the cortex is allocated to one of K states through $\boldsymbol{Z}_{v(j)}$, and we assume that the labelling process follows a Potts model (Potts, 1952) so that

$$P(\boldsymbol{Z}|\beta) = \frac{\exp\{\beta \sum_{h\sim j} \delta(\boldsymbol{Z_j}, \boldsymbol{Z_h})\}}{G(\beta)}, \quad \delta(\boldsymbol{Z_j}, \boldsymbol{Z_h}) = 2\boldsymbol{Z_j'Z_h} - 1,$$

where $G(\beta)$ is the normalizing constant for this probability mass function, $\beta \geq 0$ is a hyper-parameter which governs the strength of spatial cohesion, and $h \sim j$ indicates that voxels h and j are neighbours, with a first-order neighbourhood structure over the 3D regular grid of voxels. This $\delta(\boldsymbol{Z_j}, \boldsymbol{Z_h})$ function determines whether neighbouring voxels h and j having the same state or not.

We assume that the mean temporal dynamics follow a first-order vector autoregressive process:

$$\boldsymbol{\mu}^A(t) = \boldsymbol{A}\boldsymbol{\mu}^A(t-1) + \mathbf{a}(t), \;\; \mathbf{a}(t)|\sigma_a^2 \overset{i.i.d}{\sim} MVN(\mathbf{0}, \sigma_a^2\mathbf{I})$$

$t = 2, \ldots, T$, $\boldsymbol{\mu}^A(1) \sim MVN(\mathbf{0}, \sigma_{\mu_1}^2\boldsymbol{I})$, with $\sigma_{\mu_1}^2$ known, but σ_a^2 unknown. The hyper-parameter for the Potts model is assigned a uniform prior $\beta \sim \text{Unif}[0, \beta_{crit}]$, where β_{crit} is an approximation of the phase transition point of the K-state Potts model on a 3-dimensional regular lattice (Moores et al., 2015), $\beta_{crit} = \frac{2}{3}\log\{\frac{1}{2}[\sqrt{2} + \sqrt{4K-2}]\}$. Additional priors completing the model specification are as follows:

$$\begin{aligned}
&\alpha_l \overset{i.i.d}{\sim} \text{Inverse-Gamma}(a_\alpha, b_\alpha),\ l = 1, 2, ...K \\
&A_{ij} \overset{i.i.d}{\sim} N(0, \sigma_A^2), i = 1, ..., K-1, j = 1, ..., K-1 \\
&\sigma_q^2 \sim \text{Inverse-Gamma}(a_q, b_q), \;\; q \in \{a, M, E\}
\end{aligned}$$

The model parameters are then: $\boldsymbol{\Theta} = \{\boldsymbol{Z}, \{\boldsymbol{\mu}^A(1), \boldsymbol{\mu}^A(2), ..., \boldsymbol{\mu}^A(T)\}, \{\alpha_1, \alpha_2, ..., \alpha_k\}, \sigma_E^2, \sigma_M^2, \{S_j(t), t = 1, 2, ..., T, j = 1, 2, ..., P\}, \beta, \boldsymbol{A}, \sigma_a^2\}$ with prior distributions fully determined after specification of $a_E, b_E, a_M, b_M, a_\alpha, b_\alpha, \sigma_A^2, a_a, b_a, \sigma_{\mu_1}^2$, and these hyper-parameters are chosen so that the resulting priors are somewhat diffuse. The subscript a denotes 'active state' and $S_j(t)$ is the unknown neural activity for location j at time point t.

The posterior distribution takes the form $P(\boldsymbol{\Theta}|\boldsymbol{E}, \boldsymbol{M}) = P(\boldsymbol{\Theta}, \boldsymbol{E}, \boldsymbol{M})/P(\boldsymbol{E}, \boldsymbol{M})$

where

$$
\begin{aligned}
&P(\boldsymbol{\Theta}, \boldsymbol{E}, \boldsymbol{M}) = P(\boldsymbol{E}, \boldsymbol{M}|\boldsymbol{\Theta})P(\boldsymbol{\Theta}) = P(\boldsymbol{E}|\boldsymbol{\Theta})P(\boldsymbol{M}|\boldsymbol{\Theta})P(\boldsymbol{\Theta}) \\
&= \prod_{t=1}^{T} MVN(\boldsymbol{E}(t); \boldsymbol{X}_E \boldsymbol{S}(t), \sigma_E^2 \boldsymbol{H}_E) \times MVN(\boldsymbol{M}(t); \boldsymbol{X}_M \boldsymbol{S}(t), \sigma_M^2 \boldsymbol{H}_M) \\
&\times IG(\sigma_E^2; a_E, b_E) \times IG(\sigma_M^2; a_M, b_M) \times \left[\prod_{j=1}^{p} \prod_{t=1}^{T} \prod_{l=1}^{K} N(S_j(t); \mu_l(t), \alpha_l)^{Z_{v(j)l}} \right] \\
&\times \left[\prod_{t=2}^{T} MVN(\boldsymbol{\mu}^A(t); \boldsymbol{A}\mu^A(t-1), \sigma_a^2 \boldsymbol{I}) \right] \times MVN(\boldsymbol{\mu}^A(1); \boldsymbol{0}, \sigma_{\mu_1}^2 \boldsymbol{I}) \\
&\times \text{Potts}(\boldsymbol{Z}; \beta) \times \prod_{l=1}^{K} IG(\alpha_l; a_\alpha, b_\alpha) \times \text{Unif}(\beta; 0, \beta_{crit}) \\
&\times \left[\prod_{i=1}^{K-1} \prod_{j=1}^{K-1} N(A_{ij}; 0, \sigma_A^2) \right] \times IG(\sigma_a^2; a_a, b_a)
\end{aligned}
$$

$MVN(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the density of the $\dim(\mathbf{x})$-dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ evaluated at $\mathbf{x}$; $IG(x; a, b)$ denotes the density of the inverse-gamma distribution with parameters a and b evaluated at x; $N(x; \mu, \sigma^2)$ denotes the density of the normal distribution with mean μ and variance σ^2 evaluated at x; $\text{Potts}(\boldsymbol{Z}; \beta)$ is the joint probability mass function of the Potts model with parameter β evaluated at $\boldsymbol{Z}$; $\text{Unif}(x; a, b)$ is the density of the uniform distribution on $[a, b]$ evaluated at x.

Considering the scrambled face dataset presented in Figure 1, with $T = 161$ time points and $P = 8,196$ brain locations selected on the cortex, the dimension of $\boldsymbol{S} = (\boldsymbol{S}(1)', \ldots, \boldsymbol{S}(T)')'$ is $1,319,556$ and this high-dimensional parameter space poses challenges for Bayesian computation. In addition, the parameter space includes the discrete-valued mixture allocation variables $\boldsymbol{Z}$ and a large number of such variables will cause problems for standard MCMC sampling algorithms typically used for the required computation. We must therefore consider some approximations, and in the following section we discuss simultaneous point estimation for $\boldsymbol{\Theta}$ and model selection for the number of mixture components in the latent process using an algorithm that makes the required computation feasible and relatively efficient.

2.3 Computation and Estimation the Number of Mixture Components

The iterated conditional modes (ICM) algorithm (Besag, 1986) is a relatively straightforward approach for computing a point estimate for $\boldsymbol{\Theta}$ and has a long history of application to image restoration. The properties of the algorithm within this context were first described by Besag (1986). Iterated conditional modes is a deterministic algorithm used to obtain a configuration of a local maximum of the joint probability. It does this by iteratively maximizing the probability of each variable conditioned on the rest. The algorithm proceeds by partitioning $\boldsymbol{\Theta}$ into a set of convenient blocks $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_W\}$, and is iterative, where, given an initial value $\boldsymbol{\Theta}^{(0)}$, the i^{th} iteration proceeds by cycling through each of the blocks $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_W$ in turn, and updating block $\boldsymbol{\Theta}_j$ by maximizing the corresponding full conditional distribution. That is, the j^{th} sub-step of the i^{th} iteration is based on the following equation

$$\boldsymbol{\Theta}_j^{(i)} = \underset{\boldsymbol{\Theta}_j}{\operatorname{argmax}}\, P(\boldsymbol{\Theta}_j | \boldsymbol{E}, \boldsymbol{M}, \boldsymbol{\Theta}_1^{(i)}, \ldots, \boldsymbol{\Theta}_{j-1}^{(i)}, \boldsymbol{\Theta}_{j+1}^{(i-1)}, \boldsymbol{\Theta}_W^{(i-1)})$$

where the objective function is the probability mass/density function of the corresponding full conditional distribution $[\boldsymbol{\Theta}_j | \boldsymbol{E}, \boldsymbol{M}, \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_{j-1}, \boldsymbol{\Theta}_{j+1}, \boldsymbol{\Theta}_W]$.

Within our ICM algorithm the update for the spatial cohesion parameter β requires repeated evaluation of the normalizing constant $G(\beta)$ in the Potts model. It is well known that evaluating this normalizing constant requires fairly extensive computation and this computation is often approached using thermodynamic integration (Johnson et al., 2013). Thermodynamic integration provides one computational approach to approximating the marginal likelihood and was described in Gelman and Meng (1998). We avoid thermodynamic integration by using the pseudolikelihood approximation

$$\text{Potts}(\boldsymbol{Z}; \beta) \approx \prod_{i=1}^{N_v} P(\boldsymbol{Z}_i | \boldsymbol{Z}_{-i}, \beta) = \prod_{i=1}^{N_v} \frac{\exp\left(2\beta \sum_{j=1}^{k} Z_{ij} \sum_{l \in \delta_i} Z_{lj}\right)}{\sum_{q=1}^{k} \exp(2\beta \sum_{l \in \delta_i} Z_{lq})},$$

where δ_i denotes the set of indices corresponding to the neighbours of voxel i. While this approximation will incur some bias, it allows us to avoid lengthy computations.

Within the framework of the ICM algorithm there are several options for updating the labelling process $\boldsymbol{Z}$. In the simplest case, the variable associated with each indi-

vidual voxel $\boldsymbol{Z}_j$ is updated one-by-one in sequence, $j = 1, \ldots, n_v$. We adopt a more efficient approach that begins with partitioning $\boldsymbol{Z}$ into two blocks $\boldsymbol{Z} = \{\boldsymbol{Z}_W, \boldsymbol{Z}_B\}$ according to a 3-dimensional chequerboard arrangement, where $\boldsymbol{Z}_W$ corresponds to the 'white' voxels and $\boldsymbol{Z}_B$ corresponds to the 'black' voxels. Under the Markov random field prior with a first-order neighbourhood structure, the elements of $\boldsymbol{Z}_W$ are conditionally independent given $\boldsymbol{Z}_B$, the remaining parameters, and the data $\boldsymbol{E}$, $\boldsymbol{M}$. This allows us to update $\boldsymbol{Z}_W$ in a single step which involves simultaneously updating its elements from their full conditional distributions, and this updating can be implemented using multiple cores. Even without multiple cores, this scheme will typically result in faster convergence when compared with the sequential one-by-one updates (Ge et al., 2014). The variables $\boldsymbol{Z}_B$ are updated in the same way. This idea was recently employed by Ge et al. (2014) within the context of a Metropolis-Hastings algorithm.

As described in Besag (1986), the algorithm converges rapidly to a point estimate; however, this estimate and convergence of the algorithm will depend on the initial value of $\boldsymbol{\Theta}$. A careful choice for the initial value is thus important. To obtain the initial values for $\boldsymbol{S}$ we used regularized least squares with either a ridge or lasso penalty. The estimates for $\boldsymbol{S}_j = (S_j(1), \ldots, S_j(T))', \ j = 1, \ldots, P$ were then clustered into K groups using a K-means algorithm and these groups are then used to assign the initial values for $\boldsymbol{Z}$. Initial values for $\mu_j(t)$ were then obtained by taking the average of the initial $\boldsymbol{S}$ values assigned to each of the mixture components. The initial value for β was drawn from its prior distribution, and the initial values for the remaining parameters are set according to the mode of the associated prior.

To improve the quality of the final estimates we have found it useful to apply smoothing to the estimated source time series $\hat{S}_j(t), \ t = 1, \ldots, T$ at each location. This is accomplished by using local polynomial regression implemented via the *loess* function in the R programming language (R Development Core Team, 2017).

To reduce the dimension of the parameter space and as a result the required computation time, we use the K-means algorithm to cluster the P locations on the cortex into a smaller number of $J \leq P$ clusters, and assume that $S_j(t) = S_l(t)$ for cortical locations l, j belonging to the same cluster. Typical values for J are $J = 250, 500, 1000$, and these choices are investigated in our simulation studies.

While the value of K, the number of mixture components in equation (2.2), is fixed with no prior assigned to it, it will typically be the case that the number of mixture components will not be known beforehand. For an evoked response study,

our model specification assumes that the number of mixture components is small, no more than 10, but likely less. From our ICM algorithm we can obtain a simple estimate for the number of mixture components based on the estimated allocation variables $\hat{\boldsymbol{Z}}$. In particular, we run the algorithm with a value of K that is considerably larger than the expected number of mixture components. For example, the value of K can be set as $K = 15$ when running the algorithm. The j^{th} location on the cortex is allocated to one of the mixture components based on the value of $\hat{\boldsymbol{Z}}_{v(j)}$, where $\hat{\boldsymbol{Z}}_{v(j)} = (\hat{Z}_{v(j)_1}, \hat{Z}_{v(j)_2}, \ldots, \hat{Z}_{v(j)_K})'$ and $\hat{Z}_{v(j)_l} = 1$ if location j is allocated to component $l \in \{1, \ldots, K\}$. When the algorithm is run with a value of K that is larger than necessary, there will exist empty mixture components that will not be assigned any locations under $\hat{\boldsymbol{Z}}$. In a sense these empty components are automatically pruned out as redundant. The estimated number of mixture components is then obtained as follows:

$$\hat{K}_{ICM} = \sum_{l=1}^{K} I\{\sum_{v=1}^{n_v} \hat{Z}_{v_l} \neq 0\}. \tag{2.3}$$

This estimator is very simple and only requires us to run the ICM algorithm once for a single value of K. Multiple runs of the algorithm with different values of K are avoided altogether. Properties of the estimator $\hat{K}_{ICM}$ are investigated in Section 2.4.

The overall estimation procedure is presented in Algorithm 1 and the corresponding derivations are presented in Section A.2 . Convergence of the algorithm is monitored by examining the relative change of the Frobenius norm of the estimated sources on consecutive iterations. In Section A.3, we investigate the performance and correctness of our proposed model and algorithm on a number of test cases using synthetic data where the truth is known. The correctness is measured by comparing our estimated locations and dynamics of neural activities to the truth to see if they match with each other.

2.4 Simulation Studies

2.4.1 Evaluation of Neural Source Estimation

We evaluated the quality of the source estimates though a simulation study as the number of activated brain regions change, and we make comparisons between our approach with and without smoothing, and the mesostate-space model (MSM) of Daunizeau and Friston (2007). As with the examples of Section A.3, we generated

Algorithm 1 - ICM Algorithm for Potts Mixture Model. '$\leftarrow$' denotes assigning value.

$\Theta \leftarrow$ Initial Value

Converged $\leftarrow 0$

while Converged $= 0$ **do**

$\sigma_M^2 \leftarrow \Big[\sum_{t=1}^{T} \frac{1}{2}\big(\boldsymbol{M}(t) - \boldsymbol{X}_M \boldsymbol{S}(t)\big)' \boldsymbol{H}_M^{-1}\big(\boldsymbol{M}(t) - \boldsymbol{X}_M \boldsymbol{S}(t)\big) + b_M\Big] / \Big[a_M + \frac{TN_M}{2} + 1\Big]$

$\sigma_E^2 \leftarrow \Big[\sum_{t=1}^{T} \frac{1}{2}\big(\boldsymbol{E}(t) - \boldsymbol{X}_E \boldsymbol{S}(t)\big)' \boldsymbol{H}_E^{-1}\big(\boldsymbol{E}(t) - \boldsymbol{X}_E \boldsymbol{S}(t)\big) + b_E\Big] / \Big[a_E + \frac{TN_E}{2} + 1\Big]$

$\sigma_a^2 \leftarrow \Big[\sum_{t=2}^{T} \frac{1}{2}(\boldsymbol{\mu}^A(t) - \boldsymbol{A}\boldsymbol{\mu}^A(t-1))'(\boldsymbol{\mu}^A(t) - \boldsymbol{A}\boldsymbol{\mu}^A(t-1)) + b_a\Big] / \Big[a_a + \frac{(T-1)(K-1)}{2} + 1\Big]$

$vec(\boldsymbol{A}) \leftarrow \left(\frac{1}{\sigma_a^2}\Big(\sum_{t=2}^{T} \boldsymbol{\mu}^A(t)' \boldsymbol{Kr_t}\Big) \times \boldsymbol{C}_1^{-1}\right)'$, where $\boldsymbol{C}_1 = \frac{1}{\sigma_A^2}\boldsymbol{I}_{(K-1)^2} + \frac{1}{\sigma_a^2}\Big(\sum_{t=2}^{T} \boldsymbol{Kr_t}' \boldsymbol{Kr_t}\Big)$, and $\boldsymbol{Kr_t} = \Big(\boldsymbol{\mu}^A(t-1)' \otimes \boldsymbol{I}_{K-1}\Big)$

for $l = 1, ..., K$ **do**

$\alpha_l \leftarrow \Big[\frac{\sum_{j=1}^{P}\sum_{t=1}^{T} Z_{v(j)l}(S_j(t) - \mu_l(t))^2}{2} + b_\alpha\Big] / \Big[\frac{T\sum_{j=1}^{P} Z_{v(j)l}}{2} + a_\alpha + 1\Big]$

end for

$\boldsymbol{\mu}(1) \leftarrow \left(\Big(\sum_{j=1}^{P}(S_j(1)\vec{I}_{K-1})' \boldsymbol{D}_j + \frac{1}{\sigma_a^2}\boldsymbol{\mu}^A(2)'\boldsymbol{A}\Big) \times \boldsymbol{B}_1^{-1}\right)'$, where $\boldsymbol{B}_1 = \sum_{j=1}^{P} \boldsymbol{D}_j + \frac{1}{\sigma_a^2}\boldsymbol{A'A} + \frac{1}{\sigma_{\mu_1}^2}\boldsymbol{I}_{K-1}$, $\boldsymbol{D}_j = \mathrm{Diag}(\frac{Z_{v(j)l}}{\alpha_l}, l = 2, ..., K)$, $\vec{I}_{K-1} = (1, 1, \ldots, 1)'$ with dim $(\vec{I}_{K-1}) = K - 1$

for $t = 2, ..., T-1$ **do**

$\boldsymbol{\mu}(t) \leftarrow \left(\Big(\sum_{j=1}^{P}(S_j(t)\vec{I}_{K-1})' \boldsymbol{D}_j + \frac{1}{\sigma_a^2}(\boldsymbol{\mu}^A(t+1))'\boldsymbol{A} + \frac{1}{\sigma_a^2}(\boldsymbol{\mu}^A(t-1)'\boldsymbol{A}')\Big) \times \boldsymbol{B}_2^{-1}\right)'$

where $\boldsymbol{B}_2 = \sum_{j=1}^{P} \boldsymbol{D}_j + \frac{1}{\sigma_a^2}(\boldsymbol{A'A} + \boldsymbol{I}_{K-1})$

end for

$\boldsymbol{\mu}(T) \leftarrow \left(\Big(\sum_{j=1}^{P}(S_j(T)\vec{I}_{K-1})' \boldsymbol{D}_j + \frac{1}{\sigma_a^2}(\boldsymbol{\mu}^A(T-1)'\boldsymbol{A}')\Big) \times \boldsymbol{B}_3^{-1}\right)'$

where $\boldsymbol{B}_3 = \sum_{j=1}^{P} \boldsymbol{D}_j + \frac{1}{\sigma_a^2}\boldsymbol{I}_{k-1}$

for $j = 1, ..., P$ **do**

$\boldsymbol{S}_j \leftarrow -\frac{1}{2}\boldsymbol{\Sigma}_{S_j}\boldsymbol{W}_{2j}$ ▷ $\boldsymbol{S}_j = (S_j(1), S_j(2), ..., S_j(T))'$

$\boldsymbol{\Sigma}_{S_j}^{-1} = W_{1j}\boldsymbol{I}_T$, $\boldsymbol{W}'_{2j} = (W_{2j}(1), W_{2j}(2), ..., W_{2j}(T))$

where $W_{1j} = \frac{1}{\sigma_M^2}\Big(\boldsymbol{X}_M[,j]'\boldsymbol{H}_M^{-1}\boldsymbol{X}_M[,j]\Big) + \frac{1}{\sigma_E^2}\Big(\boldsymbol{X}_E[,j]'\boldsymbol{H}_E^{-1}\boldsymbol{X}_E[,j]\Big) + \sum_{l=1}^{K}\frac{Z_{v(j)l}}{\alpha_l}$

$W_{2j}(t) = \frac{1}{\sigma_M^2}\Big(-2\boldsymbol{M}(t)'\boldsymbol{H}_M^{-1}\boldsymbol{X}_M[,j] + 2(\sum_{v\neq j}\boldsymbol{X}_M[,v]S_v(t))'\boldsymbol{H}_M^{-1}\boldsymbol{X}_M[,j]\Big)$

$+ \frac{1}{\sigma_E^2}\Big(-2\boldsymbol{E}(t)'\boldsymbol{H}_E^{-1}\boldsymbol{X}_E[,j] + 2(\sum_{v\neq j}\boldsymbol{X}_E[,v]S_v(t))'\boldsymbol{H}_E^{-1}\boldsymbol{X}_E[,j]\Big) - 2\sum_{l=1}^{K}\frac{\mu_l(t)}{\alpha_l}$

$\boldsymbol{X}_M[,j]$, $\boldsymbol{X}_E[,j]$ denote the jth column of $\boldsymbol{X}_E$ and $\boldsymbol{X}_M$

end for

19: Let $\mathbb{B}$ denote the indices for 'black' voxels and $\mathbb{W}$ denote the indices for 'white' voxels.

20: **for** $\kappa \in \mathbb{B}$ *simultaneously* **do**

21: $Z_{\kappa q} \leftarrow 1$ and $Z_{\kappa l} \leftarrow 0,\ \forall l \neq q$ where $q = \text{argmax}_{h \in \{1,\ldots,K\}} P(h)$, and

22:
$$P(h) = \frac{\alpha_h^{-TN_{j\kappa}/2} \times \exp\left(-\frac{1}{2}\sum_{j|v(j)=\kappa} \alpha_h^{-1} \sum_{t=1}^{T} (S_j(t) - \mu_h(t))^2 + 2\beta \sum_{v \in \delta_\kappa} Z_{vh}\right)}{\sum_{l=1}^{K} \alpha_l^{-TN_{j\kappa}/2} \times \exp\left(-\frac{1}{2}\sum_{j|v(j)=\kappa} \alpha_l^{-1} \sum_{t=1}^{T} (S_j(t) - \mu_l(t))^2 + 2\beta \sum_{v \in \delta_\kappa} Z_{vl}\right)}$$

where $N_{j\kappa}$ is the number of cortical locations contained in voxel κ.

23: **end for**

24: **for** $\kappa \in \mathbb{W}$ *simultaneously* **do**

25: $Z_{\kappa q} \leftarrow 1$ and $Z_{\kappa l} \leftarrow 0,\ \forall l \neq q$ where $q = \text{argmax}_{h \in \{1,\ldots,K\}} P(h)$, and

26:
$$P(h) = \frac{\alpha_h^{-TN_{j\kappa}/2} \times \exp\left(-\frac{1}{2}\sum_{j|v(j)=\kappa} \alpha_h^{-1} \sum_{t=1}^{T} (S_j(t) - \mu_h(t))^2 + 2\beta \sum_{v \in \delta_\kappa} Z_{vh}\right)}{\sum_{l=1}^{K} \alpha_l^{-TN_{j\kappa}/2} \times \exp\left(-\frac{1}{2}\sum_{j|v(j)=\kappa} \alpha_l^{-1} \sum_{t=1}^{T} (S_j(t) - \mu_l(t))^2 + 2\beta \sum_{v \in \delta_\kappa} Z_{vl}\right)}$$

where $N_{j\kappa}$ is the number of cortical locations contained in voxel κ.

27: **end for**

28: $\hat{\beta} \leftarrow \text{argmax}_{\beta \in [0,\, \beta_{crit}]} \Psi(\beta)$, where

$$\Psi(\beta) = 2\beta \sum_{i=1}^{N_v} \sum_{j=1}^{k} Z_{ij} \sum_{l \in \delta_i} Z_{lj} - \sum_{i=1}^{N_v} \log\{\sum_{q=1}^{k} \exp(2\beta \sum_{l \in \delta_i} Z_{lq})\}$$

29: Check for convergence. Set Converged = 1 if so.

30: **end while**

both MEG and EEG data based on neural activity at 8,196 locations on the cortex, and this activity is projected onto the sensor arrays using the forward operators $\boldsymbol{X}_M$ and $\boldsymbol{X}_E$, with Gaussian noise added at each sensor.

For this study we set the number of mixture components K to be the true number of latent states (either two, three, four, or nine) in both our model as well as the MSM, so that fixing K in this study does not give either approach an advantage over the other. In the next section we present another simulation study where we focused on estimating the number of mixture components and evaluating the sampling distribution of $\hat{K}_{ICM}$. In this study, we considered four scenarios with two, three, four, and nine latent states, and in each case one of these states was inactive, while the other states have activity generated by Gaussian signals. In the simplest case we have only a single activated region, and this region is depicted in Figure A.1. The temporal signal arising from locations contained in the activated region for this case is depicted in Figure A.2, panel (a). The other three cases have two, three, and eight activated regions, and these regions are depicted in Figure A.3, panels (a) and (c), for the case of two and three active regions, and Figure A.5, for the case of eight active regions, while the corresponding temporal signals associated with the activated regions are depicted in Figure A.4, panels (a), (c) and (g).

In each case we simulated 500 replicate datasets and each of the four approaches was applied. For each replicate we estimated the correlation between the estimated sources and the true sources $\text{Corr}[(\boldsymbol{S}(1)', \boldsymbol{S}(2)', \ldots, \boldsymbol{S}(T)'), (\hat{\boldsymbol{S}}(1)', \hat{\boldsymbol{S}}(2)', \ldots, \hat{\boldsymbol{S}}(T)')]$ as a measure of agreement, and this measure was averaged over the 500 replicate datasets. For each simulated dataset we applied our algorithm with $J = 250, 500, 1000$ clusters so as to evaluate how the performance varies as this tuning parameter changes. In addition, we made comparisons between these methods based on the mean-squared error of $\hat{S}_j(t)$. In particular, for each brain location j and time point t we estimated the mean-squared error (MSE) of the estimator $\hat{S}_j(t)$ based on the $R = 500$ simulation replicates. These MSE's were then totalled over brain locations and time points in order to obtain the Total MSE (TMSE). This total was obtained separately for locations in active regions and then for the inactive region, where the active regions are depicted in Figure A.1, left column of Figure A.3 and Figure A.5.

The average correlation between the estimated values and the truth for each of the distinct settings in our study is presented in Table 2.1. Examining Table 2.1, we see that for most of the cases considered our Potts-mixture model, both with and without smoothing, yields a higher average correlation than the MSM with either

EEG or MEG. With respect to average correlation, we also see that smoothing does not improve the correlation obtained when using the Potts-mixture model. With respect to the number of clusters J, we find that using a lower number of clusters results in a higher average correlation, even in the case where $K = 9$.

The Total MSE's for the same distinct cases are depicted in Table 2.2. From the results in this table, we make several key observations :

1. When we consider active regions and our approach with a differing number of clusters, we observe that using $J = 250$ clusters yields the lowest TMSE compared with the alternatives of $J = 500$ or $J = 1000$ clusters. In addition, for the best case corresponding to $J = 250$, we see that incorporating temporal smoothing after the ICM algorithm yields optimal TMSE values among those settings considered for our approach. We also notice that these optimal TMSE values obtained from our method are uniformly lower than the TMSE obtained from MSM-EEG and MSM-MEG for all values of K.

2. When we consider inactive regions, it is clear that the MSM with either modality has lower total mean-squared error than the Potts-mixture model in all cases. This might be caused by the clustering scheme in the Potts-mixture model such that inactive regions are mistakenly assigned to active.

3. When we consider inactive regions for the specific case where $K = 9$ our approach with $J = 250$ clusters yields a TMSE that is very large. In this case, the TMSE drops significantly when the number of clusters is increased from 250 to 500 or 1000 and we see here an advantage to using a larger number of clusters.

The results for the case where $K = 9$ in Table 2.1 (Average Correlation) may at first seem contradictory to those for $K = 9$ in Table 2.2 (TMSE), where we see our approach outperforming MSM in the former and underperforming in the latter. This contradiction can be explained when considering scale. In particular, our approach appears to estimate the patterns of spatiotemporal activation more accurately than MSM in this most complicated setting leading to the higher correlation measure, while the scale of the true sources appears to be better estimated by MSM leading to the improved TMSE. It is our view that the patterns of activation are far more important than the scale.

Finally, our separation of TMSE into active and inactive regions has led a reviewer to suggest that we examine the false discovery rate of the methods being compared

in this simulation study. To be specific, let FP denote the number of vertices that are declared active by the model that are in fact truly inactive. Let TP denote the number of vertices that are declared active by the model that are in fact truly active. Then $FP + TP$ is the total number of vertices declared active. For a given dataset, the false positive rate is $p_{FP} = FP/(FP + TP)$. We have computed this quantity for each simulation replicate and then taken the average false positive rate over all simulation replicates. In a similar way, let FN denote the number of vertices declared inactive by the model that are in fact active. Let TN denote the number of vertices declared inactive by the model that are in fact inactive. The false negative rate, for a given dataset, is then $p_{FN} = FN/(FN + TN)$ and we calculated the average of this quantity over simulation replicates. The results are presented in Table 2.3.

Examining Table 2.3, we first note that the false positive rate associated with MSM is in general unacceptably high, with over 60% (and in many cases much higher) of the vertices labelled as active by MSM being inactive in all cases. The extreme false positive rate found for MSM is a result that we find particularly interesting. We note that the average false positive rate of 0.641 observed for MSM (MEG) when $K = 3$ is lower than the other values observed for MSM (MEG) and MSM (EEG) because the average is pulled down by a number of very small values in some of replicates. The performance of our approach with respect to false discoveries is considerably better than that of MSM. With respect to the false negative rate, the approaches perform equally well. However, MSM (EEG) has the best performance when the number of active regions is low. Overall, these results demonstrate a significant improvement obtained from our methodology when considering false discoveries and roughly equal performance when considering false negative rates.

2.4.2 Evaluation of Mixture Component Estimation

We perform a simulation to evaluate the sampling distribution of $\hat{K}_{ICM}$. We consider the following scenarios:

1. Two latent states as depicted in Figure A.1 with the active regions having a Gaussian signal as depicted in Figure A.2, panel (a).

2. Three latent states as depicted in Figure A.3, panel (a) with the active regions having Gaussian signals as depicted in Figure A.4, panel (a).

Table 2.1: Simulation study I - Neural Source Estimation. Average (Ave.) correlation between the neural source estimates and the true values for the Potts-Mixture model, the Potts-Mixture model without local polynomial smoothing, the Mesostate-space model with MEG data, and the Mesostate-space model with EEG data. The simulation study is based on $R = 500$ simulation replicates where each replicate involves the simulation of MEG and EEG data based on a known configuration of the neural activity. For each replicate we show the average correlation compared as a measure of agreement and this correlation is then averaged over the $R = 500$ simulation replicates in order to obtain the Ave. Correlation. "NS" refers to no smoothing for Potts model.

		$K = 2$	$K = 3$	$K = 4$	$K = 9$
Method	Clusters	Ave. Corr.	Ave. Corr.	Ave. Corr.	Ave. Corr.
Potts-Mixture	250	0.59	0.62	0.61	0.50
Potts-Mixture (NS)	250	0.59	0.62	0.59	0.50
Potts-Mixture	500	0.51	0.50	0.43	0.49
Potts-Mixture (NS)	500	0.51	0.50	0.42	0.48
Potts-Mixture	1000	0.38	0.40	0.39	0.41
Potts-Mixture (NS)	1000	0.37	0.40	0.37	0.40
MSM (MEG)	NA	0.20	0.47	0.37	0.34
MSM (EEG)	NA	0.24	0.41	0.41	0.28

Table 2.2: Simulation study I - Neural Source Estimation. Total mean-squared error of the neural source estimators for the Potts-Mixture model, the Potts-Mixture model without local polynomial smoothing, the Mesostate-space model with MEG data, and the Mesostate-space model with EEG data. The simulation study is based on $R = 500$ simulation replicates where each replicate involves the simulation of MEG and EEG data based on a known configuration of the neural activity. For each brain location j and time point t we obtain (estimate) the mean-squared error (MSE) of the estimator of $S_j(t)$ based on the $R = 500$ simulation replicates. These MSE's are then totalled over brain locations and time points in order to obtain the Total MSE indicated in the table. This total is obtained separately for locations in active regions and then for the inactive region, where the active regions are depicted in the left column of Figure A.3, Figure A.1 and Figure A.5 ."NS" refers to no smoothing for the Potts model.

		$K = 2$		$K = 3$		$K = 4$		$K = 9$	
Method	Clusters	Active	Inactive	Active	Inactive	Active	Inactive	Active	Inactive
Potts-Mixture	250	86	139	557	479	802	662	2216	16900
Potts-Mixture (NS)	250	90	138	609	471	874	702	5318	16488
Potts-Mixture	500	207	200	847	356	1451	573	2346	1001
Potts-Mixture (NS)	500	209	201	1289	536	1669	711	2734	1340
Potts-Mixture	1000	267	235	1141	398	2170	676	3050	1017
Potts-Mixture (NS)	1000	274	239	1819	615	2448	793	3605	1337
MSM (MEG)	NA	380	19	842	55	1297	65	2345	357
MSM (EEG)	NA	325	71	795	172	1094	237	2499	665

Table 2.3: Simulation study I - Neural Source Estimation. False Positive Rate (p_{FP}) of estimating active state and False Negative Rate(p_{NP}) of estimating inactive state . "NS" refers to no smoothing for the Potts model.

		$K = 2$		$K = 3$		$K = 4$		$K = 9$	
Method	Clusters	p_{FP}	p_{NP}	p_{FP}	p_{NP}	p_{FP}	p_{NP}	p_{FP}	p_{NP}
Potts-Mixture	250	0.245	0.008	0.361	0.016	0.444	0.023	0.537	0.050
Potts-Mixture (NS)	250	0.245	0.008	0.401	0.014	0.446	0.022	0.534	0.050
Potts-Mixture	500	0.291	0.010	0.322	0.023	0.385	0.035	0.426	0.060
Potts-Mixture (NS)	500	0.290	0.010	0.338	0.021	0.379	0.034	0.426	0.060
Potts-Mixture	1000	0.340	0.011	0.329	0.027	0.417	0.040	0.425	0.062
Potts-Mixture (NS)	1000	0.340	0.011	0.341	0.025	0.418	0.040	0.422	0.062
MSM (MEG)	NA	0.922	0.003	0.641	0.019	0.879	0.011	0.914	0.025
MSM (EEG)	NA	0.966	0.000	0.922	0.005	0.898	0.020	0.917	0.033

3. Four latent states as depicted in Figure A.3, panel (c) with the active regions having Gaussian signals as depicted in Figure A.4, panel (c).

4. Four latent states as depicted in Figure A.3, panel (c) with the active regions having Gaussian signals and a sinusoid signal as depicted in Figure A.4, panel (e).

5. Nine latent states as depicted in Figure A.5 with the active regions having Gaussian signals as depicted in Figure A.4, panel (g).

For each of the five cases the data were simulated with 5% Gaussian noise added at the sensors as in the previous section, with 1000 replicate datasets used in each case. The ICM algorithm was run with $K = 10$ for each of the 5000 simulated datasets with the other settings for the model and algorithm set as in previous sections. For each dataset we computed the value of the estimator (2.3) and histograms illustrating the sampling distribution of $\hat{K}_{ICM}$ are presented in Figure 2.2, panels (a) - (e) for each of the five cases above.

For the first four cases we see that the mode of the sampling distribution corresponds exactly to the true number of latent states. In the final case where we have a larger number (nine) of latent states we see that $\hat{K}_{ICM}$ is biased and underestimates the number of latent states. This is not unexpected with a mixture model having a large number of components that are not well separated as the estimation method will merge two adjacent components into one estimated component leading to under-estimation.

We repeated the simulation study for the five cases above but made the estimation problem more difficult by reducing the separation of the Gaussian signals. The modified temporal signals are depicted in Figure A.2. The sampling distribution of $\hat{K}_{ICM}$ for each of the five cases under the less well-separated setup are depicted in Figure 2.2, (f) - (j). In this case the mode of the sampling distribution does not correspond to the true number of latent states except the first case and we see that $\hat{K}_{ICM}$ tends to under-estimate the true number of states, with the mode of the sampling distribution being one-less than the true value in the case (2), (3) and (4) and we see fairly substantial under-estimation in the fifth case.

Overall, given the complexity of our model, the inherent difficulty of the problem of estimating the number of components in a high-dimensional spatial latent Gaussian mixture model, the simplicity of the proposed estimator, and its computational

efficiency, we find the performance of $\hat{K}_{ICM}$ satisfactory, though we acknowledge the tendency of the estimator to under-estimate the true number of mixture components in some settings.

2.5 Electromagnetic Brain Mapping of Scrambled Faces

We applied our algorithm to the MEG and EEG data presented in Figure 2.1, to reconstruct the associated neural activity which is constrained to lie on the cortical surface. The data are from an experiment where a single subject is repeatedly presented with pictures of scrambled faces while required to make a symmetry judegement. The experiment and related analyses are described in detail in Henson et al. (2003, 2007, 2009a, 2009b, 2010). Beginning with pictures of faces, each scrambled face is created from a single picture by 2D Fourier transformation, random phase permutation, inverse transformation and outline-masking of each face. Thus the scrambled faces are closely matched with the corresponding faces for low-level visual properties.

The experiment involves a sequence of trials each lasting 1800ms, where in each trial the subject is presented with one of the pictures for a period of 600ms while being required to make a four-way, left-right symmetry judgment while brain activity is recorded over the array. Both scrambled faces and unscrambled faces are presented to the subject; however, our analysis will focus only on trials involving scrambled faces. This produces a multivariate time series for each trial, and the trial-specific time series are then averaged across trials to create a single multivariate time series. In the case of EEG data, the average evoked response is an average of 344 trials while for MEG data, the average evoked response is an average of 336 trials. The degree of inter-trial variability is quite low. This experiment is conducted while EEG data are recorded, and then again on the same subject while MEG data are recorded. The EEG data were acquired on a 128-sensors ActiveTwo system, sampled at 2048 Hz and subsequently downsampled to 200 Hz. The resulting average evoked response to scrambled faces is depicted in Figure 2.1 (c). The MEG data were acquired on 274 sensors with a CTF/VSM system, sampled at 480 Hz and subsequently downsampled to 200 Hz. The resulting average evoked response to scrambled faces is depicted in Figure 2.1 (a). In total, each average evoked response covers roughly 805ms leading to $T = 161$ time points, where the 40^{th} time point $t = 40$ corresponds to the time

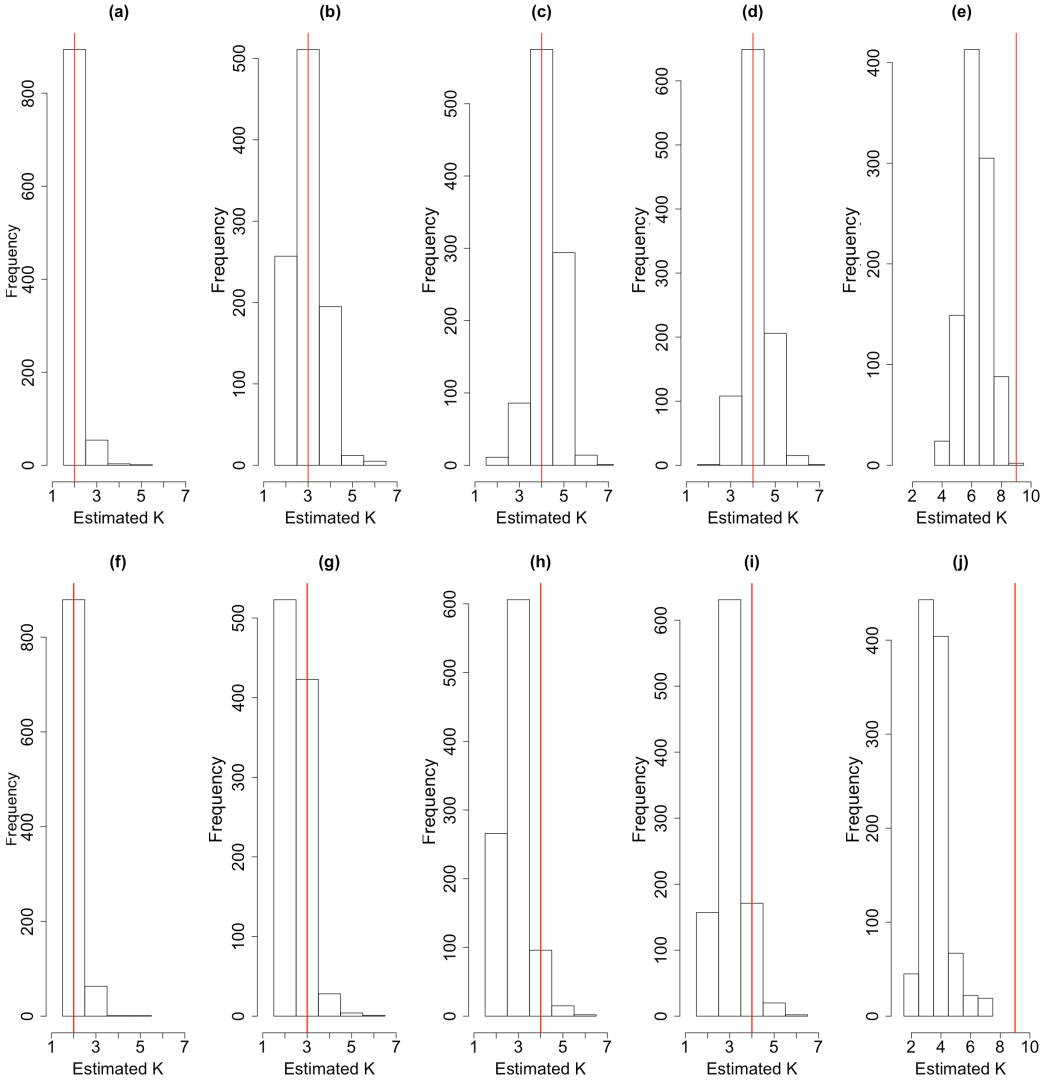


Figure 2.2: Histograms illustrating the sampling distribution of $\hat{K}_{ICM}$ obtained in the simulation study of Section 2.4. The first row corresponds to the case where the true signals are well-separated (these signals are depicted in Figure A.4, left column); (a&f), $K = 2$; (b&g), $K = 3$; (c&h), $K = 4$ with three Gaussian signals; (d&i), $K = 4$ with two Gaussian signals and one sinusoid; (e&j), $K = 9$ with eight Gaussian signals. The second row corresponds to the case where the true signals are less well-separated depicted in Figure A.2. In each case the vertical red line indicates the true number of latent states underlying the simulated data.

at which the stimulus was presented (the red vertical line in Figure 2.1). Averaging across trials is considered standard practice to increase the signal to noise ratio, but is based on the assumption that the process of interest is stationary across trials. This averaging also increases the viability of the Gaussian assumption on the data via the central limit theorem.

In the EEG data we see three peaks after the presentation of the stimulus at roughly $t = 60$ (100ms after stimulus), $t = 75$ (175ms after stimulus), and $t = 85$ (225ms after stimulus). In the MEG data we see two peaks after the presentation of the stimulus at roughly $t = 60$ (100ms after stimulus) and $t = 70$ (150ms after stimulus). In order to capture the actual neural response of interest, that is, the response to the observation of scrambled faces while the subject makes a symmetry judgment, we used a temporal segment of the data from time point $t = 50$ to $t = 100$, indicated by the vertical dashed lines in Figure 2.1, panel (a) and panel (c). The algorithm is run with $K = 10$, $n_v = 560$ voxels, and $J = 250$ clusters with initial values selected as described in Section 2.3.

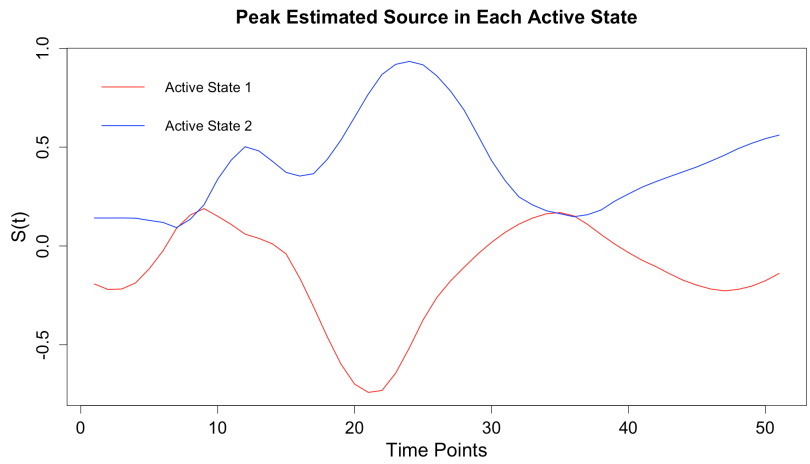


Figure 2.3: Brain Activation for Scrambled Faces - Peak source $\hat{S}_j(t)$ in each of the two active states.

From the algorithm output we found the estimated number of states to be $\hat{K}_{ICM} = 3$, indicating that there are two active states. For each of the two active states, we determined the location at which the corresponding source activity $\hat{S}_j(t)$ has the highest power. The estimated curves for each of these two locations is depicted in Figure 2.3. The first active state contains a large negative peak, while the second active state contains two peaks including a large positive peak. In both states the

largest peak (at these two high power locations) occurs within the vicinity of 175ms after the presentation of the stimulus. The large peak in the signal from the first state occurs slightly before the large peak in the signal from the second state.

The spatial patterns of the estimated neural sources $\hat{\boldsymbol{S}}(t)$ are of primary interest. The overall power ($\sum_{t=1}^{T} \hat{S}_j(t)^2$) of these estimated sources obtained from our model at each brain location j is mapped onto the cortex in the first row of Figure 2.4. Examining these results we see that the greatest power occurs on the bilateral ventral occipital cortex and the occipital cortex. More specifically, the highest power signals on both hemispheres seem to arise within Brodman areas 18 and 19 which are visual association areas (Baars and Gage, 2012). Brodman area 18 is responsible for the interpretation of images while Brodman area 19 has feature-extracting, shape recognition, attentional, and multimodal integrating functions. In general, the power map seems to represent regions that would be expected to show scrambled face-related activity. For example, Daunizeau and Friston (2007) analyze the EEG response to scrambled faces for a single subject and also found regions with high probability of being active on the bilateral ventral occipital cortex and the occipital cortex. These authors also found a region with high probability of being active in the right frontal lobe; our analysis based on the combined MEG and EEG data does not detect high power in this region.

For comparison, we also applied MSM to the MEG data only, and then applied MSM to the EEG data only. The corresponding results obtained from MSM-MEG are depicted in the second row of Figure 2.4. Broadly speaking, MSM-MEG seems to indicate similar results to those obtained from our model, in particular with respect to activation on the bilateral ventral occipital cortex, Brodman areas 18 and 19. Interestingly, the results from MSM-EEG, depicted in third row of Figure 2.4, differ strongly when compared with results of MSM-MEG and our model. In particular, the spatial spread of the high power regions on the ventral occipital cortex, Brodman area 18, is considerably smaller and Brodman area 19 is not indicated. Importantly, MSM-EEG also detects high power in a region on the right frontal lobe, Brodman area 8 (Baars and Gage, 2012), which is involved in the management of uncertainty. Recall, that the experimental paradigm requires that the subject make a symmetry judgment when presented with a scrambled face, and there may be uncertainty associated with this judgement. Relating this back to our first simulation study, we observed that MSM-EEG tends to outperform MSM-MEG in the active regions of the brain. From this observation, in addition to the previous results found in Daunizeau and Friston

(2007), we suspect that the high power detected in the right frontal lobe by MSM-EEG is a true neural signal that has not been detected by our model. Ideally, our results would show some combination of the results found by MSM-MEG and MSM-EEG, but in this particular case our solution seems to align well only with MSM-MEG.

We note that there are more MEG than EEG sensors so that there is more MEG data than EEG data in this case; however, the weighting of the two modalities depends on the relative values chosen for the two standard deviation parameters σ_E^2 and σ_M^2. Our approach to tuning these parameters has been through maximizing the posterior distribution using the ICM algorithm though alternative approaches such as cross-validation could be used to estimate these parameters.

In Figure 2.5 we examine cortical maps of $|\hat{S}_j(t)|$ at three peak time points $t = 50+10$ (100ms after presentation of the stimulus), $t = 50 + 25$ (175ms after presentation of the stimulus), and $t = 50 + 35$ (225ms after presentation of the stimulus). For each of the three selected time points we see that highest activity is observed on the left ventral surface; more specifically at Brodman area 19 at the first two peaks, and then moving into the perirhinal cortex at the third peak. The perirhinal cortex is involved in visual perception. The activation on the right ventral surface is relatively consistent across the three time points with peak activation in Brodman areas 18 and 19.

While it is not of specific interest to our brain mapping application we note that the spatial dependence parameter β of the Potts model is estimated at $\hat{\beta} = 0.44$, which is right at the upper boundary of our restricted parameter space based on the approximate phase transition point of the Potts model. Interestingly, we have found this phenomenon, that is, $\hat{\beta} = \beta_{crit}$, to occur with many datasets in the simulation studies.

2.5.1 Goodness-of-Fit to the Scrambled Faces MEG and EEG Data

Finally, to examine the goodness-of-fit for the model we compute the residuals $\hat{\boldsymbol{\epsilon}}_M(t) = \boldsymbol{M}(t) - \boldsymbol{X}_M\hat{\boldsymbol{S}}(t)$ and $\hat{\boldsymbol{\epsilon}}_E(t) = \boldsymbol{E}(t) - \boldsymbol{X}_E\hat{\boldsymbol{S}}(t)$ for each time point $t = 1, \ldots, T$. We will make here the assumption that they should be draws from a mean-zero Gaussian distribution if the assumed model generated the observed data. Figure 2.6, panels (a) and (b) show the time series plots of the residuals for EEG and MEG respectively. In the case of the EEG data, the model seems to have captured the signal at most

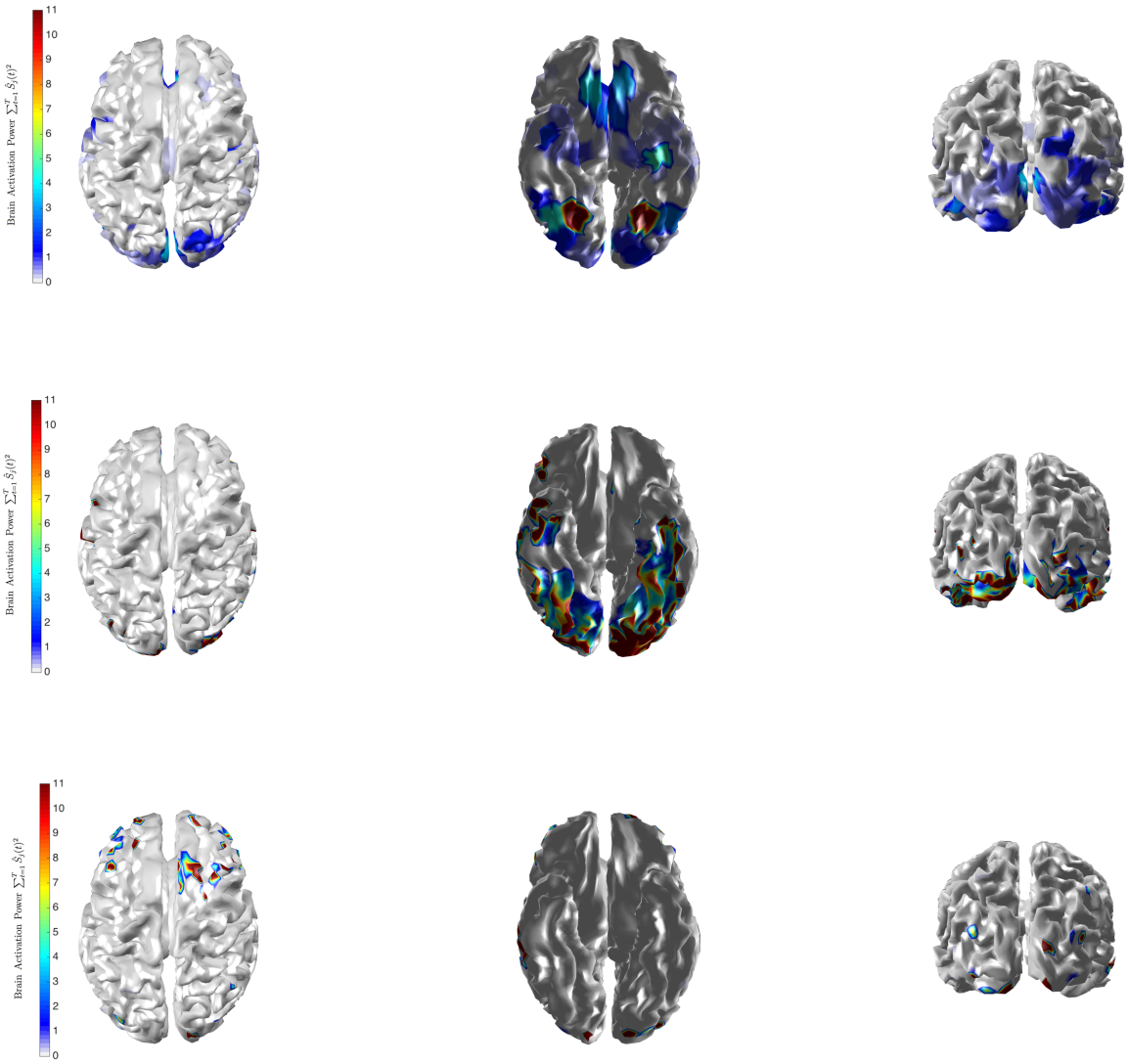


Figure 2.4: Brain Activation for Scrambled Faces - The power of the estimated source activity $\sum_{t=1}^{T} \hat{S}_j(t)^2$ at each location j of the cortical surface. From top to bottom, row 1 displays results from our proposed method applied to the combined MEG and EEG data; row 2 displays results from MSM applied to the MEG data; row 3 displays results from MSM applied to the EEG data.

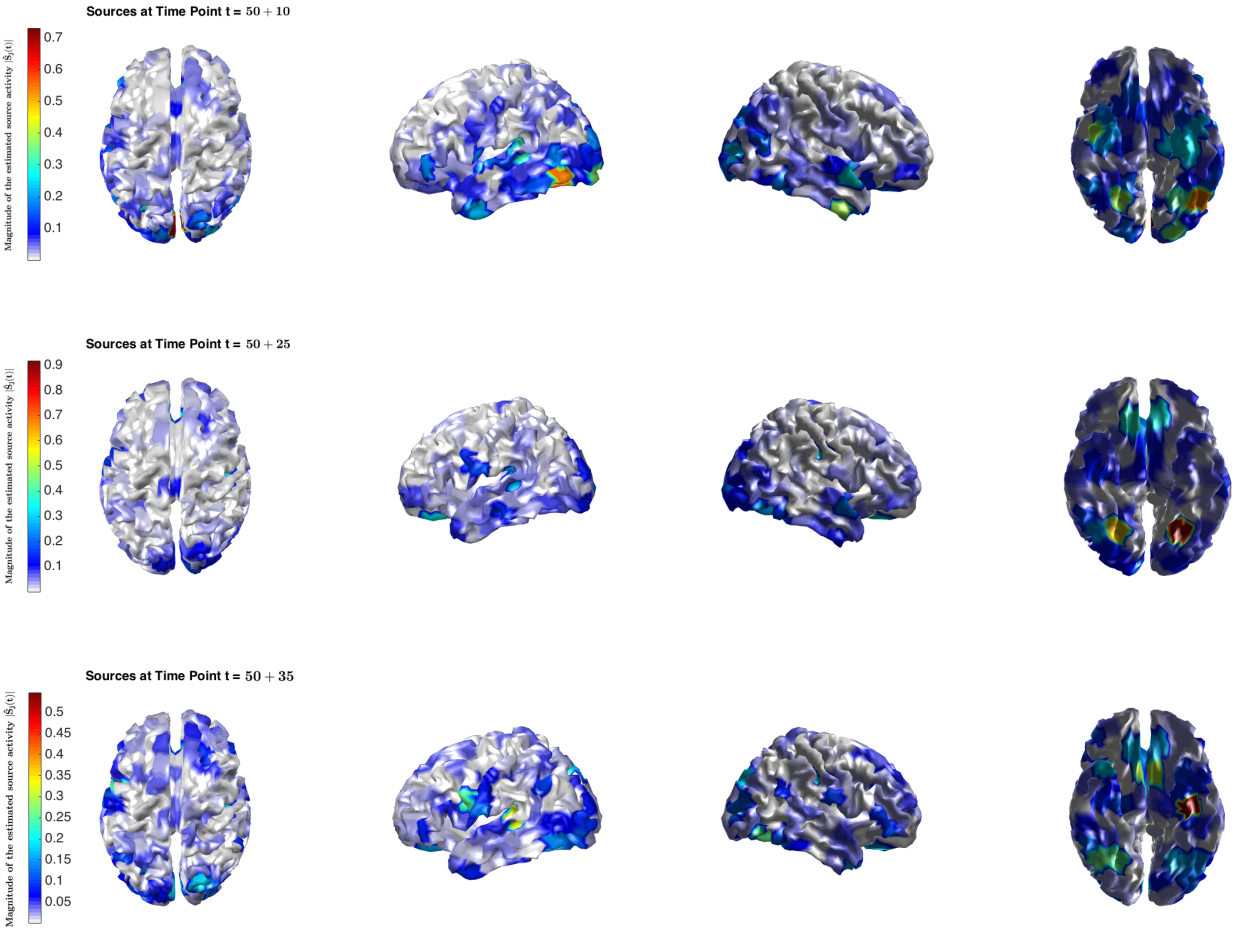


Figure 2.5: Brain Activation for Scrambled Faces - Magnitude of the estimated source activity $|\hat{S}_j(t)|$ at each location j of the cortical surface and at three different time points, $t = 50 + 10$ (Row 1; 100ms after presentation of the stimulus), $t = 50 + 25$ (Row 2; 175ms after presentation of the stimulus), and $t = 50 + 35$ (Row 3; 225ms after presentation of the stimulus).

of the sensors, though there are a few sensors where it appears that some part of the evoked signal has not been captured and remains in the residuals. The same holds true for the MEG data, and in addition, the residuals for the MEG data exhibit a periodic signal. This periodic signal is not part of the evoked response but is rather a property of the brain noise and so we are not overly concerned to see this pattern in the residuals. More concerning are the few sensors where large peaks still remain, indicate parts of the evoked response that have not been captured adequately by the model.

Figure 2.6, (c) and (d) show plots of the residuals versus fitted values for EEG and MEG respectively. For the EEG data there are no striking patterns, while for the MEG data we see higher values to the left of zero and lower values to the right of zero. These high and low values likely arise from the peaks of the periodic signal not captured by the fitted mean.

Finally, Figure 2.6, (e) and (f) show normal quantile-quantile plots for the EEG and MEG residuals respectively. The Gaussian assumption seems somewhat tenable for the EEG data while this assumption does not seem reasonable for the MEG data. In particular, we see a strong deviation from normality in the left tail of the distribution.

Overall, the residual analysis indicates a number of modelling assumptions that might be questionable for the data at hand. This is particularly true for the MEG data while the model yields a relatively better fit for the EEG data. Examining robustness to model misspecification and developing a more flexible model that can accommodate some of the features of the data not adequately captured by the current model (e.g., non-Gaussian distribution for the MEG data; temporally correlated residuals) is an important avenue for future work.

2.6 Discussion

Motivated by a study examining the neural response to scrambled faces, we have developed a new approach for solving the inverse problem associated with combined EEG and MEG data. We view our methodology as primarily applicable both to situations where the MEG and EEG data are collected simultaneously and also to situations where the data are collected sequentially in a situation where the data mimic a simultaneous recording paradigm. Our new model incorporates two simple ideas. First, it can be beneficial to combine complementary sources of information

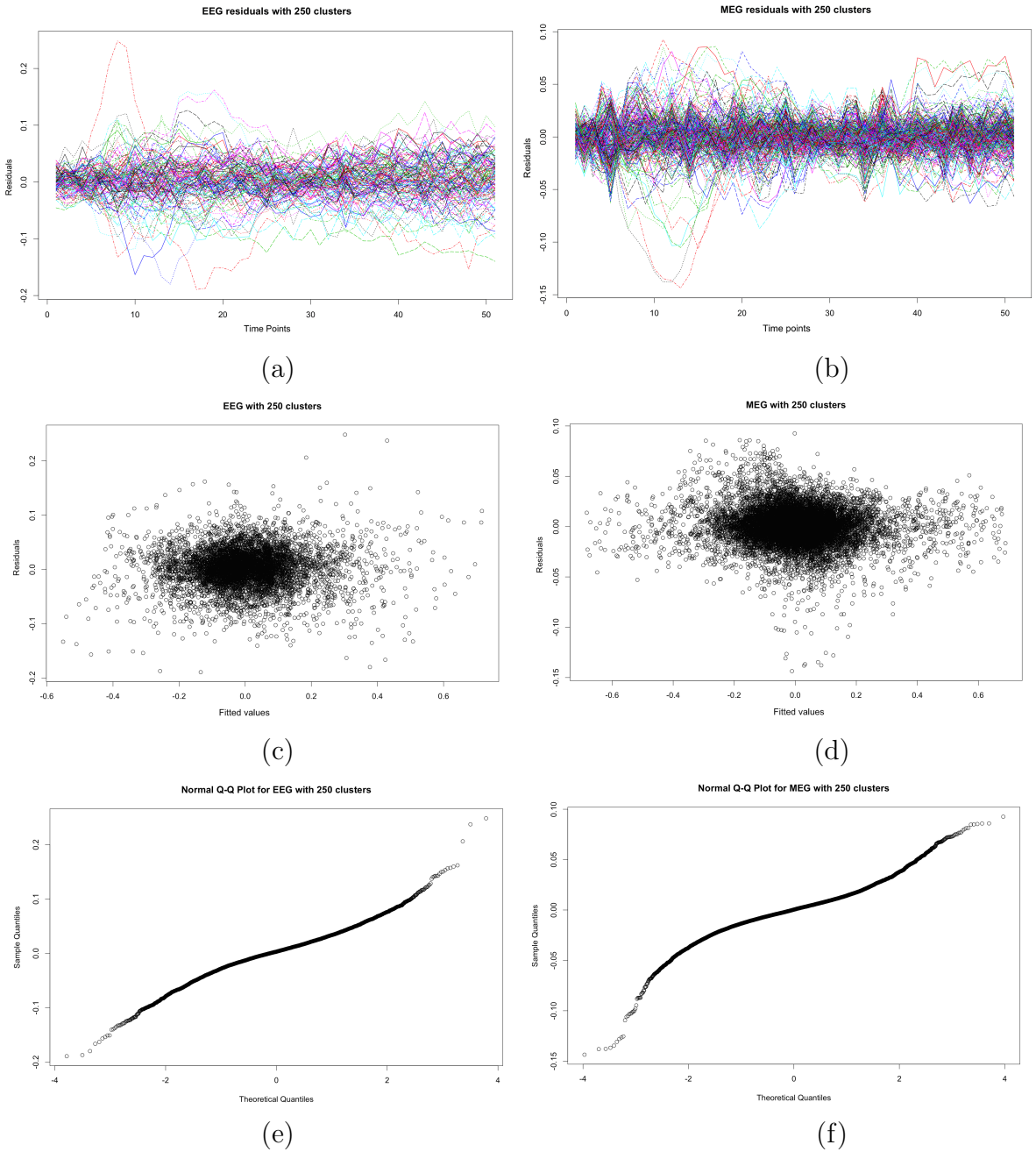


Figure 2.6: Brain Activation for Scrambled Faces - Residual Diagnostics: Time series of residuals, (a) EEG, (b) MEG; Residuals versus fitted values, (c) EEG, (d) MEG; Residual normal quantile-quantile plots, (e) EEG, (f) MEG.

when estimating unknown parameters, in particular when in the high-dimensional setting. We thus develop a joint model that links together MEG, EEG, and MRI data. Second, the hidden states of the brain are spatially correlated and we incorporate this prior information into a model that estimates the states of the brain, the number of such states, and their dynamics. Combining these two ideas together we have developed an approach for source localization that appears to result in some improvements over and above the original MSM mixture model, for the settings considered. It is worth noting that it is not only the models but also the model fitting algorithms that differ between the two approaches. Our algorithm employs ICM whilst Daunizeau and Friston (2007) employ the mean-field variational Bayes algorithm.

Our methodology makes the very strong assumption that neural activity is generated by a small number of latent states. While many neuroscientists might hesitate to claim such low dimensions, there are certainly many situations where the bulk of neural activity may turn out to be closely approximated by a low dimensional process with linear interaction. On a coarser time scale, most fMRI studies of tasks find dozens of distinct small regions activated, although they may not be statistically independent. Extending our approach to accommodate a large number of sources is an open problem.

For $\hat{K}_{ICM}$, our empirical investigations suggest that our estimator for the number of mixture components $\hat{K}_{ICM}$ exhibits a reasonable performance and is relatively stable with respect to initial values. Our simulation results indicate that $\hat{K}_{ICM}$ can be biased, under-estimating the number of mixture components when the separation between components is reduced, and in particular when the number of latent states increases. When the mixture components are sufficiently well separated the estimator seems to perform well as indicated in Figure A.6. Further investigations, both empirical and theoretical, and also for other latent mixture models, where estimation of the number of mixture components is based on the automatic pruning achieved by the ICM or an alternative optimization algorithm will be of interest. For complex mixture models with many parameters the appeal of $\hat{K}_{ICM}$ is its simplicity and the efficiency associated with its computation. Establishing conditions for convergence $\hat{K}_{ICM} \xrightarrow{D} K_{true}$ would be useful in providing a theoretical justification for this estimator, both within our context and also for application to latent mixture models in general. Aside from the inverse problem considered in this paper, the idea of simultaneous estimation and model selection for the latent Gaussian mixture model developed here has considerable potential for further development.

Chapter 3

A Spatial Model for Imaging Genetics

3.1 Introduction

We consider multivariate multiple regression modeling within the context of imaging genetics where interest lies in uncovering the associations between genetic variations and neuroimaging measures as quantitative traits (QTs). This problem has received a great deal of attention recently and is challenging because it combines the analysis of neuroimaging data with genetic data (see e.g., Vounou et al., 2010; Stein et al., 2010; Silver et al., 2011; Inkster et al., 2010; Hibar et al., 2011; Ge et al., 2012; Thompson et al., 2013; Stingo et al., 2013; Zhu et al., 2014; Hibar et al., 2015; Huang et al., 2015; Huang et al., 2017; Lu et al., 2017). Recent reviews of statistical issues in this area are discussed in Liu and Calhoun (2014) and Nathoo et al. (2018).

The neuroimaging measures can serve as endophenotypes (John and Lewis, 1966), which could be used to separate behavioural symptoms into more stable phenotypes with a clear genetic connection, for neurological disorders such as Alzheimer's disease (AD). AD has been considered widely as an application in imaging genetics with many recent studies focussing on the ADNI database. As described in Szefer et al. (2017), the estimated heritability of late-onset AD is 60 - 80 percent (Gatz et al., 2006). The largest susceptibility allele is the $\epsilon 4$ allele of APOE (Corder et al. 1993), which may play a role in 20 to 25 percent of AD cases. The remaining heritability of AD may be explained by many additional genetic variants and these may have a small effect.

Data analysis within this setting can range from studies considering a specific candidate region of interest (ROI) within the brain and a specific candidate genetic marker in the simplest case, to massive brain-wide genome-wide analyses in the most challenging case. In our work, we consider the intermediary setting where interest lies in assessing the association between a moderate number of brain imaging phentoypes (e.g., 111 ROIs in Vounou et al., 2010; 12 ROIs in Wang et al., 2012; 93 ROIs in Zhu et al., 2014; 56 ROIs in Greenlaw et al., 2017) and with the number of SNPs ranging from between a few hundred to a few thousand. Within this setting a multivariate model with a regression matrix jointly characterizing the associations between all ROIs and genetic markers is feasible; although, as detailed in the aforementioned references, we still face a challenging multivariate potentially high-dimensional regression problem.

Greenlaw et al.(2017) recently proposed a Bayesian group sparse multi-task regression model where the primary focus was the use of a new shrinkage prior based on a product of multivariate Laplace kernels developed following the ideas of Park and Casella (2008) and Kyung et al. (2010). The specific prior developed in Greenlaw et al.(2017) was motivated by the penalized multi-task regression estimator proposed by Wang et al. (2012). This development was an effort to move from point estimation to Bayesian credible intervals in a generalization where the mode of the posterior distribution in the model of Greenlaw et al. (2017) is exactly the estimator proposed by Wang et al. (2012).

While Greenlaw et al. (2017) demonstrated the advantage of uncertainty quantification in their imaging genetics application to the ADNI study, their model makes a simplifying assumption for the covariance matrix of c imaging phenotypes, where the first level of their model assumes:

$$\mathbf{y}_\ell \,|\mathbf{W}, \sigma^2 \overset{ind}{\sim} MVN_c(\mathbf{W}^{\mathrm{T}}\mathbf{x}_\ell \,,\, \sigma^2 I_c) \quad \ell = 1, \ldots, n \tag{3.1}$$

where $\mathbf{y}_\ell = (\mathbf{y}_{\ell 1}, \ldots, \mathbf{y}_{\ell c})^T$ denotes the vector of imaging phenotypes for subject ℓ, where $\ell = 1, \ldots, n$; $\mathbf{W}$ is the regression matrix; $\mathbf{x}_\ell = (\mathbf{x}_{\ell 1}, \ldots, \mathbf{x}_{\ell d})^T$, where $\mathbf{x}_\ell$ denotes the vector of genetic markers for subject ℓ. The assumed covariance structure $\sigma^2 I_c$ ignores spatial correlation as well as bilateral correlation across brain hemispheres. By the latter we mean correlation in similar structures on opposite hemispheres of the brain (e.g., a priori we expect the volume of the right hippocampus to be correlated with the volume of the left hippocampus).

We develop a new model that allows for this type of correlation by adopting a

proper bivariate conditional autoregressive process (BCAR; see, e.g., Gelfand and Vounatsou, 2003; Jin et al., 2005) for the errors in the regression model. While spatial models for functional magnetic resonance imaging (fMRI) and other neuroimaging modalities have been developed to a large extent (see, e.g., Penny et al., 2005; Bowman, 2005; Bowman et al., 2008; Derado et al., 2013; Teng et al., 2018a; Teng et al., 2018b), to our knowledge there has been very little development of explicitly spatial models for imaging genetics. One exception is the mixture model developed by Stingo et al. (2013) where an Ising prior, a binary Markov random field, is used for Bayesian variable selection. Our model is rather different in both its aims and structure as it is based on a continuous bivariate Markov random field that is specified at the first level of the model for the imaging data directly.

Typically, models incorporating multivariate CAR specifications are used for modelling observations (in the case of a proper CAR model) or spatially-varying parameters when multiple observations or parameters appear at each spatial site. For our application the use of this process is non-standard in the sense that we do not model multiple observations at each site, but rather, we pair corresponding observations on opposite hemispheres of the brain and use the bivariate spatial process to model a combination of the bilateral correlation across the left and right brain hemispheres as well as the spatial correlation within each hemisphere. As a matter of fact for the MRI data considered in our application the bilateral correlation is the stronger signal in the observed data and so it is important to account for it.

For the bivariate spatial model, we use a separable BCAR process as it is reasonable in our application to assume (as it might be in many neuroimaging studies) that the spatial structure on the two hemispheres of the brain is similar. Non-separable multivariate spatial models (see, e.g., Gelfand and Banerjee, 2010; MacNab 2016) could be adopted for more flexibility allowing the spatial structure on the two hemispheres to be different; however, we do not expect that this additional flexibility would be useful in the current context. This spatial process is combined with a group Lasso prior for the regression coefficients, where each group corresponds to a single row of $\mathbf{W}$. Each row in this case represents the associations between a given SNP and the phenotypes across all ROIs. We employ a bivariate Gaussian scale mixture representation of a group Lasso prior in order to facilitate Bayesian computation.

To compute the posterior distribution we develop two algorithms, both of which are implemented in our R package *bgsmtr* for imaging genetics regression modelling. The first is a Gibbs sampling algorithm and the second is a faster mean-field vari-

ational Bayes (VB) approximation to the posterior distribution (see e.g., Ormerod and Wand, 2010; Nathoo et al., 2013; Teng et al., 2018a; Teng et al., 2018b). Within the context of hierarchical models for spatial data, mean-field VB inference has been considered by Ren et al. (2011). In addition to the computation of the posterior distribution, the *bgsmtr* package now incorporates Bayesian FDR procedures (Morris et al., 2008) for SNP selection. This can be used alongside or as an alternative to SNP selection based on credible intervals.

The overall contribution of our work is four-fold. First, we develop an explicitly spatial model for imaging genetics based on the BCAR process. Second, we develop both an MCMC algorithm and a mean-field VB algorithm for approximating the posterior distribution. Third, we incorporate Bayesian FDR procedures for SNP selection within the new spatial model. Fourth, our new developments are implemented in the latest version of the *bgsmtr* R package that is available for download on CRAN.

3.2 Bayesian Spatial Regression Model

Let $\mathbf{y}_\ell = (\mathbf{y}_{\ell 1}, \ldots, \mathbf{y}_{\ell c})^T$ and $\mathbf{x}_\ell = (\mathbf{x}_{\ell 1}, \ldots, \mathbf{x}_{\ell d})^T$ denote the imaging measures at c ROIs and the genetic data respectively for subject ℓ, $\ell = 1, \ldots, n$, where $\mathbf{x}_{\ell j} \in \{0, 1, 2\}$ represents the number of minor alleles of the j^{th} SNP for subject ℓ, where $j = 1, \ldots, d$. The regression model takes the form $E(\mathbf{y}_\ell) = \boldsymbol{W}^T \mathbf{x}_\ell,\ \ell = 1, \ldots, n$, where $\boldsymbol{W}$ has dimensions $d \times c$ and W_{ij} represents the association between the i^{th} SNP, where $i = 1, \ldots, d$ and the j^{th} ROI imaging phenotype, where $j = 1, \ldots, c$

Our model is developed for settings where the imaging data are symmetric with the same measures collected on each hemisphere of the brain. This is true when the neuroimaging data are considered at the voxel level and it is also the case for the study considered here where we analyze MRI data from the ADNI-1 database preprocessed using the FreeSurfer V4 software.

As described in Szefer et al. (2017), potential confounders in the analysis are population stratification and APOE genotype. Since true population structure is not observed, a set of principal coordinates from multidimensional scaling are used to derive proxy variables for population stratification in the data. We also adjust for APOE genotype, since it can account for the population stratification in the data, over and above the principal components or principal coordinates (Lucotte et al. 1997).

The response imaging measures at each brain ROI are first adjusted for the ten

principal coordinates, as well as for dummy variables representing APOE genotype, using weighted ordinary least squares regression. The residuals from each regression are then used as the adjusted neuroimaging phenotypes (Szefer et al., 2017).

Let $\boldsymbol{y}_{\ell i} = (y_{li}^{(L)}, y_{li}^{(R)})'$ be the brain summary measures obtained at the i^{th} ROI in the left hemisphere (L) and the right hemisphere (R). Then $\boldsymbol{y}'_{\ell} = (\boldsymbol{y}'_{\ell 1}, \ldots, \boldsymbol{y}'_{\ell c/2})'$ is the imaging data ordered so that left-right imaging phenotype pairs are adjacent in the response vector. There are thus $c/2$ ROIs on each hemisphere and we let $\boldsymbol{A}$ denote a $c/2 \times c/2$ symmetric neighbourhood matrix which in the simplest case can have binary elements, where $A_{ij} = 1$ indicates that ROI i and j are neighbours $i \neq j$, or more generally $A_{ij} \geq 0$ and $A_{ii} = 0$, $i = 1, \ldots, c/2$. The *bgsmtr* R package allows the user to specify the neighbourhood matrix $\boldsymbol{A}$, or in the absence of user input takes A_{ij} to be the average of the absolute value of the sample correlation between ROI i and ROI j, where the average is taken over left/right hemispheres. The regression model then takes the form

$$\boldsymbol{y}_{\ell} = \boldsymbol{W}^T \boldsymbol{x}_{\ell} + \boldsymbol{\epsilon}_{\ell} \tag{3.2}$$

and the model for the errors $\boldsymbol{\epsilon}_{\ell}$ is a mean-zero multivariate normal distribution of dimension c, which can be specified through a set of $c/2$ compatible bivariate conditional distributions for $\boldsymbol{\epsilon}_{li} = (\epsilon_{li}^{(L)}, \epsilon_{li}^{(R)})'$, specified as follows:

$$\boldsymbol{\epsilon}_{li} | \boldsymbol{\epsilon}_{l\{-i\}}, \rho, \boldsymbol{\Sigma} \sim \text{BVN}\big(\frac{\rho}{A_{i.}} \sum_{j=1}^{c/2} A_{ij} \boldsymbol{\epsilon}_{\boldsymbol{lj}}, \frac{1}{A_{i.}} \boldsymbol{\Sigma}\big)$$

where $\rho \in [0, 1)$ characterizes spatial dependence with $\rho = 0$ corresponding to independence across all ROI pairs and $\boldsymbol{\Sigma}$ is a 2×2 matrix where $\kappa = \Sigma_{12}/\sqrt{\Sigma_{11}\Sigma_{22}} \in (-1, 1)$ quantifies within pair dependence, and with $\kappa = 0$ corresponding to independence within ROI pairs. As far as we are aware, this spatial model for neuroimaging data is one of the first to explicitly model dependence across brain hemispheres in addition to accounting for local spatial dependence.

Under this new specification the first level of the regression model takes the following form:

$$\boldsymbol{y}_{\ell} | \boldsymbol{W}, \boldsymbol{\Sigma} \overset{ind}{\sim} \text{MVN}_c(\boldsymbol{W}^T \boldsymbol{x}_{\ell}, (\boldsymbol{D}_{\boldsymbol{A}} - \rho \boldsymbol{A})^{-1} \otimes \boldsymbol{\Sigma}) \tag{3.3}$$

where $\boldsymbol{D}_{\boldsymbol{A}} = \text{diag}\{A_{i.}, i = 1, ..., c/2\}$ and $A_{i.} = \sum_{j=1}^{\frac{c}{2}} A_{ij}$. The covariance structure now is the kronecker product of two parts. The first term $(\boldsymbol{D}_{\boldsymbol{A}} - \rho \boldsymbol{A})^{-1}$ with dimension $c/2 \times c/2$ incorporates the spatial correlation between neighbouring brain regions for

imaging phenotypes while second term $\boldsymbol{\Sigma}$ with dimension 2×2 characterizes the bilateral correlation across brain hemispheres.

For the regression coefficients, we let $\tilde{W}_{ij^*} = (W_{ij}, W_{ij+1})'$, $j = 2j^* - 1$, $j^* = 1, ..., \frac{c}{2}$, and we adopt a shrinkage prior based on a bivariate scale mixture

$$\tilde{W}_{ij^*}|\omega_i^2, \Sigma \overset{ind}{\sim} \text{BVN}(\mathbf{0}, \omega_i^2\Sigma),$$

$$\omega_i^2|\lambda^2 \overset{iid}{\sim} \text{Gamma}(\frac{c+1}{2}, \lambda^2/2),\ \Sigma \sim \text{Inv-Wishart}(v, \boldsymbol{S}),$$

where ρ and λ^2 are tuning parameters controlling spatial dependence and regression sparsity respectively. These can be varied across a coarse grid and selected using information criteria. In our application we select these parameters using the Watanabe?Akaike information criterion (WAIC) as recommended in similar contexts by Greenlaw et al. (2017) and Nathoo et al. (2016). Alternatively, ρ can be fixed at a default value of $\rho = 0.95$ corresponding to a relatively high level of spatial correlation when this is a reasonable assumption, and λ^2 can be varied over a range of values with the number of active SNPs recorded for each such value. The results can then be summarized based on a desired or expected level of sparsity. The remaining hyperparameters v and $\boldsymbol{S}$ are set at $v = 2$ and $\boldsymbol{S} = \boldsymbol{I}$ to yield a prior that is somewhat vague, and they can be varied as part of a sensitivity analysis.

3.3 Computation and SNP Selection

3.3.1 Bayesian Computation

Bayesian inference for our proposed model is based on the posterior distribution $P(\boldsymbol{\Theta}|\mathbf{Y})$, where $\boldsymbol{\Theta} = \{\boldsymbol{W}, \boldsymbol{\Sigma}, \boldsymbol{\omega^2}\}$ and $\mathbf{Y}$ denotes the imaging data for all n subjects. Posterior computation can be implemented using Gibbs sampling. The update steps for this algorithm are listed in Algorithm 2 and their derivations are given in the Appendix B.

Algorithm 2 Gibbs Sampling Algorithm

1. Set tuning parameters λ^2 and ρ.

2. Initialize $\boldsymbol{W}, \boldsymbol{\Sigma}, \boldsymbol{\omega^2}$ and repeat steps (3) - (6) below to obtain the desired Monte Carlo sample size after burn-in.

3. For $k = 1, ..., d$, update $\boldsymbol{W}^{(k)T}$ as:

$$\boldsymbol{W}^{(k)T} \sim \text{MVN}_{m_k c}(\underset{\sim}{\mu}_{\text{k}}, \boldsymbol{\Sigma}_k),$$

4. Where:

$$\underset{\sim}{\mu}_{\text{k}} = \boldsymbol{\Sigma}_k \left(-\sum_{\ell=1}^{\text{n}} (\boldsymbol{x}_\ell^{(k)} \otimes I_c)[(D_A - \rho A) \otimes \boldsymbol{\Sigma}_{(t-1)}^{-1}] \times (\boldsymbol{x}_\ell^{(-k)T} \otimes I_c)(\boldsymbol{W}_{(t-1)}^{(-k)T}) + \sum_{\ell=1}^{\text{n}} (\boldsymbol{x}_\ell^{(k)} \otimes I_c)[(D_A - \rho A) \otimes \boldsymbol{\Sigma}_{(t-1)}^{-1}] \boldsymbol{y}_\ell \right),$$

$$\boldsymbol{\Sigma}_k = \left(\boldsymbol{H}_k + \sum_{\ell=1}^{\text{n}} (\boldsymbol{x}_\ell^{(k)} \otimes I_c)[(D_A - \rho A) \otimes \boldsymbol{\Sigma}_{(t-1)}^{-1}] \ (\boldsymbol{x}_\ell^{(k)T} \otimes I_c) \right)^{-1},$$

$$\boldsymbol{H}_k = \left[\left\{ \frac{1}{\omega_{k(t-1)}^2} \right\} \otimes I_{\frac{c}{2}} \otimes \Sigma^{-1} \right].$$

5. Update $\boldsymbol{\Sigma}$ as:

$$\boldsymbol{\Sigma} \sim \text{Inverse-Wishart}(S^*, v^*)$$

where:

$$S^* = \sum_{l=1}^{n} \sum_{i=1}^{\frac{c}{2}} \sum_{j=1}^{\frac{c}{2}} b_{ij} \tilde{y_{li}^*} \tilde{y_{li}^*}^T + \sum_{i=1}^{d} \sum_{j^*=1}^{\frac{c}{2}} \frac{\tilde{W}_{ij^*} \tilde{W}_{ij^*}^T}{\omega_i^2} + S,$$

$$v^* = 2n + \frac{cd}{2} + v, \ \ y_l^* = y_l - \boldsymbol{W}^T x_l,$$

$$\tilde{y^*}_{lj^*}^T = (y_{lj}^*, \ y_{lj+1}^*), \ \ \tilde{W}_{ij^*} = (W_{ij}, \ W_{ij+1}).$$

6. For $i = 1, \ldots, d$ update ω_i^2, through

$$1/\omega_i^2 \sim \text{Inverse-Gaussian} \left(\sqrt{\frac{\lambda^2}{c_i^*}}, \ \ \lambda^2 \right)$$

where:

$$c_i^* = tr(\sum_{j^*=1}^{\frac{c}{2}} \tilde{w_{ij^*}} \tilde{w_{ij^*}}^T \boldsymbol{\Sigma}_{(t)}{}^{-1})$$

As a faster albeit more approximate approach to computing the posterior distribution, we also develop a mean-field VB algorithm. As opposed to Monte Carlo sampling, variational inference is based on solving an optimization problem. The approximation $q(\boldsymbol{\theta})$ to the posterior distribution $P(\boldsymbol{\theta}|\mathbf{Y})$ is based on constructing and optimizing a lower bound on the marginal likelihood $P(\mathbf{Y})$.

Assuming that $q(\boldsymbol{\theta})$ has the same support as $P(\boldsymbol{\theta}|\mathbf{Y})$, the log-marginal likelihood can be written as

$$\begin{aligned}\log P(\mathbf{Y}) &= \int q(\boldsymbol{\theta}) \log\{\frac{P(\mathbf{Y},\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\} d\boldsymbol{\theta} + \int q(\boldsymbol{\theta}) \log\{\frac{q(\boldsymbol{\theta})}{P(\boldsymbol{\theta}|\mathbf{Y})}\} d\boldsymbol{\theta} \\ &= E_q[\log\{\frac{P(\boldsymbol{\theta},\mathbf{Y})}{q(\boldsymbol{\theta})}\}] + E_q[\log\{\frac{q(\boldsymbol{\theta})}{P(\boldsymbol{\theta}|\mathbf{Y})}\}] \\ &= \mathfrak{F}(q,\mathbf{Y}) + KL(q||p) \geq \mathfrak{F}(q,\mathbf{Y}).\end{aligned}$$

Here, $KL(q||p)$ denotes the Kullback-Leibler divergence from $q(\boldsymbol{\theta})$ to $P(\boldsymbol{\theta}|\mathbf{Y})$ and the final inequality is true since $KL(q||p) \geq 0$. The approximation to $P(\boldsymbol{\theta}|\mathbf{Y})$ by $q(\boldsymbol{\theta})$ is obtained by restricting $q(\boldsymbol{\theta})$ to a manageable class of distributions and maximizing the lower bound $\mathfrak{F}(q,\mathbf{Y})$ (which is equivalent to minimizing $KL(q||p)$) over that class. In the case of mean-field VB, the restriction of $q(\boldsymbol{\theta})$ is to a product form $q(\boldsymbol{\theta}) = \prod_{j=1}^{J} q_j(\boldsymbol{\theta}_j)$. In the specific context of our model the assumed product form is a follows

$$P(\boldsymbol{\Theta}|Y) \approx \Big[\prod_{k=1}^{d} q(\boldsymbol{W}^{(k)})\Big]\Big[\prod_{i=1}^{d} q(\omega_i^2)\Big] q(\Sigma) \tag{3.4}$$

where $\boldsymbol{W}^{(k)}$ is the k^{th} row $\boldsymbol{W}$.

We maximize the functional $\mathfrak{F}(q_1,\ldots,q_J,\mathbf{Y})$ over the q_j's using a coordinate ascent procedure. The update steps for this procedure take the form (see, e.g., Ormerod and Wand, 2010)

$$q_i(\boldsymbol{\theta}_i) = \frac{\exp\{E_{\theta_{-i}}[\log P(\boldsymbol{\theta}_i|\mathbf{Y},\boldsymbol{\theta}_{-i})]\}}{\int \exp\{E_{\theta_{-i}}[\log P(\boldsymbol{\theta}_i|\mathbf{Y},\boldsymbol{\theta}_{-i})]\} d\boldsymbol{\theta}_i}$$

where the expectation is taken with respect to $q_{-i}(\boldsymbol{\theta}_{-i}) = \prod_{l\neq i} q_l(\boldsymbol{\theta}_l)$. This leads to a set of update equations related to the EM algorithm (Beal, 2003) that are iterated until convergence to a local optimum. These update equations are presented in Algorithm 3 and their derivations are detailed in the Appendix B. On convergence the

approximation to the posterior distribution is based on (3.4) as well as the solutions

$$\begin{aligned} q(\boldsymbol{W}^{(k)}) &\equiv \text{MVN}(\boldsymbol{\mu}_{q_{\boldsymbol{W}^{(k)}}}, \boldsymbol{\Sigma}_{q_{\boldsymbol{W}^{(k)}}}) \\ q(\omega_i^2) &\equiv \text{Reciprocal Inverse Gaussian}(\mu_{q(\eta_i)}, \lambda_{q(\eta_i)}) \\ q(\Sigma) &\equiv \text{Inverse-Wishart}(S_{q(\Sigma)}, v_{q(\Sigma)}) \end{aligned}$$

where the statistics $\{\boldsymbol{\mu}_{q_{\boldsymbol{W}^{(k)}}}, \boldsymbol{\Sigma}_{q_{\boldsymbol{W}^{(k)}}}, k = 1, \ldots, d\}$; $\{\mu_{q(\eta_i)}, \lambda_{q(\eta_i)}, i = 1, \ldots, d\}$; $S_{q(\Sigma)}$, $v_{q(\Sigma)}$ are obtained as the output of the iterative Algorithm 3.

3.3.2 Bayesian False Discovery Rate

The Bayesian false discovery rate (FDR) procedure applied in our work for SNP selection follows the approach developed in Morris et al. (2008), but it has been adapted and implemented for the current spatial model.

We assume that we have N samples $W_{ij}^{(1)}, \ldots, W_{ij}^{(N)}$ from the posterior distribution (obtained through Gibbs sampling or through simply simulating from the variational approximation) for each of the regression coefficients W_{ij}, $i = 1, \cdots, d$, $j = 1, \cdots, c$. Let c^* be a known critical value. Given this value, we compute a posterior tail probability for the i-th SNP at region j as $p_{ij} = Pr(|W_{ij}| > c^*|\mathbf{Y}), i = 1, \ldots, d; j = 1, \ldots, c$, which can be approximated by $p_{ij} \approx N^{-1} \sum_{i^*=1}^{N} I\left\{|W_{ij}^{(i^*)}| > c^*\right\}$ and we replace any $p_{ij} = 1$ with $1 - (2N)^{-1}$(Morris et al., 2008). Given these posterior tail probabilities and a desired global FDR-bound α, we denote by ϕ_α the corresponding threshold chosen so that a SNP-region pair (i, j) is selected if $p_{ij} > \phi_\alpha$. The cutoff ϕ_α can be computed by sorting $\{p_{ij}, i = 1, \cdots, d; j = 1, \ldots, c\}$ in descending order $\{p(i), i = 1, \cdots, d \times c\}$, then $\phi_\alpha = p(\lambda)$, with $\lambda = \max\left\{l^* : (l^*)^{-1} \sum_{l=1}^{l^*} (1 - p(l)) \leq \alpha\right\}$ where l^* is the index. The threshold ϕ_α is a cutpoint on the posterior probabilities that controls the expected Bayesian FDR below level α.

The value of c^* can be chosen based on prior knowledge of what constitutes an effect size of interest or, in the absence of such knowledge, it can be chosen empirically based on the data. For example, posterior quantities such as the average or minimum posterior standard deviation taken over all regression coefficients are possible choices. The latter is the default choice in *bgsmtr*.

Algorithm 3 mean-field Variational Bayes Algorithm

1. Set tuning parameters λ^2 and ρ.

2. Initialize $q_{(\boldsymbol{W})}, q_{(\boldsymbol{\Sigma})}, q_{(\boldsymbol{\omega^2})}$ and cycle through steps (3) - (5) below until the increase in the lower bound $\mathcal{L}(q)$ is negligible.

3. For $k = 1, ..., d$, update

$$\boldsymbol{\Sigma}^{-1}_{q(W^k)} \leftarrow \Bigg(\Big[diag\left\{\mu_{q(\eta_i)}\right\}_{i\in\pi_k} \otimes I_{\frac{c}{2}} \otimes (v_{q(\Sigma)} S_{q(\Sigma)}) \Big] + \sum_{\ell=1}^{\mathrm{n}} (\boldsymbol{x}_\ell^{(k)} \otimes I_c)[(D_A - \rho A)) \otimes (v_{q(\Sigma)} S_{q(\Sigma)})](\boldsymbol{x}_\ell^{(k)T} \otimes I_c) \Bigg)^{-1},$$

$$\boldsymbol{\mu}_{q(W^k)} \leftarrow \boldsymbol{\Sigma}_{q(W^k)} \Bigg(- \sum_{\ell=1}^{\mathrm{n}} (\boldsymbol{x}_\ell^{(k)} \otimes I_c)[(D_A - \rho A) \otimes E_{q(\Sigma)}(\Sigma^{-1})](\boldsymbol{x}_\ell^{(-k)T} \otimes I_c)\big(\boldsymbol{\mu}_{q(W^{-k})}\big) + \sum_{\ell=1}^{\mathrm{n}} (\boldsymbol{x}_\ell^{(k)} \otimes I_c)[(D_A - \rho A) \otimes E_{q(\Sigma)}(\Sigma^{-1})]\boldsymbol{y}_\ell \Bigg)$$

4. Update $S_{q(\Sigma)}$ as

$$S_{q(\Sigma)} \leftarrow \sum_{l=1}^{n} \sum_{i=1}^{\frac{c}{2}} \sum_{j=1}^{\frac{c}{2}} E_q(b_{ij}) \tilde{y^*_{li}} \tilde{y^*_{li}}^T + \sum_{i=1}^{d} \sum_{j^*=1}^{\frac{c}{2}} E_q\big(\tilde{W}_{ij^*} \tilde{W}^T_{ij^*}\big) \mu_{q(\eta_i)} + S$$

where:

$$E_q(b_{ij}) = [D_A - \rho A]_{ij}$$

5. for $i = 1, ..., d$, update $\mu_{q(\eta_i)}$

$$\mu_{q(\eta_i)} \leftarrow \sqrt{\frac{\lambda^2}{E_q(c_i^*)}}$$

where:

$$E_q(c_i^*) = E_q \left(tr(\sum_{j^*=1}^{\frac{c}{2}} \tilde{W_{ij^*}} \tilde{W_{ij^*}}^T \Sigma^{-1}) \right)$$

Update:

$$\lambda_{q(\eta_i)} \leftarrow \lambda^2$$

$$\mu_{q(\omega_i^2)} \leftarrow \frac{1}{\mu_{q(\eta_i)}} + \frac{1}{\lambda_{q(\eta_i)}}$$

$$Var_{q(\omega_i^2)} \leftarrow \frac{1}{\mu_{q(\eta_i)} \lambda_{q(\eta_i)}} + \frac{2}{\lambda^2_{q(\eta_i)}}$$

3.4 ADNI-1 Study of MRI and Genetics

We applied our spatial model as well as the group sparse multi-task regression model of Greenlaw et al. (2017) to MRI and genetic data collected from $n = 632$ subjects from the ADNI-1 database.

The response measures were obtained by preprocessing the MRI data using the FreeSurfer V4 software which conducts automated parcellation to define volumetric and cortical thickness values from the 28 ROIs considered in Shen et al. (2010), Szefer et al. (2017), and Greenlaw et al. (2017) on each hemisphere of the brain, leading to $c = 56$ brain measures in total. These ROIs are chosen based on prior knowledge that they are related to Alzheimer's Disease and are described in detail in Table 2 of Greenlaw et al. (2017). Each of the response variables were adjusted for age, gender, education, handedness, baseline total intracranial volume (ICV), potential population stratification and APOE genotype and centered to have zero-sample-mean and unit-sample-variance. To adjust for these potential confounding variables, the phenotype and genotype data were regressed separately onto these variables in advance, and the residuals were used in subsequent analyses (Szefer, 2014).

The genetic data comprise SNPs belonging to the top 40 Alzheimer's Disease (AD) candidate genes listed on the AlzGene database (Bertram et al., 2007) as of June 10, 2010. The data presented here are queried from the genome build as of December 2014, from the ADNI-1 data. After quality control and imputation steps, the genetic data used for this study include 486 SNPs from the 33 targeted genes discussed in Szefer et al. (2017) and Greenlaw et al. (2017). The freely available software package PLINK (Purcell et.al., 2007) was used for genomic quality control. Thresholds used for SNP and subject exclusion are the same as in Wang et. al. (2012), with the exception that we require a more conservative genotyping call rate of at least 95% (Ge et al. 2012).

We fit our new spatial model to these data using both Algorithm 2 (Gibbs sampling) and Algorithm 3 (VB). In addition, we fit the model developed in Greenlaw et al. (2017) using the MCMC sampler derived therein. In all cases, MCMC sampling was run for 10,000 iterations with the initial 5,000 iterations discarded. The required computation time for the spatial model (MCMC) was 50 hours on a single core (2.66-GHz Xeon x5650) with 20GB of RAM, while the computation for the model of Greenlaw et al. (2017) was 5hrs. The VB algorithm was run to convergence and requires only 0.5hrs.

To compare the spatial and non-spatial models and to choose values for the tuning parameters, we used the WAIC. This criterion can be computed from posterior simulation output and takes the form

$$WAIC = -2\sum_{l=1}^{n} \log E_{\mathbf{W},\mathbf{\Sigma}}[p(\mathbf{y}_\ell \,|\, \mathbf{W}, \mathbf{\Sigma})|\mathbf{y}_1, \ldots, \mathbf{y}_n]$$

$$+2\sum_{l=1}^{n} VAR_{\mathbf{W},\mathbf{\Sigma}}[\log p(\mathbf{y}_\ell \,|\, \mathbf{W}, \mathbf{\Sigma})|\mathbf{y}_1, \ldots, \mathbf{y}_n]$$

where $p(\mathbf{y}_\ell \,|\, \mathbf{W}, \mathbf{\Sigma})$ is the multivariate normal density function associated with the conditional autoregressive model (3.3), and the expectation and variance are taken with respect to the posterior distribution. This quantity can be seen as an approximation to generalized leave-one-out cross-validation error and only requires a single fit of the model, with lower values being preferred.

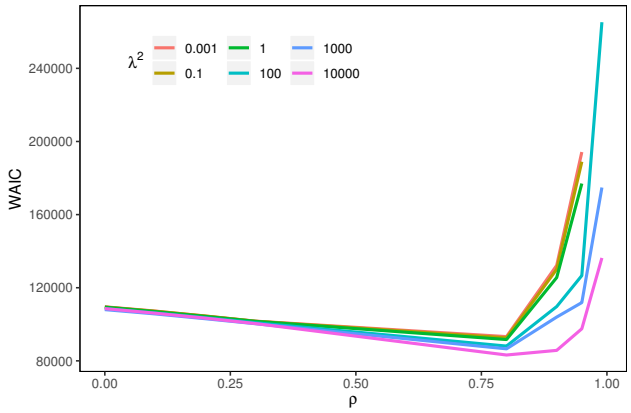


Figure 3.1: ADNI-1 Data - Relationship between WAIC and ρ for different values of λ^2.

Figure 3.1 presents the WAIC computed for a number of different choices of the tuning parameters ρ and λ^2 and suggests using values of $\rho = 0.8$ and $\lambda^2 = 10,000$ for the ADNI-1 data. In this case, the value of the WAIC is $83,170$. The WAIC obtained for the non-spatial model of Greenlaw et al. (2017) with tuning parameters selected according to their implementation is $108,745$. This relatively large difference in WAIC suggests that our proposed spatial model has superior performance in our study.

Figure 3.2 presents the number of SNPs chosen by the spatial model for each ROI using Bayesian FDR as a function of the tuning parameter λ^2 for both Gibbs sampling

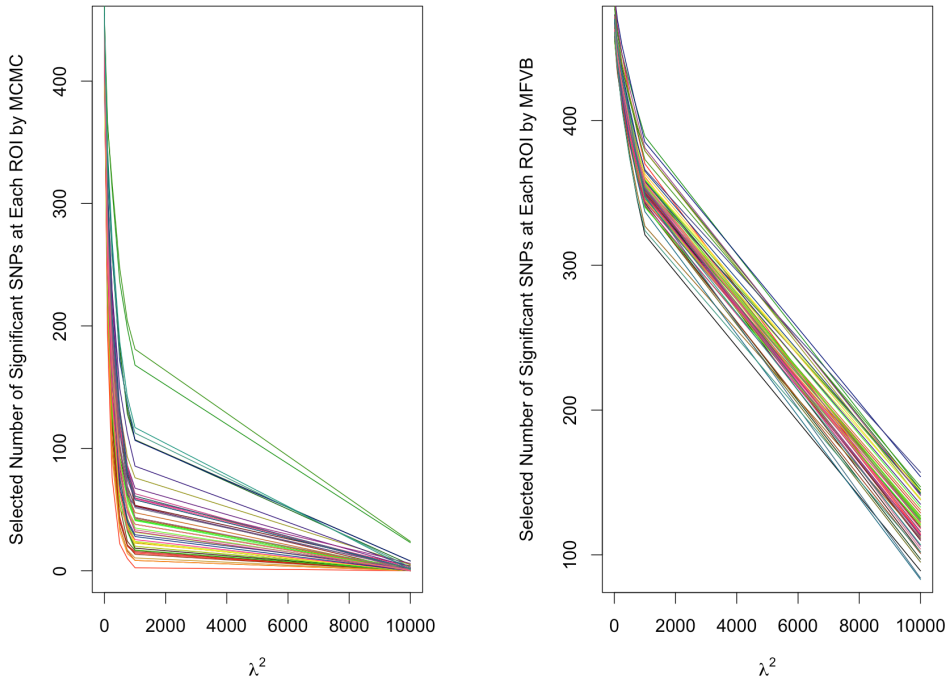


Figure 3.2: ADNI-1 Data - Relationship between the number of selected SNPs in each region and λ^2. Each region is represented with a curve in each panel of the figure. The left panel shows this relationship for MCMC combined with Bayesian FDR ($\alpha = 0.05$) while the right panel shows the same relationship for VB with Bayesian FDR ($\alpha = 0.05$).

and VB. As expected, the curves are monotone decreasing but it is interesting to note that their shapes differ when comparing the algorithms. In particular, VB selects a larger number of SNPs at all values of λ^2. This suggests that the VB algorithm is a fairly rough approximation since MCMC is a gold standard and it has associated consistency guarantees (Robert and Casella, 2004). The VB algorithm is thus best suited for obtaining starting values to initialize the MCMC, and it can also be used as a tool to gain some initial insight (based on the mean-field approximation) into the data while the MCMC sampler runs to completion.

In the MCMC algorithm, for the values of the tuning parameters selected by WAIC ($\rho = 0.8$, $\lambda^2 = 10,000$) the average number of SNPs selected per ROI was 2, while more than half of the ROIs have no SNPs selected. In total, 75 SNPs across all 56 ROIs were selected and these are listed in Table B.1 along with the corresponding phenotypes that they were associated with. With the VB approximation, 150 SNPs were selected, and the set of SNP-ROI pairs selected by MCMC is a proper subset of the set selected by VB. In addition, the subset of SNPs and phenotypes also selected

Table 3.1: ADNI-1 Study: Estimated posterior mean and 95% equal-tail credible intervals for the ROIs selected by Bayesian FDR for APOE SNP rs405509.

	Spatial Model		**Original Model**	
Region	Mean	95% CI	Mean	95% CI
Left_Postcentral	0.13	[0.03, 0.24]	0.12	[0.02, 0.22]
Right_SupFrontal	0.12	[0.02, 0.22]	0.15	[0.05, 0.25]
Left_Supramarg	0.13	[0.02, 0.23]	0.15	[0.05, 0.26]
Right_Supramarg	0.14	[0.04, 0.24]	0.13	[0.02, 0.22]

by the approach of Greenlaw et al. (2017) where the marginal posterior 95% credible interval was used for SNP selection are also highlighted in bold in Table B.1.

Considering all three approaches, the most consistent signal is found at the APOE gene, where all three methods select SNP rs405509 and find associations with *right-Midtemporal* (thickness of the middle temporal gyrus), *right-Supramarg* (thickness of the supramarginal gyrus), *right-MeanFront* (mean thickness of Fusiform, parahippocampal, and lingual gyri, temporal pole and transverse temporal pole), and finally *right-MeanLatTemp* (mean thickness of the Inferior temporal, middle temporal, and superior temporalgyri). It is pertinent to note that this SNP is the only APOE SNP contained within the set of 486 targeted SNPs included in our study. It is also interesting to note that the selected associations for this SNP all correspond to ROIs in the right brain hemisphere. The associated point estimates and 95% credible intervals for the four ROIs are given in Table 3.1, for both the spatial and original model. Both models yield very similar posterior summaries and all of these indicate a positive association between the imaging phenotypes and the number of APOE rs405509 minor alleles.

Another consistent signal was found at the ACE gene with SNP rs4311, which was found associated with 12 ROIs. We note that all but one of these ROIs is in the right hemisphere, and three of these ROIs (all of which are in the right hemisphere) are in common with the ROIs selected for this SNP by Greenlaw et al. (2017).

In Figure 3.3 we indicate the SNPs chosen for each ROI, where the SNPs are grouped on the x-axis by gene and the ROIs are grouped in left/right pairs on the y-axis. The selected SNPs for each ROI are shown for the case where the tuning parameter $\lambda^2 = 1000$ and also for the case where $\lambda^2 = 10,000$. In both cases the value of the spatial tuning parameter is set at $\rho = 0.8$ as suggested by Figure 3.1.

Figure 3.3: ADNI-1 Data: SNPs chosen with the spatial model fit using Gibbs sampling and Bayesian FDR ($\alpha = 0.05$) are highlighted in red for each phenotype. The black ticks on y-axis indicate the phenotypes from the left/right hemisphere, and the SNPs from same gene are indicated by the ticks on x-axis. The top panel corresponds to the case $\lambda^2 = 1000$ while the bottom panel corresponds to the case $\lambda^2 = 10,000$.

Examining Figure 3.3, two ROIs stand out has having a relatively broad genetic signal that persists even as the tuning parameter increases from $\lambda^2 = 1000$ to $\lambda^2 = 10,000$. These are *Left-Supramarg* (thickness of the left supramarginal gyrus) and *Left-SupTemporal* (thickness of the left superior temporal gyrus). For the case where $\lambda^2 = 1000$ phenotype *Left-Supramarg* is associated with 188 SNPs (top panel of Figure 3.3) and this decreases to 24 SNPs (bottom panel of Figure 3.3) when $\lambda^2 = 10,000$. When $\lambda^2 = 1000$ phenotype *Left-SupTemporal* is associated with 188 SNPs and this decreases to 23 SNPs when $\lambda^2 = 10,000$.

3.5 Conclusion

We have developed a spatial multi-task regression model for relating genetic data to multivariate imaging phenotypes. The model uses a shrinkage prior with group penalization for the coefficients of each SNP (rows of $\mathbf{W}$) in the regression structure. The error structure, for the imaging phenotype is based on a proper bivariate conditional autoregressive model, which allows for both spatial correlation as well as bilateral correlation across brain hemispheres. Our model is one of the first explicitly spatial hierarchical models for imaging genetics and neuroimaging to account for both spatial correlation and bilateral correlation. The new model along with Bayesian FDR procedures and both VB and Gibbs sampling algorithms are implemented in the latest version of the *bgsmtr* R package.

With regards to the two computational algorithms, we recomend that the approximate VB procedures be used to initialize the MCMC algorithm and also to obtain an initial insight into the data while the MCMC sampler runs. It appears that VB combined with Bayesian FDR tends to be too liberal in selection of SNPs, and in our application the SNP-ROI pairs selected by MCMC + Bayesian FDR are a proper subset of that selected by VB + Bayesian FDR.

Our analysis of the ADNI-1 data found a consistent signal from APOE SNP rs405509 as well as ACE SNP rs4311. In both cases phenotypes in the right hemisphere of the brain seem to be favoured. In terms of having a broad genetic signal, the thickness of the left supramarginal gyrus and the thickness of the left superior temporal gyrus seem to be associated with the largest number of SNPs.

While our current methodology is best suited for situations where the analysis is focussed on a relatively small set of targeted SNPs (no more than a few thousand) and a moderate number of ROIs (no more than 100), these are settings in which

a full multivariate model for imaging genetics can be specified and fit. Extending applicability of the methodology to settings with massive numbers of genetic and neuroimaging variables is an avenue for future work. Divide and conquer strategies such as the consensus Monte Carlo algorithm (Scott et al., 2016) as well as splitting up the brain into a smaller number of sub-regions might lead to feasible implementations for such settings. Their design and implementation for imaging genetics should prove to be a substantial challenge.

Chapter 4

A Bayesian Approach to the Mixed-Effects Analysis of Accuracy Data in Repeated-Measures Designs

4.1 Introduction

Many types of behavioural data generated by experimental investigations of human language, memory, and other cognitive processes processes entail the measurement of response accuracy. For example, in studies of word identification, error rates in word-naming or lexical-decision tasks are analyzed to determine whether manipulated variables or item characteristics influence response accuracy (e.g., Chateau and Jared, 2003; Yap et al,. 2008). Similarly, in experiments on memory topics such as false memory and the avoidance of retroactive and proactive interference on recall, response errors or probability of accurate responding are the critical measures of performance (e.g., Arndt and Reder, 2003; Jacoby et al., 2015).

For example, we were investigating the influence of a semantic context on the identification of printed words shown either under clear (high contrast) or degraded (low contrast) conditions. The semantic context consisted of a prime word presented in advance of the target item. On critical trials, the target item was a word and on other trials the target was a nonword. The task was to classify the target on each trial as a word or a nonword (this is called a "lexical decision" task). Our interest

was confined to trials with word targets. The prime word was either semantically related or unrelated to the target word (e.g., granite-STONE vs. attack-FLOWER), and the target word was presented either in clear or degraded form. Combining these two factors produced four conditions (related-clear, unrelated-clear, related-degraded, unrelated-degraded). For the current analysis, accuracy of response was the dependent measure.

The common treatment of accuracy or error-rate data has, and to a large extent continues, to consist of aggregating data across trials within each condition for each subject to generate the equivalent of a proportion correct or incorrect score, ranging from 0 to 1. These scores are then analyzed using repeated-measures analysis of variance (ANOVA) or, in the simplest cases, a t test. Although this standard approach, hereafter termed 'standard aggregating approach', has serious problems that have repeatedly been pointed out to researchers, it continues to be used. Here we illustrate a solution to these problems offered by Bayesian data analysis. We first discuss the problems of the standard aggregating approach. Then we summarize one approach to this problem that has gained traction over the last decade (non-Bayesian Generalized Linear Mixed Models), followed by a brief review of some of the general pros and cons of Bayesian approaches. This chapter presents a Bayesian statistical modeling framework for repeated-measures accuracy data, simulation studies evaluating the proposed methodology, and an application to actual data arising from a single-factor repeated-measures design.

4.1.1 The Standard Aggregating Approach

To assess the validity of our characterization of how researchers typically analyze accuracy, error, or other classification data, we examined articles published in recent issues of four of the leading journals in the field of cognitive psychology: the *Journal of Memory and Language (JML)*, the *Journal of Experimental Psychology: Learning, Memory, and Cognition (JEP)*, *Cognition*, and *Cognitive Psychology*. All articles appearing in issues with a publication date of January to August 2016 (up to the October 2016 issue for *JML* because later issues of that journal were available at the time the survey was conducted) were considered. Articles in which accuracy was analyzed using a transformed measure such as d', receiver operating characteristic curves, or parameters of computational models based on simulation of accuracy data were not included due to no real data application. A total of 180 articles across the

four journals reported data expressed as proportions or the equivalent (e.g., accuracy, error, classification responses). Among these articles, 69 were on a topic related to language processing and the remaining 111 addressed other issues in memory and cognition. For each article, we determined whether the authors used standard methods of analyzing data that included aggregating performance across items or across subjects or whether generalized linear mixed models were used in which individual trials were the units of analysis. We included in the standard-analysis category any standard univariate method of analysis, such as analysis of variance, t tests, correlation, and regression in which data were aggregated over items or over subjects. The application of analysis of variance using subjects and items as random effects in separate analyses and reporting F_1 and F_2 was also classified as using a method of aggregation, where F_1 is the computed F-ratio when condition effect is tested against the conditions by subjects interaction and F_2 is the computed F-ratio when condition effect is tested against the item within condition effect. This approach, used widely since Clark's (1973) seminal paper on item variability, relies on an analysis that aggregates across defined subsets of trials (items for F_1 and subjects for F_2), rather than analyzing data at the level of individual trials. Our assessment indicated that for articles on language-related topics, 37 articles (54%) applied some form of the standard aggregating approach (of these, 15 articles used methods that reported effects aggregated over subjects and effects aggregated over items; i.e., F_1 and F_2). For articles on other topics of memory and cognition, 99 (89%) relied on the standard aggregating approach (two of these reported F_1 and F_2 analyses). Overall, then, 76% of recently published articles in these four leading cognitive psychology journals analyzed accuracy or other binomial data in the historically standard way, which involves aggregating performance across items for at least a subset of the analyses. The remaining articles used generalized linear mixed models to analyze the data, which does not aggregate across items and which we discuss in detail below.

The shortcomings of what continues to be a widely applied method of analyzing accuracy data, and binomial data in general (i.e., aggregating across items), have been known for some time (Cochran, 1940) and have been reiterated in recent accounts of alternative approaches (e.g., Dixon, 2008; Jaeger, 2008; Quené and Van den Bergh, 2008). For instance, the proportions generated from binary observations (correct versus incorrect) need not be normally distributed, which violates one of the

In four of the articles reporting linear mixed model regression analyses of accuracy, it was not clear whether logistic regression was used or whether raw accuracy was the dependent measure.

fundamental assumptions of ANOVA and t tests. Moreover, the variance of accuracy scores will depend on the mean, with larger variance when the mean is closer to .5 and variance vanishing to zero as the mean approaches 0 or 1. This dependency implies that if effects are present (i.e., means vary across conditions), the assumption of homogeneity of variance, on which ANOVA depends, is likely violated. By aggregating data across trials, error variance is likely reduced, leading to an elevation of type I error probability in null-hypothesis significance testing (Quené and Van den Bergh, 2008). Also, the dependence across trials leads to the violation of independent experiment assumption for using binomial distribution with likelihood ration test. Finally, because the proportion correct is bounded by 0 and 1, confidence intervals created from such data may well extend outside that range when the relevant mean approaches one of these limits, meaning that probability mass is being assigned to impossible values (Dixon, 2008; Jaeger, 2008).

A common strategy that is adopted to avoid these problems is the application of a data transformation such as the square-root arcsine transformation. Unfortunately, this approach makes the interpretation of the analysis more difficult as the hypothesis tests then correspond to the means of the transformed data and not to the actual accuracy data. Jaeger (2008) also shows that these transformations do not fix the problem when the mean proportions are close to 0 or 1. Furthermore, transforming the data after aggregating across items precludes the investigation of item effects.

4.1.2 Generalized Linear Mixed Models

A viable solution to these difficulties with the standard aggregating approach to analyzing accuracy data involves using generalized linear mixed-models of logistic regression (Dixon, 2008; Jaeger, 2008; Quené and Van den Bergh, 2008). In this setting a hierarchical model based on two levels is specified for the data, where, at the first level the response variables are assumed to be generated from a Bernoulli distribution. At the second level of the model the accuracy or error rates are converted to a logit scale (the logarithm of the odds of success or failure): $\text{logit}(p) = \ln(p/(1-p))$ and the variability in the log-odds across subjects, items, and conditions is based on a mixed effects model. We emphasize here that p is not computed from the data and does not correspond to the proportion of accurate responses aggregated over items for a given condition and subject; rather, p is an unknown parameter representing the probability of an accurate response for a given subject, item, and experimental

condition. Rather than aggregating data over trials to obtain a single estimate of the proportion correct in a given condition for each subject, the individual binary accuracy trial scores are the unit of measurement. This level of granularity allows the assessment of possible random effects for both subjects and items. That is, effects of a manipulation may not be consistent from subject to subject or item to item and a mixed-effects analysis can characterize the extent of these differences. Variance in effects across items can thus be assessed, which addresses the concern raised by Clark (1973) about the "language-as-a-fixed-effect fallacy" (Jaeger, 2008; Quené and Van den Bergh, 2008).

The proposed use of mixed-effects logistic regression for the analysis of accuracy data can be implemented either with or without significance tests. In the latter case, information criteria such as the Akaike Information Criterion (AIC) proposed by Akaike (1973) can be used for model selection. In the former case, these analyses continue to rely on the basic principles of null-hypothesis significance testing (NHST) for making decisions about whether independent variables are producing effects on performance. A number of recent reports in the psychological literature have highlighted potential deficiencies associated with NHST (e.g., Kruschke, 2013; Rouder et al., 2009; Rouder et al., 2012; Wagenmakers, 2007). We will briefly mention only a few of those difficulties here.

First, the probability value associated with a significance test reflects the probability of obtaining an observed result, or one more extreme, on the assumption that the null hypothesis is true. An inference must then follow, establishing one's belief that the null hypothesis is false on those occasions where the obtained probability value is very low. Many researchers mistakenly interpret that probability as the likelihood that the null hypothesis is true, given the observed data (e.g., Haller and Krauss, 2002). That inference is not available through NHST, but it can, for example, be generated by a Bayesian analysis. Second, NHST is, by design, capable of providing evidence in favour of only the alternative hypothesis. When evidence does not allow rejection of the competing null hypothesis, no strong conclusion can be reached. Another potential advantage of the Bayesian approach is that it allows the strength of evidence in favor of either a null or an alternative hypothesis to be quantified. Although such reasoning can, and has been, accommodated under some non-Bayesian approaches, it follows naturally from the Bayesian perspective and Bayesian methods provide one principled approach to doing this. Finally, although not an inherent problem of non-Bayesian approaches but rather a potential pitfall of their misuse,

when using NHST researchers are susceptible to the problems caused by optional stopping during data collection. One may be tempted, for example, to collect additional data if a NHST applied to data currently in hand generates a p value that is just shy of significance. It has been clearly demonstrated that this approach to data collection substantially raises the probability of a type I error (e.g., Wagenmakers, 2007), whereas Bayesian analysis is not susceptible to this problem and will only yield increasingly accurate results as more data accumulate (Berger and Berry, 1988; Wagenmakers, 2007), assuming the methods are used adequately. Again, we emphasize that this is not an inherent problem of non-Bayesian approaches but one that can and often does arise when these procedures are misused.

4.1.3 Bayesian Approaches

As a solution to this problem with NHST and to advance the use of mixed-effects analyses of binomial data, we propose the use of a Bayesian version of mixed-effects models of logistic and probit regression for the repeated-measures case. The nature of these modified versions of regression analysis is discussed in detail below. Although Bayesian analysis of generalized linear mixed-models has been developed extensively in the statistical literature over the past decade, our interest is specifically on the application to behavioural data arising from accuracy studies where a method combining the use of Bayesian analysis with generalized linear mixed-models for binomial data has not been considered previously. We investigate two options for selecting between null and alternative hypotheses (models) using Bayesian analysis: the Bayes factor computed using a Bayesian Information Criterion (BIC) by approximation (Wagenmakers, 2007), and the Watanabe-Akaike Information Criterion (WAIC) by Watanabe (2010). We used R software (R Development Core Team, 2017) to implement these approaches in addition to computing posterior distributions.

Although the Bayesian approach offers an exciting avenue for the analysis of memory and language data, there is a large body of work that debates the pros and cons of a Bayesian analysis. Efron (1986) discussed the potential problems with the Bayesian approach and provided examples where the frequentist approach provides easier solutions. The use of Bayesian approaches can lead to an increase in conceptual complexity of some aspects of the data analysis. Users must take the time to acquire the necessary background in order to use Bayesian methods appropriately. In addition, an important issue that has received a lot of attention in the literature is the choice

of the prior distribution, which can have an influence on the results. Informative prior distributions can be chosen to reflect prior knowledge on certain parameters of a model, though formulating such priors can be very difficult and is often impractical. In our work, we favour the use of weakly informative priors that have high or infinite prior variance so that the prior plays a limited role in the inference. For Bayesian logistic regression the issue of priors and the development of weakly informative priors is discussed extensively in Gelman et al. (2008).

The use of logistic mixed models is an alternative to methods that involve aggregation over items or subjects. As demonstrated in the literature, this alternative is a more effective methodology for the analysis of repeated-measures accuracy studies. In the memory and language literature this has been considered primarily from a classical, non-Bayesian perspective (Dixon 2008; Jaeger 2008; Quenè and Van den Bergh 2008). In parallel, there is currently a shift towards the use of Bayesian methods for the analysis of cognitive studies (Wagenmakers 2007; Rouder et al. 2009; Rouder et al. 2012). To date, much of this shift has focused on the analysis of continuous response variables. The goal of this project is to provide tools for memory and language researchers to combine the advantages of both logistic/probit mixed models and Bayesian methods for the analysis of repeated-measures studies of accuracy. In doing so, we allow for the evaluation of posterior distributions for the effects of interest, and we also offer two possible approaches for Bayesian model selection (a) the Bayes factor based on the BIC approximation (Wagenmakers, 2007); and (b) the more recently proposed WAIC. These two approaches are motivated based on different utilities, the former assesses model performance using the marginal likelihood while the latter assesses model performance by estimating a measure of out-of-sample prediction error.

Given that we offer two different approaches for Bayesian model selection it is naturally of interest to consider comparisons between them. We therefore make these comparisons by evaluating the operating characteristics of both approaches using simulation studies. In order to place these comparisons within the larger field of methods that can be applied to repeated-measures accuracy data we also evaluate two other approaches. One is a Bayesian analysis based on item aggregation with model selection determined by the Bayes factor. The other is logistic mixed modeling within the classical (non-Bayesian) setting with model selection based on the standard Akaike Information Criterion (AIC) by Akaike (1973). The latter is arguably the current non-Bayesian state-of-the-art. We use these evaluations to inform a discussion on the pros

and cons of the Bayesian approach and its combination with logistic/probit mixed modelling.

The primary contributions of our work are: (a) Facilitating the combination of binomial mixed-effects modeling and Bayesian inference for repeated-measures analysis of accuracy studies, (b) simulation studies evaluating this approach relative to the standard aggregating approach and classical logistic mixed models, and (c) making available easy-to-use R software with examples facilitating such analyes.

4.2 Method

4.2.1 Statistical Models for Repeated-Measures Accuracy Studies using Single-Factor Designs

Data

We first consider a repeated-measures design involving K subjects, I experimental conditions corresponding to a single factor, and J items. Generally the number of experimental conditions will be much smaller than the number of items, $I \ll J$. For each subject the data consist of a binary response measuring accuracy on each of the J items. For example, item could be a single word that given to a subject. The response on each item occurs in a particular experimental condition and these conditions are randomly assigned across items. The structure of the data are illustrated in Table 4.1 for the case where there are $I = 3$ experimental conditions with labels b, g, or r, and these labels are indicated as subscripts on each of the binary measurements, the latter taking values either 0 or 1. It should be noted that our framework can accommodate cases where items are seen by the same subject under multiple conditions. In that case the data could be represented by adding additional columns for each item in Table 4.1. Furthermore, the studies considered here may use counterbalancing, so that for some subjects a particular item will be assigned to a particular condition whereas for other subjects that same item will be assigned to a different condition. We also do not require that a response be obtained on every item for each subject, so it is possible for different subsets of items to be presented to different subjects.

Table 4.1: Example data structure: $I = 3$ conditions, K subjects, J items where conditions are indicated as subscripts b, g, or r of each binary data value.

	Item 1	Item 2	...	...	Item J
Subject 1	1_b	1_r	1_g	0_b	1_b
Subject 2	0_g	0_b	0_r	1_g	0_g
$\vdots$	0_r	0_r	0_g	1_r	0_g
$\vdots$	0_r	0_g	0_r	1_b	0_r
Subject K	0_b	0_b	0_b	1_r	0_b

Model Development

We let Y_{ijk} denote the binary response obtained from subject k when item j is assigned to condition i, where $i = 1, \ldots, 3,\ j = 1, \ldots, J$ and $k = 1, \ldots, K$. The standard aggregating approach often applied in the analysis of such data, begins by averaging the response variables across the items for each condition to obtain accuracy scores corresponding to each subject and condition. Referring to Table 4.1, this averaging results in three response variables for each row of the table, one for each of the three conditions. A repeated-measures ANOVA is then applied to the aggregated data. In contrast, trial-based analyses like those pursued here (and in Dixon, 2008) avoid averaging over items and model each binary score using a Bernoulli distribution. We let p_{ijk} denote the probability of an accurate response ($Y_{ijk} = 1$) when item j is assigned to condition i and subject k. The model assumes

$$Y_{ijk}|p_{ijk} \overset{ind}{\sim} \text{Bernoulli}(p_{ijk}),$$

and we emphasize that our modeling approach is specified at the level of binary observations (i.e., Y_{ijk} is either 0 or 1) and the accuracy probability p is a parameter of the corresponding Bernoulli distribution with $p = Pr(Y = 1)$. Each of these binary observations is specific to a particular subject, item, and level of the experimental condition. The analysis we propose evaluates a number of models each corresponding to different assumptions on how this probability varies across items, subjects, and conditions. We note that the experimental design is such that we expect lack of independence of the data within subject (the rows of Table 4.1) and also within item (the columns of Table 4.1). As a result all of the models that we consider include random effects for both subjects and items (Clark, 1973) to account for possible

dependence.

These models rely on first transforming the accuracy probability (p_{ijk}) using a link function $g(p) : [0, 1] \to \mathbb{R}$. That is, the link function, g, takes an value of p from 0 to 1 and maps it onto an element of the set of real numbers. We consider two common link functions $g(p) = \log\left(\frac{p}{1-p}\right)$ which corresponds to a logistic model, and $g(p) = \Phi^{-1}(p)$ which corresponds to a probit model, where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. In the latter model it is useful to think of the probability values, p, as being converted to a corresponding Z-score. The two link functions are depicted in the left panel of Figure 4.1.

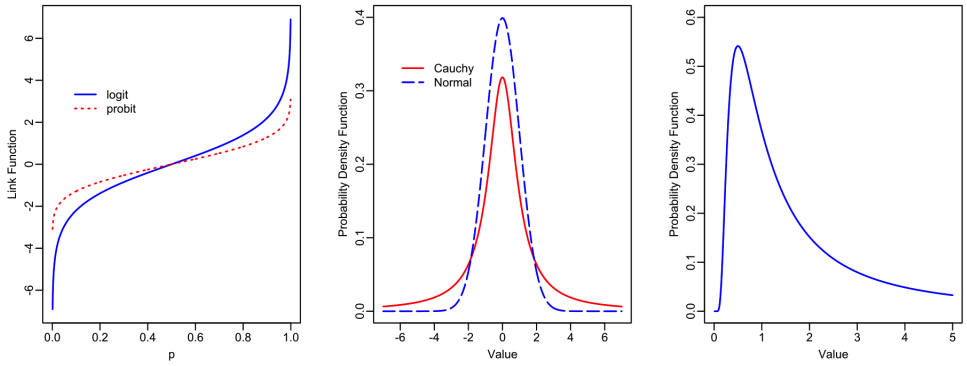


Figure 4.1: Left Panel - the logistic and probit link functions; Centre Panel - the probability density function of a Cauchy distribution and a normal distribution; Right Panel - the probability density function of an inverse-Gamma distribution.

The hierarchical Bayesian models for accuracy are presented below. The different models represent different possible effects. Following the recommendations of Gelman (2008), Cauchy prior distributions are assigned to the fixed effects in all of the models for computation convenience. The centre panel of Figure 4.1 provides an illustration of the Cauchy prior distribution in relation to the normal distribution. As is typical in Bayesian generalized linear mixed models, the random effects are assigned normal distributions and the variance components corresponding to the random effects are assigned inverse-gamma distributions (Gelman, 2014). The latter distribution is illustrated in the right panel of Figure 4.1 and is a convenient choice as it means that the algorithm used to fit the model to the data has an analytical form that is easy to work with.

1. *(LM_0 - Logit/PM_0 - Probit) Baseline model with random subject and item ef-*

fects with no effect of the experimental condition:

$$g(p_{ijk}) = \beta_0 + a_j^{(R)} + b_k^{(R)}$$

$$a_j^{(R)} \overset{iid}{\sim} N(0, \sigma_a^2), \quad b_k^{(R)} \overset{iid}{\sim} N(0, \sigma_b^2), \quad \beta_0 \sim \text{Cauchy}(0, \sigma_\beta = 10)$$

$$\sigma_a^2 \sim \text{Inverse-Gamma}(\kappa_{\sigma_a}, \tau_{\sigma_a}), \quad \sigma_b^2 \sim \text{Inverse-Gamma}(\kappa_{\sigma_b}, \tau_{\sigma_b})$$

where $i = 1, \ldots, I,\ j = 1, \ldots, J$ and $k = 1, \ldots, K$. Here β_0 is the model intercept, superscript (R) denotes random effect, $a_j^{(R)}$ is the item random effect with variance σ_a^2, and $b_k^{(R)}$ is the subject random effect with variance σ_b^2. The hyperparameters $\kappa_{\sigma_a}, \tau_{\sigma_a}, \kappa_{\sigma_b}, \tau_{\sigma_b}$ are fixed to values that make the inverse-gamma prior distributions weakly informative, with infinite variance. We reiterate that this model assumes that there is no effect of the experimental condition on the probability of an accurate response and this is the only model where this assumption is made. This model corresponds to the null hypothesis that there is no effect of the experimental condition in the data.

2. *(LM_F - Logit/PM_F - Probit) Fixed effect for the experimental condition:*

$$g(p_{ijk}) = \beta_0 + \alpha_i + a_j^{(R)} + b_k^{(R)}$$

with $\alpha_1 = 0,\ \alpha_i \overset{iid}{\sim} \text{Cauchy}(0, \sigma_\alpha = 2.5),\ i = 2, \ldots, I$, and with all other priors identical to model 1. The constraint $\alpha_1 = 0$ is imposed for model identification and as a result one arbitrarily selected experimental condition is considered a baseline condition and the remaining fixed effects α_i represent the effect of condition i relative to that baseline. The value of the scale parameter $\sigma_\alpha = 2.5$ in the Cauchy prior distribution for the fixed effects is different from the corresponding value in the prior for the intercept $\sigma_\beta = 10$ based on the work of Gelman et al. (2008) who recommend these choices for Bayesian logistic regression as a weakly informative prior. The Cauchy prior is centered around zero so that there is no preference for either a positive or negative effect. Using cross-validation, Gelman et al. (2008) show that this class of priors outperform Gaussian and Laplace priors and we use it here for both logistic and probit regression. This model extends the first model by assuming that the probability of an accurate response depends on the experimental condition through a fixed effect α_i.

3. *(LM_{R_S} - Logit/PM_{R_S} - Probit) The effect of experimental condition varies across subjects:*

$$g(p_{ijk}) = \beta_0 + \alpha_i + a_j^{(R)} + b_k^{(R)} + (\alpha b)_{ik}^{(R)}$$

where, in this case, the effect of the experimental condition is represented by both the fixed effects α_i and the random effects $(\alpha b)_{ik}^{(R)}$ which represent an interaction between subject and condition, implying that the effect of condition varies across subjects. As before, constraints are imposed so that one arbitrarily selected condition is taken as a baseline condition, $(\alpha b)_{1k}^{(R)} = 0$, and the remaining random effects are assumed to be normally distributed $(\alpha b)_{ik}^{(R)} \overset{iid}{\sim} N(0, \sigma^2_{\alpha b})$, $i = 2, \ldots, I; k = 1, \ldots, K$. An inverse-gamma prior distribution is assumed for the corresponding variance component, $\sigma^2_{\alpha b} \sim$ Inverse-Gamma$(\kappa_{\sigma_{\alpha b}}, \tau_{\sigma_{\alpha b}})$ with hyper-parameters $\kappa_{\sigma_{\alpha b}}, \tau_{\sigma_{\alpha b}}$ fixed to values that make the inverse-gamma prior distribution weakly informative, with infinite variance.

4. *(LM_{R_i} - Logit/PM_{R_i} - Probit) The effect of experimental condition varies across items:*

$$g(p_{ijk}) = \beta_0 + \alpha_i + a_j^{(R)} + b_k^{(R)} + (\alpha a)_{ij}^{(R)}$$

with the constraint, $(\alpha a)_{1j}^{(R)} = 0$, imposed so that one arbitrarily selected condition is taken as a baseline condition and the remaining random effects are assumed to be normally distributed, $(\alpha a)_{ij}^{(R)} \overset{iid}{\sim} N(0, \sigma^2_{\alpha a})$, $i = 2, \ldots, I; j = 1, \ldots, J$. An inverse-gamma prior distribution is assumed for the corresponding variance component, $\sigma^2_{\alpha a} \sim$ Inverse-Gamma$(\kappa_{\sigma_{\alpha a}}, \tau_{\sigma_{\alpha a}})$ with $\kappa_{\sigma_{\alpha a}}, \tau_{\sigma_{\alpha a}}$ fixed to values that make the inverse-gamma prior distribution weakly informative, with infinite variance. All other prior distributions are identical to model 2. In this case the effect of the experimental condition is represented by both the fixed effects α_i and the random effects $(\alpha a)_{ij}^{(R)}$ which represent an interaction between item and condition (items potentially vary in the extent to which they exhibit effects of the experimental conditions).

5. *($LM_{R_{S,i}}$ - Logit/$PM_{R_{S,i}}$ - Probit) The effect of experimental condition varies across items and subjects:*

$$g(p_{ijk}) = \beta_0 + \alpha_i + a_j^{(R)} + b_k^{(R)} + (\alpha a)_{ij}^{(R)} + (\alpha b)_{ik}^{(R)}$$

with distributions for random effects (condition-by-item and condition-by-subject effects) and hyper-priors set as in models 3 and 4. This is the most general of the models considered for a single-factor repeated measures design.

Each of the five models presented above represents different assumptions on the effect of the experimental conditions while explicitly modeling the binary response through the Bernoulli distribution and accounting for between-subject and between-item variability with random effects. Considering the two possible choices for the link function, logit or probit, there are ten possible models and these models are summarized in Table 4.2.

Table 4.2: The full set of Bernoulli mixed models for single-factor designs representing different assumptions about effects on the accuracy probability.

Model	**Link**	**Condition Effect**
LM_0	Logistic	Null
LM_F	Logistic	Fixed
LM_{R_s}	Logistic	Varies across subjects
LM_{R_i}	Logistic	Varies across items
$LM_{R_{s,i}}$	Logistic	Varies across subjects and items
PM_0	Probit	Null
PM_F	Probit	Fixed
PM_{R_s}	Probit	Varies across subjects
PM_{R_i}	Probit	Varies across items
$PM_{R_{s,i}}$	Probit	Varies across subjects and items

4.2.2 Analysis of Two-Factor Designs

In the case of a design with two experimental factors we let Y_{hijk} denote the binary response obtained from subject k when item j is assigned to condition i of the first factor and condition h of the second factor, $k = 1, \ldots, K$, $j = 1, \ldots, J$, $i = 1, \ldots, I$, $h = 1, \ldots, H$. We assume

$$Y_{hijk}|p_{hijk} \overset{ind}{\sim} \text{Bernoulli}(p_{hijk})$$

where p_{hijk} is the corresponding probability of an accurate response. Different models correspond to different assumptions on how this probability varies across items,

subjects, and the levels of the two experimental factors. The most general model we consider for a two-factor design takes the form

$$g(p_{hijk}) = \beta_0 + \gamma_h + \alpha_i + (\gamma\alpha)_{hi} + a_j^{(R)} + b_k^{(R)} + (\gamma a)_{hj}^{(R)} + (\alpha a)_{ij}^{(R)} + (\gamma b)_{hk}^{(R)} + (\alpha b)_{ik}^{(R)}$$

where α_i, γ_h, and $(\gamma\alpha)_{hi}$ are fixed effects corresponding to the first factor, the second factor, and their interaction respectively. As before $a_j^{(R)}$ and $b_k^{(R)}$ are random effects for items and subjects while the random effects $(\gamma a)_{hj}^{(R)}$, $(\alpha a)_{ij}^{(R)}$, $(\gamma b)_{hk}^{(R)}$ and $(\alpha b)_{ik}^{(R)}$ allow the effects of the two experimental factors to vary across items and subjects. Just as with the single-factor case, we assume Cauchy priors for the fixed effects and adopt the identification constraints $\alpha_1 = \gamma_1 = (\gamma\alpha)_{1i} = (\gamma\alpha)_{h1} = 0$ so that the first level of both factors are taken to be baseline conditions. The random effects are assigned normal priors as before and the corresponding variance components are assigned weakly informative inverse-gamma hyper-priors.

For simplicity, we do not consider models where the interaction between the two experimental factors itself interacts with either items or subjects. Although allowing such terms increases the flexibility of the model, the associated parameters can be only weakly estimatable in practice and will therefore exhibit a low degree of Bayesian learning (prior-to-posterior movement) in particular with binary data. We note that there is some controversy in the literature with regards to the inclusion of high-order random effects in mixed models. For example, Barr et al. (2013) suggest that linear mixed models generalize best when the maximal random effects structure supported by the design is employed; however, Bates et al. (2015) indicate problems with this suggestion, including convergence problems of numerical algorithms for fitting mixed models and the potential lack of model interpretability. Our choice to exclude models with random effects that represent three-way interactions is inline with the discussion in Bates et al (2015).

Many different models can be obtained by removing certain terms in the model equation above, and either the logit or probit link can be employed. For a single-factor design, the set of ten possible models is summarized in Table 4.2, where, for example, LM_{R_i} (PM_{R_i}) denotes the logistic (probit) model where the effect of the experimental condition is represented as a random effect that varies across items, and we assume the presence of a fixed effect in any model that contains a corresponding random effect for the experimental condition. In the case of two factors, the set of models obtainable by removing appropriate terms from the general model equation above is considerably

larger. For the logistic (probit) link we refer to specific models using the notation $LM_xN_yI_z$ ($PM_xN_yI_z$), where $x \in \{0, F, R_s, R_i, R_{s,i}\}$ denotes the model structure for the first factor as defined for single-factor designs in Table 4.2, $y \in \{0, F, R_s, R_i, R_{s,i}\}$ similarly denotes the model structure for the second factor, and $z \in \{0, 1\}$ is used to specify the presence or absence of an interaction between the two factors, with $z = 1$ ($z = 0$) indicating presence (absence). For example, $LM_{R_{s,i}}N_{R_{s,i}}I_1$ denotes the most general model specified in the equation above with a logit link, $PM_0N_0I_0$ denotes the null probit model where all terms corresponding to the effects of the two factors have been removed, and $LM_{R_s}N_{R_{s,i}}I_1$ denotes a logistic model where the effect of the first factor varies across subjects, the effect of the second factor varies across subjects and items, and the interaction between the factors is included.

4.2.3 Model Fitting and Software

The posterior distribution of the model parameters (fixed effects, random effects, and variance components) associated with each model can be computed using standard Markov chain Monte Carlo (MCMC) sampling algorithms. These procedures can be implemented in the R (R Development Core Team, 2017) and JAGS (Plummer, 2003), programming languages in conjunction with the R package 'rjags' (Plummer, 2013) which provides an interface between the two. We have developed an R function 'BinBayes.R' that allows these models and algorithms to be used in a relatively straightforward manner requiring only very basic knowledge of the R language. The software along with a detailed illustration of its use for single-factor and two-factor repeated-measures designs, sample data, and examples are available for download at: `https://v2south.github.io/BinBayes/`.

4.2.4 Bayesian Model Comparison

For a given dataset, we are able to summarize the posterior distribution for any of the models for single-factor and two-factor designs. An arguably more important task is the comparison of models, as each model represents different assumptions regarding the effect of the experimental condition on the probability of response accuracy. For example, and in reference to Table 4.2, a comparison of models LM_0 and LM_F corresponds to testing for a fixed effect of the experimental condition, whereas a comparison of models LM_0 and LM_{R_i} corresponds to testing for an effect of the experimental condition that allows for this effect to vary across items. As the

link function is a modelling choice, logit and probit models can also be compared (e.g. LM_F and PM_F) to determine which is more appropriate for the data at hand.

The traditional approach for model comparison in the Bayesian framework is based on the Bayes factor. Given two models denoted by $M0$ and $M1$ the Bayes factor comparing $M0$ to $M1$ (Wagenmakers, 2007) is defined as

$$BF_{01} = \frac{Pr(\mathbf{y}|M0)}{Pr(\mathbf{y}|M1)}$$

where $\mathbf{y}$ denotes the data, and $Pr(\mathbf{y}|M)$ denotes the probability of the data under M. A value of $BF_{01} > 1$ can be viewed as evidence in favour of model $M0$ over $M1$ in the sense that the probability of the data is higher under $M0$. Kass and Raftery (1995) provide a comprehensive review of the Bayes factor including information about its interpretation where it is suggested that a value of $BF_{01} \geq 3$ corresponds to positive evidence in favour of model $M0$ over $M1$, whereas, decisive evidence corresponds to $BF_{01} > 150$.

In general, the Bayes factor can be difficult to compute and a great deal of research in the area of statistical computing has been dedicated to this problem (see e.g. Chib and Jeliazkov, 2001; Meng and Wong, 1996; Meng and Schilling, 2003; Chen, 2005; Raftery et al., 2007). A number of Monte Carlo algorithms can be applied for the computation of the Bayes factor; however, for the Bernoulli mixed models under consideration in this article, we have found that stable estimation of the Bayes factor is extremely time consuming (e.g. several hours to days on a fast laptop for datasets of standard to large size). As a more practical alternative that is easy to compute in just a few seconds, we use the Bayesian information criterion (BIC) defined for a given model M by

$$BIC(M) = -2 \log \hat{L} + p \log n,$$

where $\hat{L}$ is the maximized likelihood function for model M, p is the number of parameters in the model, and n is the sample size. In the case of generalized linear mixed models for repeated-measures designs, both the number of parameters p and the sample size n are not straightforward to define (Jones, 2011; Spiegelhalter et al., 2002). We will assume that p excludes the random effects but includes the corresponding variance components and the number of fixed effects. Indeed, this definition for p is the default used in the computation of the BIC in the R package lme4 (Bates et al., 2014). For the sample size, we assume that $n = K$, the number of subjects. This

is considered in detail and suggested by Berger et al., (2014); see also Nathoo and Masson, (2016).

Given the value of BIC for two competing models the Bayes factor is approximated by $BF_{01} \approx \exp\{(BIC(M1) - BIC(M0))/2\}$ where this expression assumes $Pr(M0) = Pr(M1)$ a priori and the accuracy of approximation will increase with the sample size. The approximation is based on a unit information prior for the model parameters (see e.g. Kass and Raftery, 1995; Masson, 2011; Nathoo and Masson, 2016; Wagenmakers, 2007). Alternatively, given a set of competing models such as those listed in Table 4.2, the BIC for each model can be computed in order to rank the models, with lower values corresponding to preferred models. Typically we require a difference in the BIC scores of two models to be at least $|\Delta BIC| = 2$ (Kass and Raftery, 1995) in order to claim that there is positive evidence in favour of one model over the other, which corresponds to a Bayes factor of $BF \approx 2.72$.

Rather than comparing models based on Bayes factors, an alternative second approach that can be used to compare models is based on an evaluation of how well each model can predict new data or heldout data. By heldout data, we mean a subset of the data that is not used in the process of fitting the model but is subsequently used to evaluate the predictive ability of the model. In this context, cross-validation is a common approach for estimating the out-of-sample prediction error which can then be used to compare models (Gelman et al., 2014). As cross-validation requires splitting the data into multiple partitions and then repeatedly fitting the model to subsets of the data, which is computationally demanding, alternative measures of predictive accuracy that in some sense approximate cross-validation have been proposed for model selection. One such approximation that has been applied extensively for model selection is AIC (Akaike, 1973) which is computed using the maximum likelihood estimator. The computation of AIC, like BIC, requires a value for the number of model parameters p which is not clearly defined in the case of hierarchical models as described above. An alternative, fully Bayesian approximation to cross-validation that avoids this problem, is WAIC, which has been proposed by Watanabe (2010) and takes the form

$$WAIC = -2 \sum_{k=1}^{K} \log E_{\theta}[p(\mathbf{y}_k|\theta)|\mathbf{y}_1, \ldots, \mathbf{y}_K]$$

$$+2\sum_{k=1}^{K} Var_\theta[\log p(\mathbf{y}_k|\theta)|\mathbf{y}_1,\ldots,\mathbf{y}_K]$$

where $\mathbf{y}_k$ is the data collected from subject k, θ denotes the set of all unknown parameters, $p(\mathbf{y}_k|\theta)$ denotes the probability mass function of $\mathbf{y}_k$ conditional on θ, and the expectation $E_\theta[\cdot]$ and variance $Var_\theta[\cdot]$ are taken with respect to the posterior distribution of the model parameters. The first part of the formula for WAIC provides a measure of the fit of the model against the data and the second part is meant to capture the inherent complexity of the model, where the reasoning is that more complex models are a priori less likely (i.e., it is a form of Occam's razor).

In practice, these are computed using an MCMC algorithm that is used to fit the model. The WAIC is thus easily computed as a by-product of fitting the model. As discussed in Gelman et al. (2014), the WAIC has the desirable property of averaging over the posterior distribution of the model parameters, that is, considering many possible values of the model parameters weighted by their posterior density, rather than conditioning on just a single value of the parameter, the maximum likelihood estimator as is done with AIC. This is desirable because it captures the estimated uncertainty into the estimates of the model given the assumptions the researcher was willing to make about the prior and model structure.

More importantly, the WAIC works well with so called singular models, that is, models with hierarchical structures where the number of parameters increases with sample size. It is thus particularly well suited for the random effect models we are considering, and unlike the penalties used by BIC and AIC, the penalty used in WAIC has been rigorously justified for model comparison in this setting (Watanabe, 2010).

Although a generally applicable calibration for differences in the WAIC scores of two models is currently lacking, a reasonable heuristic is to apply the criteria often used for AIC (Burnham and Anderson, 1998; Dixon, 2008). One such criterion is to require that $|\Delta WAIC| = 2$ in order to claim that there is positive evidence in favour of one model over the other, though other criteria may also be used. We evaluate the operating characteristics of this decision rule are evaluated in comparison to BIC and AIC through our simulation studies.

We emphasize that BIC and WAIC, although both Bayesian in their formulation, are constructed with different utilities in mind. BIC is based on the notion of posterior model probabilities and the Bayes factor, whereas WAIC is an estimate of the expected out-of-sample-prediction error. Given the differing utilities it is certainly possible that

the two criteria will disagree, and we view the approaches as complimentary. Detailed comparisons are made in the next section. In addition to computing the posterior distribution for a given model, our R function BinBayes.R will also compute the BIC and WAIC for any of the models in Table 4.2 or for models associated with the two-factor designs discussed above.

4.3 Simulation Studies

In order to evaluate the methodology described in the previous section we conducted two simulation studies.The studies examined the type I error and the power of specific decision rules based on the Bernoulli mixed models and the BIC or the WAIC for evaluating the effect of experimental conditions. These were compared with the type I error and the power of the standard aggregating approach after aggregating over items. Although we make our comparisons based on frequentist criteria, we note that it is not uncommon to evaluate Bayesian methods using such criteria (see e.g. Carlin and Louis, 2008). In the non-Bayesian context, simulation studies somewhat similar to those presented here are conducted in Dixon (2008). For the standard aggregating approach, we let $\tilde{Y}_{ik}$ denote the proportion of accurate responses at the i^{th} condition in the experiment for subject k. This is obtained by averaging that subject's binary response variables Y_{ijk} over the items at condition i. The statistical model considered in this case is

$$\tilde{Y}_{ik} = \mu_i + b_k + \epsilon_{ik}, \;\; \epsilon_{ik} \overset{iid}{\sim} N(0, \sigma^2_\epsilon), \;\; i = 1, \ldots, I; \;\; k = 1, \ldots, K \tag{4.1}$$

where $b_k \overset{iid}{\sim} N(0, \sigma^2_b)$ are subject-specific random effects that are assumed independent of the model errors ϵ_{ik}, and μ_i is the fixed effect of the experimental condition. To evaluate the effect of the experimental condition, this model was compared with the simpler model that assumes no effect of the experimental condition and thus has $\mu_i = \mu, i = 1, \ldots, I$. For the standard aggregating approach we made this evaluation within the Bayesian paradigm and computed the Bayes factor using the BayesFactor package (Rouder et al., 2012) in the R programming language. The BayesFactor package implements Monte Carlo techniques for computing the Bayes factor for a class of Gaussian error models, of which (4.1) is a special case. Finally, we compared the three Bayesian approaches for model selection (the first Bayesian approach is the standard aggregating approach under a Bayesian framework) with the standard AIC

criterion applied to logistic mixed models.

In each study, we simulated binary response accuracy data of the type depicted in Table 4.1, with $K = 72$ subjects, $J = 120$ items, and $I = 4$ experimental conditions manipulated as a repeated-measures factor. In the first study we simulate data where the effect of the experimental condition does not depend on items or subjects, and in the second study this effect is not held constant, but instead varied across the items. In each setting, 1,500 simulation replicates were used to estimate each point of the power curve for comparing the null and alternative models as the effect size varied. To create power curves when the BIC approximation to the BF was used, our decision rule was to reject the null model whenever $BF > \exp(1)$ (which occurs when $\Delta BIC = BIC_{null} - BIC_{alt} > 2$) and the same rule was used for both the WAIC and AIC. For the standard item aggregated approach using the BF, the null model was rejected whenever the Bayes factor favouring the alternative model was greater than $BF > \exp(1)$. Additionally, we also created power curves for each of the four methods where the decision rule was chosen for each so that the type I error rate was fixed at 0.05. Fixing the type I error rate for all four methods to have the same value has the advantage of making the power curves more directly comparable.

4.3.1 Simulation Study I

We generated the simulated data using the logistic mixed model LM_F which contains a fixed effect for the experimental condition. The simulated data could also be generated from a probit mixed model; however, we simulated from the logistic model as it is more commonly used in practice. In this case, the effect of the experimental condition does not depend on subjects or items. The fixed effect is represented by α_i, and we set $\alpha_1 = \alpha_2 = \alpha_3 = 0$ and $\alpha_4 = C$, where C ranged over a series of values from $C = 0$ to $C = 1.5$. This particular pattern of fixed effects where the effect of only a single condition varies was chosen so that the results are easier to summarize and visualize (i.e., in a two-dimensional plot). The intercept was set at $\beta_0 = 3.22$ corresponding to a baseline accuracy rate of 96% (representative of performance in speeded word identification experiments, for example). We also considered and will summarize later results from simulations in which performance was not near ceiling or floor. The variance components were set according to $\sigma_b = 1.045$ for the standard deviation of the subject random effects, and we considered three values for the standard deviation of the item random effects $\sigma_a = 1.5, 3$, or 5. These choices for the

simulation parameter values are guided by the model estimates obtained from the single-factor real-data example analyzed in the next section.

For each value of the fixed effect C and the item standard deviation σ_a, we simulated $1,500$ datasets, and for each dataset we fit model LM_F which contains the fixed effect of the experimental condition, as well as model LM_0 which has no effect for the experimental condition (the null model). The BIC approximation to the BF, AIC, and WAIC for both models were computed, and in addition, the standard model Eq. (4.1) was applied after aggregating over items, and the Bayes factor comparing the models with and without a fixed effect for condition was computed. In all four cases, the decision rules described in the previous section were applied to create power curves for the different model selection criteria.

The results of the simulation study are presented in Figure 4.2 which shows the result of assessing the significance of the experimental manipulation by comparing models with and without fixed effects for each approach. In the case where the decision rules were set so that the type I error rate was fixed at 0.05 for all three Bayesian approaches as well as the AIC (second column of Figure 4.2), a clear pattern emerges. When the between-item variability is at its lowest level, with $\sigma_a = 1.5$, all four approaches have power curves that are virtually identical. In this case there appears to be no advantage to applying the Bernoulli mixed model over the standard aggregating approach. However, as the between-item variability increases, the Bernoulli mixed models with BIC approximation to BF, AIC, or WAIC outperform the standard aggregating approach with BF and have uniformly higher power. Interestingly, the BIC approximation to the BF, AIC, and WAIC have identical power curves when they are calibrated to have the same type I error rate.

In practice, outside of a simulation study, it may not be possible to calibrate the decision rule so as to ensure a specific value for the type I error rate. The first column of Figure 4.2 depicts the power curves when the decision rules $\Delta AIC > 2$, $\Delta WAIC > 2$, and $BF > \exp(1)$ are applied. In this case, it must be understood that a comparison of the power curves is not an 'apples-to-apples' comparison as the type I error rates are not the same. For all values of σ_a the type I error rate for the Bernoulli mixed model with $BF > \exp(1)$ ($\Delta BIC > 2$) is 0. This produces a conservative rule that has uniformly lower power than the Bernoulli mixed model with $\Delta AIC > 2$ and $\Delta WAIC > 2$. As a tradeoff, the latter have a higher type I error rate and we note that the power curves for WAIC and AIC are virtually identical in this case. The standard aggregating approach with rejection of the null when $BF > \exp(1)$ also has

a type I error rate of 0 in all cases. It should be noted that although both the BIC approximation to the BF and the standard item aggregated approach with BF have a type I error rate of 0, the corresponding adjusted power curves differ due to the fact that the two approaches require different decision rules in order to fix the type I error rates at 0.05.

The power curve associated with the standard aggregating approach is always below that of the WAIC and AIC, and is above the power curve of the BIC approximation to the BF when σ_a is set to its lowest value of 1.5. As the value of σ_a increases, the power of the Bernoulli mixed model with BIC approximation to the BF begins to improve relative to the standard aggregating approach. In the case where $\sigma_a = 5$ both the BIC approximation to the BF and standard aggregating approach have a type I error rate of 0, but the power of the Bernoulli mixed model with BIC apprximation to the BF is generally higher than that of the standard aggregating approach, particularly for higher values of the effect size C.

4.3.2 Simulation Study II

We next consider the situation where the effect of the experimental condition varies across the items and where the objective is again to evaluate the data for the existence of a condition effect. We generated data sets from the model LM_{R_i} which contains random effects for both subjects and items, a fixed effect for the experimental condition, and a random effect representing the interaction between experimental condition and items. The parameter values for the simulation were again guided by model estimates obtained from the single-factor dataset considered in the next section. For simulating data, we set the intercept to be $\beta_0 = 3.21$ corresponding to a baseline accuracy rate of 96%, the variance components for the subject and item random effects were set based on $\sigma_b = 1.04$ and $\sigma_a = 0.44$ respectively, with these values being based on estimates obtained from fitting the model to the real data example discussed in the next section. The fixed effect for the experimental condition was again set based on $\alpha_1 = \alpha_2 = \alpha_3 = 0$ and $\alpha_4 = C$, and we considered three possible values, $C = 0, 0.25, 0.5$. The effect of the experimental condition varied across items through the random effect $(\alpha a)_{ij}^{(R)} \overset{iid}{\sim} N(0, \sigma_{\alpha a}^2)$ and the magnitude of this variability depended on the parameter $\sigma_{\alpha a}$, which we varied over a series of values ranging from 0 to 2. The values of C and $\sigma_{\alpha a}$ were varied factorially so that overall there can be a random effect even without a fixed effect ($C = 0$) and there can be a fixed effect

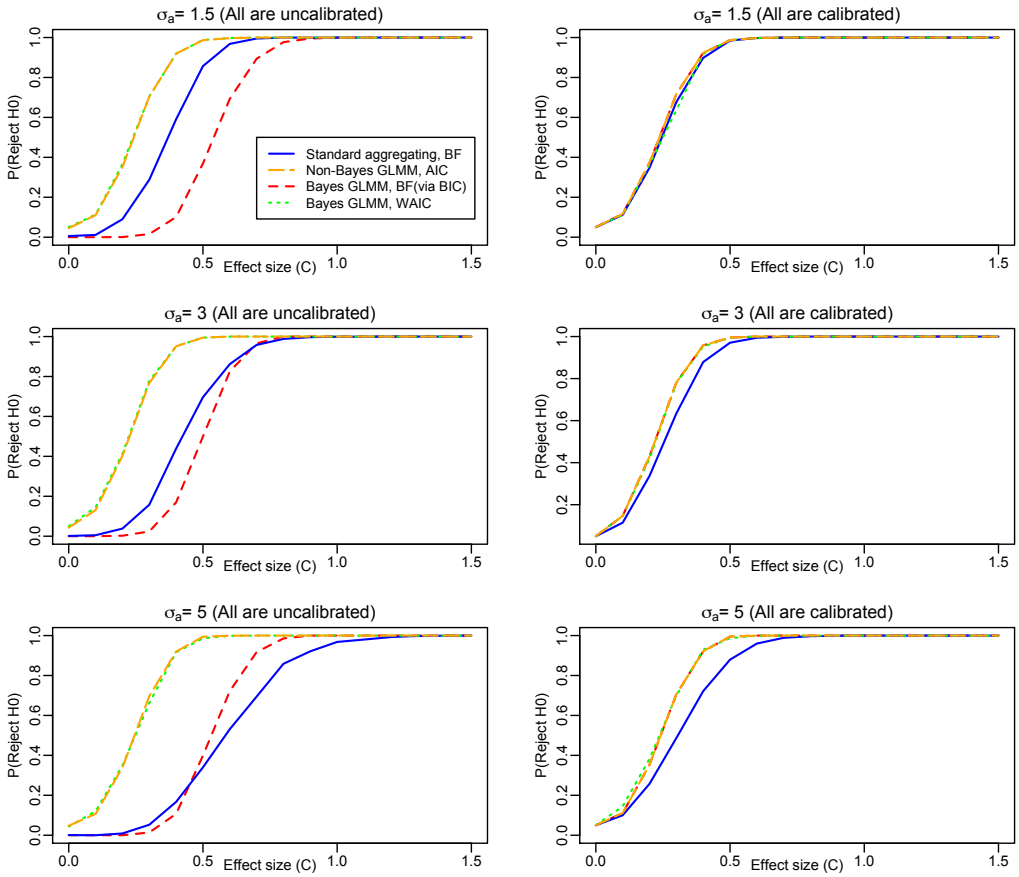


Figure 4.2: Results from simulation study I. The left column corresponds to the decision rules $\Delta BIC > 2$ (for Bayes GLMM, BF via BIC), $\Delta AIC > 2$ (for non-Bayes GLMM, AIC), $\Delta WAIC > 2$ (for Bayes GLMM, WAIC), and $BF >$ exp(1) (for standard aggregating BF), whereas in the right column the decision rules were chosen to ensure that all three methods had a type I error rate of 0.05. The rows correspond to different values of the between item variability $\sigma_a = 1.5, 3, 5$. Values of C represent the strength of the effect of the experimental conditions.

without a random effect ($\sigma_{\alpha a} = 0$). When both $C = 0$ and $\sigma_{\alpha a} = 0$ there is no effect present.

For each value of C and $\sigma_{\alpha a}$ we again simulated $1,500$ datasets, and for each dataset we fit the model LM_{R_i} which contains an effect for the experimental condition that varies across the items, as well as the model LM_0 which has no effect for the experimental condition (the null model). The comparison of models LM_{R_i} and LM_0 then corresponds to an overall test for an effect of the experimental condition, where this effect can be either random and varying across items, fixed, or both. After aggregating over items, the standard aggregating approach was applied as in the previous section. We note that the standard aggregating approach is not sufficiently flexible to allow the condition effect to depend on items, and so as before, the effect of experimental condition was evaluated through the model Eq. (4.1) and the Bayes factor comparing the models with and without a fixed effect for condition was computed. In all cases the decision rules used in the first simulation study were applied again.

The results are depicted in Figure 4.3. In this case the effect of the experimental condition is represented by both the fixed effect C and the standard deviation of the random condition-by-item interaction $\sigma_{\alpha a}$. We clarify that in this figure the power curves vary on the x-axis by $\sigma_{\alpha a}$; whereas, they vary on the x-axis by C in Figure 4.2. As we are testing for an overall effect of the experimental conditions, a type I error can occur only when $C = 0$ and $\sigma_{\alpha a} = 0$, that is, when both the fixed and random components of the experimental effect are absent, which is possible only in the left-most data point in the top panels of Figure 4.3, where both $C = 0$ and $\sigma_{\alpha a} = 0$. The second column of Figure 4.3 compares the power curves when the decision rules were chosen so as to fix the probability of a type I error at 0.05. In these cases, when $C = 0$ (so that the fixed effect of condition is absent but the random condition-by-item effect is not) the Bernoulli mixed model with any of the BIC approximation to the BF, AIC, or WAIC outperforms the standard aggregating approach rather significantly with respect to power. In this case the latter approach does not contain parameters in the model for a random effect, and can only detect this through the fixed effect for condition, though it is interesting to note that the model is able to do this once $\sigma_{\alpha a}$ is sufficiently large. This phenomenon can be thought of as a type of false alarm, because the approach detects the existence of a fixed effect when in fact the real effect that is present is a random effect. It also appears that WAIC has higher power than both the AIC and BIC approximation to the BF in general. *Thus*

we see here an advantage in terms of power for the fully Bayesian approach compared with the non-Bayesian approach when the same generalized linear mixed models are applied and it is only the model selection criterion that varies.

When C is increased so that $C = 0.25$ (so that there is now both a fixed effect of condition as well as a random condition-by-item effect) the power of all four approaches increases; however, the Bernoulli mixed model still generally outperforms the standard aggregating approach with a fairly large difference in power at most values of $\sigma_{\alpha a}$. Considering the same model comparisons based on generalized linear mixed models (LM_{R_i} versus LM_0) *we again see that WAIC has higher power than both AIC and BIC approximation to the BF in this case.* When $C = 0.5$ the fixed effect is now sufficiently large that all four methods have very high power to detect an effect of experimental condition and the power curves are generally flat as $\sigma_{\alpha a}$ varies.

Turning to the first column of Figure 4.3 where the methods are not calibrated to have the same type I error rate, we see that the Bernoulli mixed model with decision rule $BF > \exp(1)$ ($\Delta BIC > 2$) results in an extremely conservative procedure, where the type I error is 0 but the power is generally lower (particularly with $\sigma_{\alpha a}$ small) than the other procedures, with the exception of AIC which is also very conservative and has a power curve matching that of the BIC approximation to the BF when $C = 0.25$ and $C = 0.5$. Interestingly, when $C = 0$ (so that the fixed effect of the experimental condition is absent but the random condition-by-item effect is not) the power curve for the AIC is higher than that of both the BIC approximation to the BF and the standard aggregating approach but is uniformly lower than that of the WAIC. The power of the standard aggregating approach improves as the values of C increases, but its power is uniformly dominated by the Bernoulli mixed model with decision rule $\Delta WAIC > 2$ in all cases. We emphasize again that this is not an 'apples-to-apples' comparison as the two approaches are not calibrated to have the same type I error rate. The type I error rate of both approaches is indicated in the left-most data point in the first row and first column of Figure 4.3 when $\sigma_{\alpha a} = 0$, and we note that the Bernoulli mixed model with decision rule $\Delta WAIC > 2$ has a slightly higher type I error than the other approaches (Dashed line in Figure 4.3 when $C = 0$). Nevertheless, its power to detect an experimental effect is much higher for most values of $\sigma_{\alpha a}$, and this is particularly evident within a neighbourhood of $\sigma_{\alpha a} = 0.5$. *Thus it is interesting to note that fully Bayesian WAIC has higher power than AIC when the same logistic mixed models are being compared.*

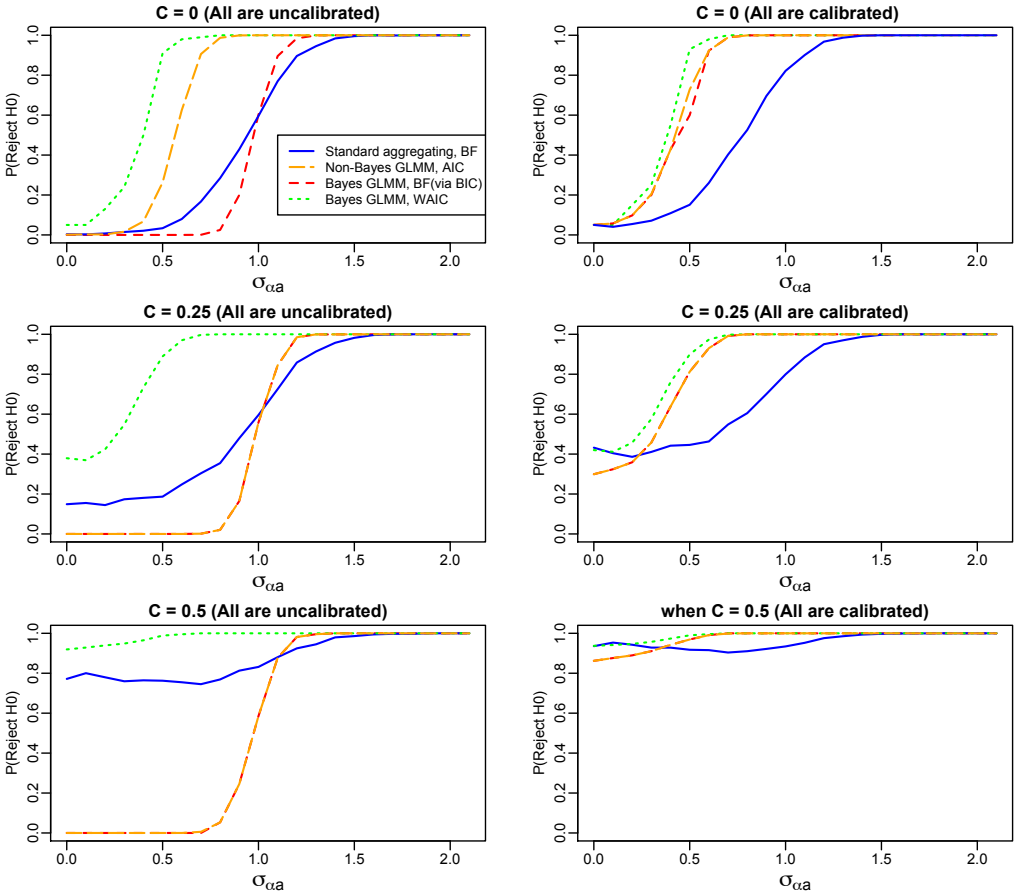


Figure 4.3: Results from simulation study II. The left column corresponds to the decision rules $\Delta BIC > 2$ (for Bayes GLMM, BF via BIC), $\Delta AIC > 2$ (for non-Bayes GLMM, AIC), $\Delta WAIC > 2$ (for Bayes GLMM, WAIC), and $BF > \exp(1)$ (for standard aggregating BF), whereas in the right column the decision rules were chosen to ensure that all four methods had a type I error rate of 0.05. Note that a type I error can occur only when $C = 0$ and $\sigma_{\alpha a} = 0$ (since otherwise the null is false) so the calibration of the type I error rates is based on the $C = 0$ and $\sigma_{\alpha a} = 0$ case for all three panels in the right column.

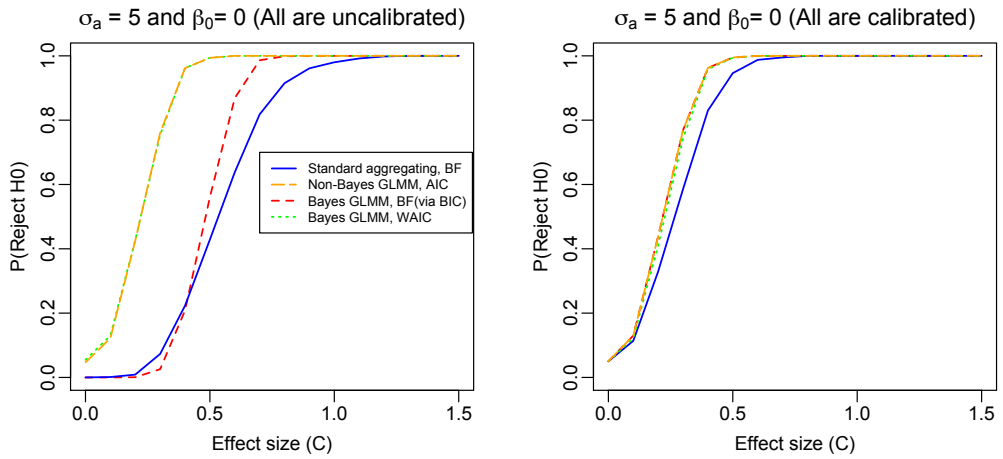


Figure 4.4: Results from simulation study I with $\beta_0 = 0$ corresponding to a baseline accuracy of 50%. The left panel corresponds to the decision rules $\Delta BIC > 2$ (for Bayes GLMM, BF via BIC), $\Delta AIC > 2$ (for non-Bayes GLMM, AIC), $\Delta WAIC > 2$ (for Bayes GLMM, WAIC), and $BF >$ exp(1) (for standard aggregating BF), whereas in the right panel the decision rules were chosen to ensure that all three methods had a type I error rate of 0.05. These settings correspond to the third row of Figure 4.2 where both simulation studies had a 96% baseline accuracy rate.

In both simulation studies we have assumed that the true value of the intercept in the logistic model is $\beta_0 = 3.22$. This is a fairly large value corresponding to a baseline accuracy rate of approximately 96%. We have also conducted additional simulation studies where the baseline accuracy rate was not so extreme, based on setting the true value to $\beta_0 = 0$, corresponding to a baseline accuracy rate of approximately 50%. The results for this more moderate accuracy rate are depicted for study I in Figure 4.4 which corresponds to the third row of Figure 4.2, and for study II in Figure 4.5 which corresponds to the first row of Figure 4.3.

Figure 4.4 and Figure 4.5 indicate that the comparison of the power curves at a baseline accuracy rate of 50% yields results that are quite similar to those already presented where the baseline accuracy rate was 96%. The primary difference is that the relative performance of the standard aggregating approach appears to drop in the case where the baseline accuracy rate is 50%. A baseline rate of 50% was also the value considered in Dixon (2008) where the intercept was taken to be zero in the simulations that evaluated generalized linear mixed models.

4.4 Example Application: Single-Factor Design

We now present an example application illustrating the use of Bernoulli mixed models for the analysis of response accuracy for a single-factor design. The analysis presented

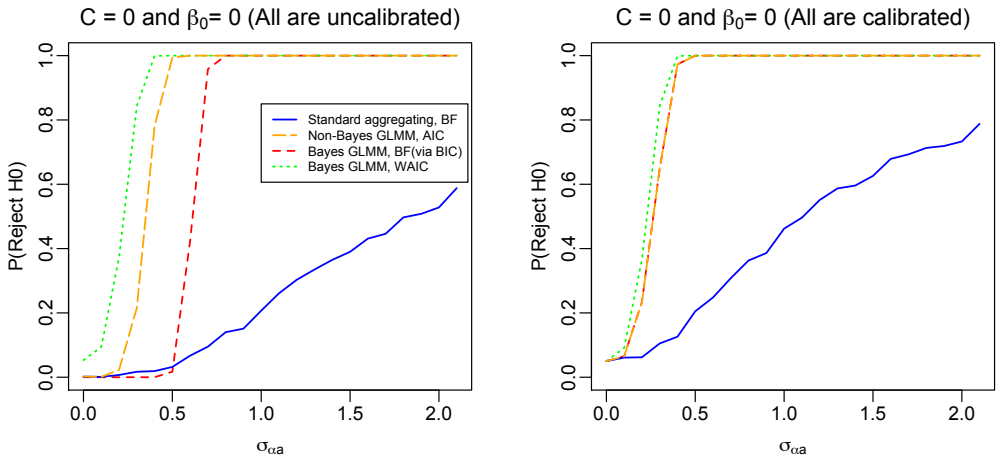


Figure 4.5: Results from simulation study II with $\beta_0 = 0$. The left figure corresponds to the decision rules $\Delta BIC > 2$ (for Bayes GLMM, BF via BIC), $\Delta AIC > 2$ (for non-Bayes GLMM, AIC), $\Delta WAIC > 2$ (for Bayes GLMM, WAIC), and $BF >$ exp(1) (for standard aggregating BF), whereas in the right figure the decision rules were chosen to ensure that all four methods had a type I error rate of 0.05. These settings correspond to the first row of Figure 4.3 where the baseline accuracy rate is 96%.

here can be reproduced using the software, data, and examples provided via `https://v2south.github.io/BinBayes/`. The data considered here were taken from a study that investigated the influence of a semantic context on the identification of printed words shown either under clear (high contrast) or degraded (low contrast) conditions. The semantic context consisted of a prime word presented in advance of the target item. On critical trials, the target item was a word and on other trials the target was a nonword. The task was a lexical-decision procedure where the subject was instructed to classify the target on each trial as a word or a nonword, and the response was either accurate or inaccurate. Our interest was confined to trials with word targets. The prime word was either semantically related or unrelated to the target word (e.g., granite-STONE vs. attack-FLOWER), and the target word was presented either in clear (high contrast) or degraded (low contrast) form. Combining these two factors produced four conditions: related-clear (RC), unrelated-clear (UC), related-degraded (RD), unrelated-degraded (UD). The two factors were treated as a single factor with four levels for this example. For the current analysis, the accuracy of the response was the dependent measure.

The study comprised $K = 72$ subjects, $I = 4$ conditions, and $J = 120$ items, and the total number of binary observations was $KJ = 8,640$. The overall rate of response accuracy was 95.4%. We took the UD (unrelated-degraded) level of the experimental condition as the baseline condition and fit each of the ten Bernoulli mixed models listed in Table 4.2, and the resulting WAIC and BIC scores were obtained for each model. The model comparisons are presented in Table 3.

According to WAIC, the optimal model is the logistic model with random subject and item effects, and where the effect of the experimental condition depends on items. According to BIC the optimal model is the logistic model with random subject and item effects, and where the effect of condition is fixed and does not depend on items. Taken together both criteria point to the existence of an effect for the experimental condition; however, the fixed condition effect has the highest (approximate) posterior model probability (BIC), whereas the model with random condition effects depending on item is expected to make the best out-of-sample predictions (WAIC). Both model selection criteria taken together seem to provide evidence in support of the logistic link as opposed to the probit link. This is primarily the case with BIC as the WAIC scores are more neutral towards the link function but do show some support in favour of the logistic link.

The BIC scores can be used to compute an approximate Bayes factor for comparing

Table 4.3: The BIC and WAIC values for each of the ten binomial mixed models presented in Table 4.2 after application to the study data. Note: the lowest (i.e., best) scores are in bold.

Model	**Link**	**Condition Effect**	**WAIC**	**BIC**
LM_0	Logit	Null	2827	2928
LM_F	Logit	Condition: Fixed	2803	**2925**
LM_{R_s}	Logit	Condition*Subject: Random	2804	3010
LM_{R_i}	Logit	Condition*Item: Random	**2800**	2997
$LM_{R_{s,i}}$	Logit	Condition*Subject + Condition*Item: Random	2802	3078
PM_0	Probit	Null	2827	2935
PM_F	Probit	Condition: Fixed	2803	2934
PM_{R_s}	Probit	Condition*Subject: Random	2806	3015
PM_{R_i}	Probit	Condition*Item: Random	2803	3007
$PM_{R_{s,i}}$	Probit	Condition*Subject + Condition*Item: Random	2806	3088

models. For example, comparing LM_0 (null condition) versus LM_F (fixed condition) we obtain a Bayes factor of

$$BF \approx \exp\{(BIC(LM_0) - BIC(LM_F))/2\} = 4.48$$

which indicates substantial evidence in favour of the model with a fixed condition effect when compared to the model with no condition effect.

Because it was chosen as the optimal model by the WAIC, we summarize the posterior distribution of model LM_{R_i} in more detail. Hence, the effect of experimental condition varies across items. The condition UD (unrelated-degraded) is taken as the baseline condition so that the item-dependent condition effects associated with the remaining three conditions are then interpreted relative to UD.

The posterior distributions for the item-dependent condition effects represent the information obtained about these effects from the data, the model, and Bayes rule. These are depicted as box-plots in Figures 4.6, 4.7, and 4.8, for the conditions UC, RD, and RC, respectively. These plots relate to the best linear unbiased predictors (BLUPs) plots for standard generalized linear mixed models. Overall, we see that all three conditions increase the probability of response accuracy relative to the baseline UD. This increase is roughly the same for RD and UC; however, it appears that RC leads to higher accuracy rates compared to all other conditions. Examining all three conditions, we see that the variability in the condition effects across the items is not substantial; however, it does appear that some items have relatively lower or higher effects, particularly in the RD condition. Examining Figure 4.8, there is also one item in the RC condition that appears to have a substantially lower effect relative to all other items. Thus, it seems that the variability in the effect size across items picked up by the WAIC is driven primarily by just a few items. Our analysis enables the user to identify such items through the posterior distribution.

Bayesian interval estimates can be obtained by selecting those values of highest posterior probability density including the mode. This is sometimes called the 95% highest posterior density interval when the posterior probability associated with the interval is 0.95. With respect to the variance components for the random effects, the between-item effect standard deviation σ_a has a 95% highest posterior density (HPD) interval of $(0.863, 1.261)$, the between-subject effect standard deviation σ_b has a 95% HPD interval of $(0.280, 0.625)$, and the standard deviation for the condition-by-item random effects $\sigma_{\alpha a}$ has a 95% HPD interval of $(0.109, 0.587)$.

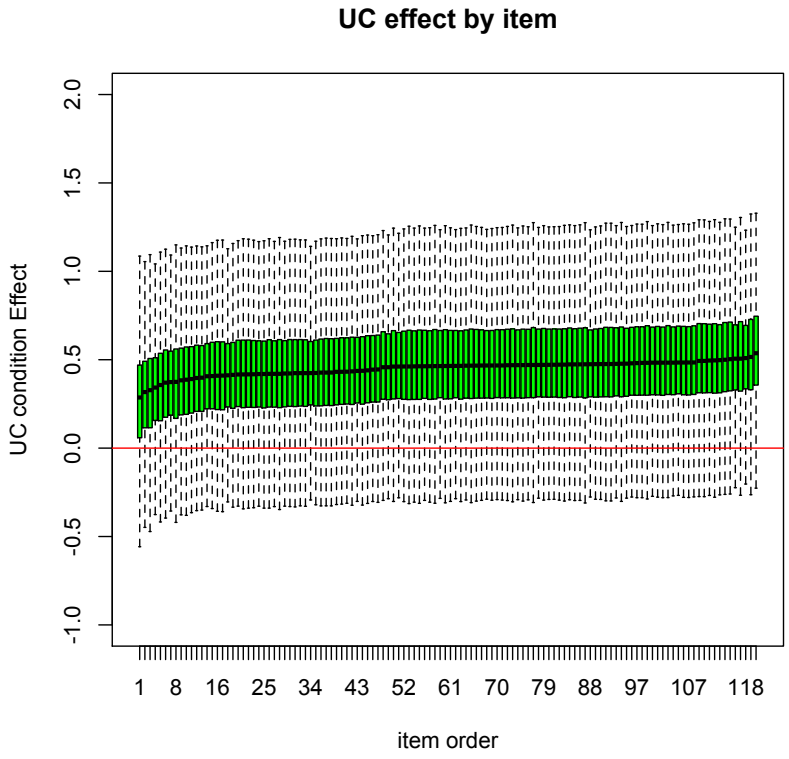


Figure 4.6: The posterior distributions for the effects of the unrelated-clear (UC) condition on the probability of response accuracy. In this case there is a separate effect for each of the 120 items used in the experiment and the figure depicts the posterior distribution for condition UC across the items as a boxplot summarizing Markov chain Monte Carlo samples drawn from the posterior distribution. Items are ordered according to their estimated effect sizes. The effects depicted are on the logit scale. The black line represents the posterior median for each item, the green region represents the set of values between the first and third quartile of the posterior distribution, and the dotted bars extend out to the extreme values.

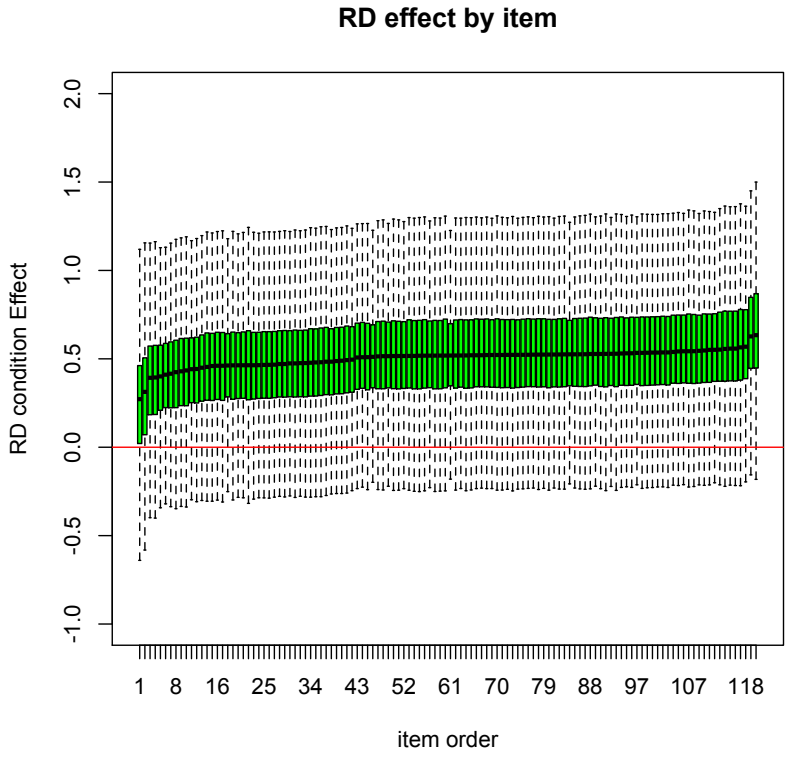


Figure 4.7: The posterior distributions for the effects of the related-degraded (RD) condition on the probability of response accuracy. In this case there is a separate effect for each of the 120 items used in the experiment and the figure depicts the posterior distribution for condition RD across the items as a boxplot summarizing Markov chain Monte Carlo samples drawn from the posterior distribution. Items are ordered according to their estimated effect sizes. The effects are depicted on the logit scale. The black line represents the posterior median for each item, the green region represents the set of values between the first and third quartile of the posterior distribution, and the dotted bars extend out to the extreme values.

4.5 Conclusions and Recommendations

We have introduced a collection of Bernoulli mixed models that can be used for the analysis of accuracy studies in the Bayesian framework. The set of models repre-

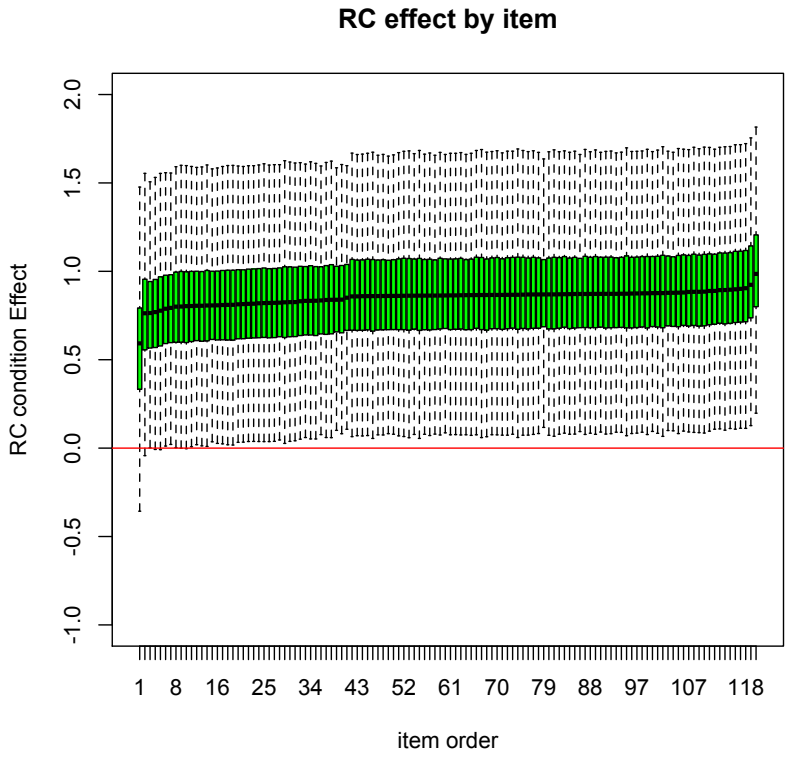


Figure 4.8: The posterior distributions for the effects of the related-clear (RC) condition on the probability of response accuracy. In this case there is a separate effect for each of the 120 items used in the experiment and the figure depicts the posterior distribution for condition RC across the items as a boxplot summarizing Markov chain Monte Carlo samples drawn from the posterior distribution. Items are ordered according to their estimated effect sizes. The effects are depicted on the logit scale. The black line represents the posterior median for each item, the green region represents the set of values between the first and third quartile of the posterior distribution, and the dotted bars extend out to the extreme values.

sents a number of different assumptions for the effect of the experimental condition on accuracy. These assumptions range from a null model to a model where the experimental effect varies across both items and subjects. The models we consider are based on random effects that are assumed normally distributed. Although one may consider generalized linear mixed models where the random effects are not normally distributed (see e.g. Nathoo and Ghosh, 2013) it is generally the case that inference in generalized linear mixed models is fairly robust to misspecification of the random effects distribution (e.g. McCulloch and Neuhaus, 2011).

For the models we have considered here, we have generally assumed the presence of a fixed effect in any model that contains a corresponding random effect. Jaeger (2008) has suggested the alternative of assessing the significance of effects by comparing to a model without the fixed effect but, critically, with the subject- or item-based random slopes for the fixed effects. This comparison makes sure that the assessment of significance (e.g., through model comparison) does not confound the effect of the fixed effect predictor with the variance captured by the random slopes for that predictor. The extent to which such confounding can cause problems for the model comparisons considered here is unclear but presents a potentially interesting avenue for investigation.

To compare possible models, we investigated AIC and both the BIC as a large-sample approximation to the Bayes factor and the WAIC as a large-sample approximation to Bayesian cross-validation. We compared these to the standard, item-aggregated approach using simulation studies. An alternative approach for model selection that we did not consider in our simulation studies is that of nested comparison of models via a chi-square over differences in model deviances. This approach, although arguably a current standard for model comparison across generalized linear mixed models (in the field), is less flexible than the approaches we considered as it requires the models being compared to be nested and is not appropriate for testing random effects because the asymptotic chi-square distribution under the null is not valid when the null hypothesis lies on the boundary of the parameter space (as it does when testing random effects where the null corresponds to setting a variance component to zero, see e.g., Lin, 1997).

Overall, we recommend the use of the WAIC as it appears to have power that is higher or at least as high as the BIC approximation to the BF, AIC, and the standard aggregating approach with BF when applied with industry-standard decision rules. The BIC approximation to the BF can be used alongside WAIC and the results treated

as complimentary when Bayes factors are of interest. While not typical, there could be specific cases where the choice of random effects structure and link function could be driven by theoretical considerations. In these cases, these considerations could be used to narrow the class of possible models and then combined with Bayesian model selection.

For the simulation settings considered here, we found that the performance of the Bernoulli mixed model approach improved relative to the standard repeated-measures approach as both the between-item variability σ_a (study I) and the item-condition variability $\sigma_{\alpha a}$ (study II) increased. In the latter case, the observed difference in performance when comparing our approach to the standard item-aggregated approach was rather substantial for a fairly large range of values for $\sigma_{\alpha a}$. In general, we see that as the variability across items increases, the application of the proposed approach becomes increasingly more valuable and likely to detect effects, relative to the standard aggregating approach to evaluating the effects of independent variables on accuracy.

With respect to comparisons between AIC and the fully Bayesian WAIC, when the same logistic mixed models were applied the two approaches had identical power curves when testing for fixed effects of the experimental condition in study I. Interestingly, the WAIC had higher power than the AIC when testing for random effects of the experimental conditions in study II. As with any simulation study, the conclusions drawn are specific to the particular simulation settings adopted. The conclusions we have drawn regarding the power of the Bayesian approach relative to its competitors can be guaranteed to hold only under the conditions assumed for the simulation studies. Nevertheless, these initial findings are very instructive.

We note that the WAIC is based on the priors included in our model specifications, whereas the BIC approximation to the BF is based on the unit information prior. The differences in performance seen when comparing these approaches is due in part to the differences in priors. It was clear that the BIC when used with the decision rule based on $\Delta BIC = 2$ results in a conservative approach. Aside from comparisons based on standard decision rules, we also made comparisons after controlling for type I error, in which case the performance of the BIC approximation to the BF improved with respect to power. Practically speaking, outside of a simulation study, it is currently not possible to control the type I error of a decision rule when using information criteria for model comparison; however, we are currently investigating an approach for doing so based on the parametric bootstrap as an avenue for future work.

We again note that there is a large body of work that debates the pros and cons

of a Bayesian analysis. The use of Bayesian approaches requires researchers to digest new ideas that can be conceptually difficult. Users must take the time to acquire the necessary background in order to use Bayesian methods appropriately and we refer the interested reader to the introductory textbook of Gelman et al. (2014). We also note that our approach, which requires MCMC sampling, is more computationally intensive than either the standard aggregating approach or the generalized linear mixed model fit by maximum likelihood with AIC used for model comparisons. We have demonstrated that WAIC has power that is higher or at least as high as AIC under certain settings, and we believe that this justifies the additional computation required. In addition, the advantages of a Bayesian analysis in the ability to construct posterior distributions for parameters of interest, and the availability of a user-friendly software implementation for both single-factor and two-factor designs make our approach an exciting alternative to standard item aggregation approaches and contemporary mixed logit/probit procedures for the analysis of repeated-measures accuracy studies.

Chapter 5

Conclusion

In this thesis we have developed new statistical methods for the analysis of neuroimaging data and repeated measures designs in cognitive science.

In Chapter 2, we describe a new approach for solving the inverse problem associated with combined EEG and MEG data that appears to result in some improvements over and above the original MSM mixture model. The new approach can be applied to situations where the MEG and EEG data are collected simultaneously and also to situations where the data are collected sequentially in a situation where the data mimic a simultaneous recording paradigm. Our methodology makes the very strong assumption that neural activity is generated by a small number of latent states. Extending our approach to accommodate a large number of sources is an open problem. In addition, our analysis of the residuals in our application suggests that developing a model with errors having more flexible tail behaviour may be useful. In fact most model-based solutions to the electromagnetic inverse problem assume Gaussian errors at the first level. Relaxing this assumption may be a useful avenue for future work.

We note that our approach, as with most in the literature, is only applicable to data from a single subject. Typically the algorithm is applied to each subject's data separately to solve the inverse problem in a first stage and then a group level analysis is run on the estimated sources at the second stage. An interesting challenge and likely useful development is an extension of our model to handle multi-subject data where the inverse problem is solved simultaneously for a group of subjects. Here, the solution still varies from subject to subject but the group data are combined to better estimate hyper-parameters under the assumption that different subjects share the same hyper-parameters. This sort of model would be well suited to divide and

conquer computing strategies such as the consensus Monte Carlo algorithm (Scott et al., 2016).

In Chapter 3, we describe a spatial multi-task regression model for relating genetic data to multivariate imaging phenotypes. The model inherits some of the features of the methodology developed by Greenlaw et al. (2017) which introduced Bayesian shrinkage priors for imaging genetics, but it enhances this through the use of a bivariate conditional autoregressive spatial model, which allows for both spatial correlation as well as bilateral correlation across brain hemispheres. Ours is one of the first explicitly spatial hierarchical models for imaging genetics, and thus our model formulation breaks new ground. In addition, our model is also one of the first to explicitly model dependence in corresponding measures across the two brain hemispheres. A limitation of the proposed methodology is that it is currently only applicable to settings where the dimension of the imaging phenotype is fairly moderate corresponding to a fairly coarse brain atlas. Extending the approach to accommodate imaging phenotypes of much higher dimension is an avenue for future work that will prove extremely challenging. Again, some sort of divide and conquer strategy that involves partitioning the image into smaller images and applying some variant of the current algorithm on each element of the partition might be a useful strategy. In addition, moving beyond a linear model to a model that allows for complex interactions between genetic variables should prove useful.

In Chapter 4, we have introduced a collection of Bernoulli mixed models that can be used for the analysis of accuracy studies in the Bayesian framework. The set of models represents a number of different assumptions for the effect of the experimental condition on accuracy. These assumptions range from a null model to a model where the experimental effect varies across both items and subjects. Overall, we recommend the use of the WAIC as it appears to have power that is higher or at least as high as the BIC approximation to the BF, AIC, and the standard aggregating approach with BF when applied with industry-standard decision rules. The BIC approximation to the BF can be used alongside WAIC and the results treated as complimentary when Bayes factors are of interest. A potentially important extension of the proposed work would be the development of Bayesian within-subject inference for logistic and probit mixed regression models. Recent work by Nathoo, Kilshaw and Masson (2018) has developed within-subject Bayesian inference for the Gaussian mixed model. Extending this idea

to the logistic and probit mixed models may be useful and such a formulation will be considered for future work.

Appendix A

Appendix for Chapter 2

A.1 Data Transformations and Supplementary Figures

Given the original MEG and EEG data collected at the sensor arrays

$$\tilde{\boldsymbol{M}}(t) = (\tilde{M}_1(t), \tilde{M}_2(t), ..., \tilde{M}_{n_M}(t))', \;\; t = 1, \ldots, T$$
$$\tilde{\boldsymbol{E}}(t) = (\tilde{E}_1(t), \tilde{E}_2(t), ..., \tilde{E}_{n_E}(t))', \;\; t = 1, \ldots, T,$$

we let $\tilde{\boldsymbol{M}}$ and $\tilde{\boldsymbol{E}}$ denote the corresponding data matrices of dimensions $n_M \times T$ and $n_E \times T$, respectively. Similarly, we let $\tilde{\boldsymbol{X}}_E$ and $\tilde{\boldsymbol{X}}_M$ denote the $n_E \times P$ and $n_M \times P$ EEG and MEG forward operators computed based on Maxwell's equations, the sensor array locations, the pre-specified locations on the cortex, and other assumptions on the conductivity of the fluids and tissues within the head.

Our model assumes that these data have been transformed as suggested by Henson et al. (2009b) as follows:

$$\boldsymbol{M} = \frac{\tilde{\boldsymbol{M}}}{\sqrt{\frac{1}{n_M} tr(\tilde{\boldsymbol{M}} \tilde{\boldsymbol{M}}^T)}}, \qquad \boldsymbol{E} = \frac{\tilde{\boldsymbol{E}}}{\sqrt{\frac{1}{n_E} tr(\tilde{\boldsymbol{E}} \tilde{\boldsymbol{E}}^T)}},$$
$$\boldsymbol{X}_M = \frac{\tilde{\boldsymbol{X}}_M}{\sqrt{\frac{1}{n_M} tr(\tilde{\boldsymbol{X}}_M \tilde{\boldsymbol{X}}_M^T)}}, \qquad \boldsymbol{X}_E = \frac{\tilde{\boldsymbol{X}}_E}{\sqrt{\frac{1}{n_E} tr(\tilde{\boldsymbol{X}}_E \tilde{\boldsymbol{X}}_E^T)}},$$

and the joint model is then specified for the transformed data as described in Section 2 of the Chapter 2.

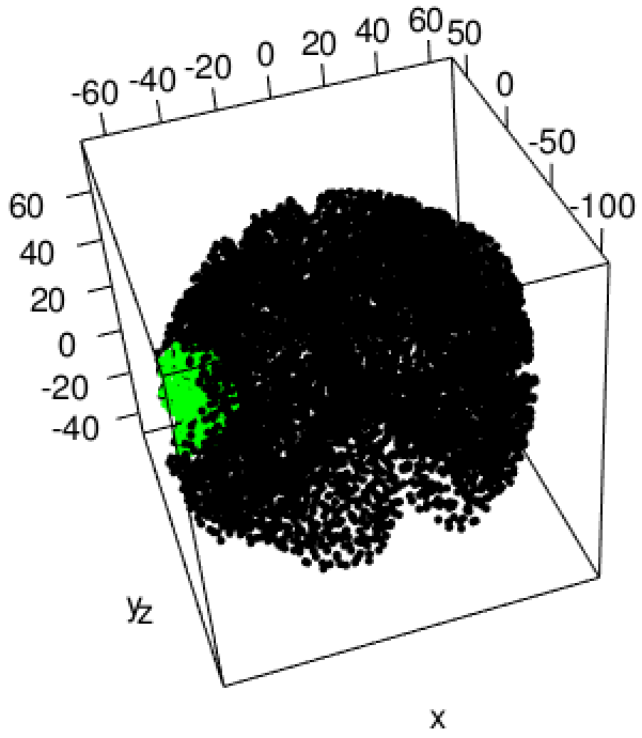


Figure A.1: The true allocation of cortical locations to mixture components in the case where $K_{true} = 2$. Locations coloured green correspond to active locations while the other locations are inactive. The signal at active locations is based on the Gaussian curve depicted in Figure A.2 panel (a). In total, there are 8,196 cortical locations used in this example.

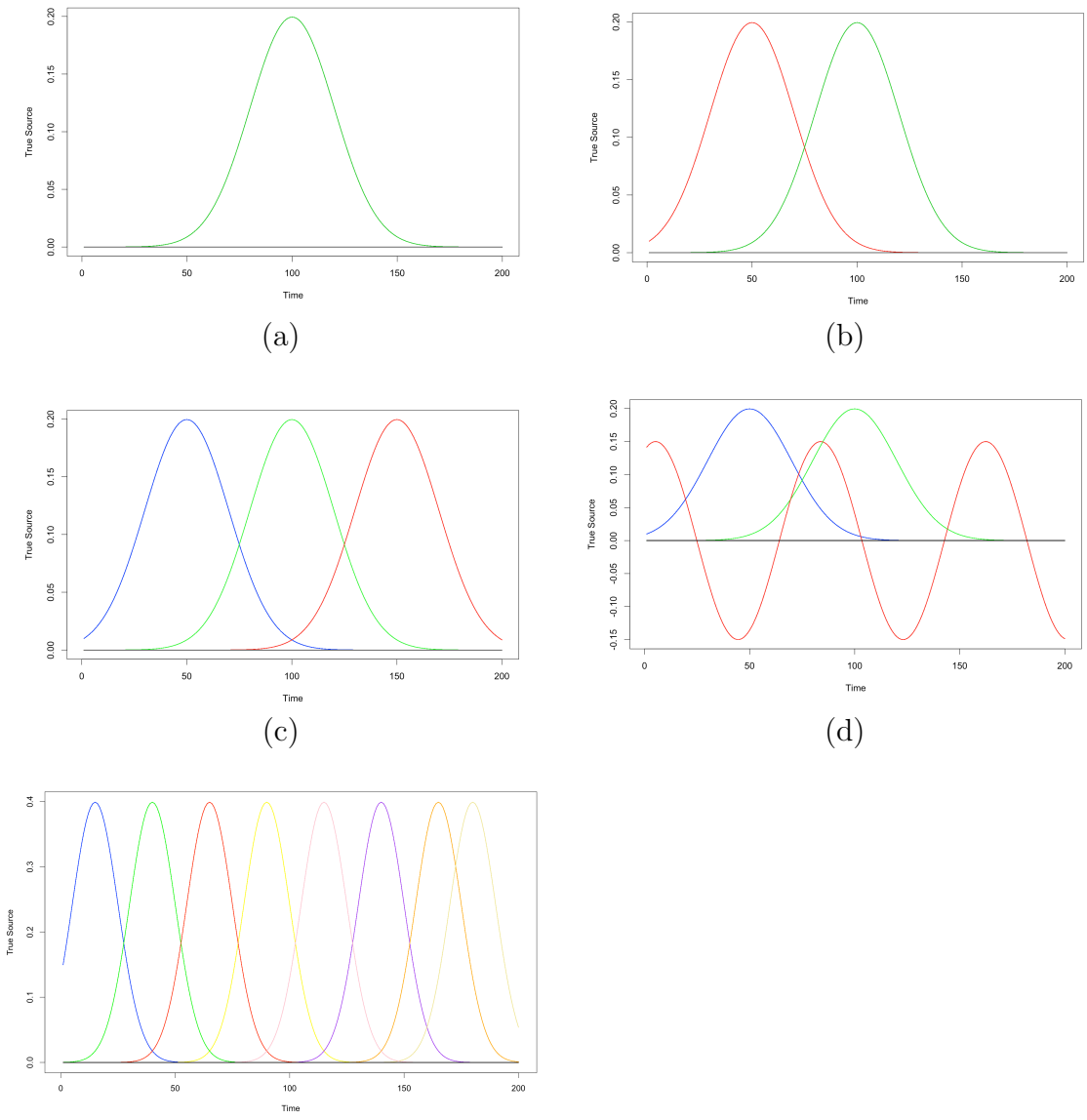


Figure A.2: Figure (a) to (e) represent the true signal $S_j(t)$ used in in each of the distinct active and inactive regions in the second part of simulation study of Section 2.4, where the mixture components ($K = 1, 3, 4, 9$) are less well separated.

A.2 Derivations for ICM algorithm

The ICM algorithm requires the full conditional distribution for each model parameter. The mode of this distribution is then used in the update step for that parameter. Here we derive the required full conditional distributions.

The full conditional distributions for σ_M^2 and σ_E^2 are obtained as follows:

$$
\begin{aligned}
P(\sigma_M^2|Rest) &\propto \prod_{t=1}^{T}[P(M(t)|S(t),\sigma_M^2)] \times P(\sigma_M^2) \\
&\propto \prod_{t=1}^{T}\Big[|\sigma_M^2 H_M|^{-1/2}\exp\Big\{-\frac{1}{2}\big(M(t)-X_M S(t)\big)^T(\sigma_M^2 H_M)^{-1}\big(M(t)-X_M S(t)\big)\Big\}\Big] \\
&\times (\sigma_M^2)^{-(a_M+1)}\exp(-b_M/\sigma_M^2) \\
&\propto (\sigma_M^2)^{-TN_M/2}\exp\Big\{\sum_{t=1}^{T}-\frac{1}{2}\big(M(t)-X_M S(t)\big)^T(\sigma_M^2 H_M)^{-1}\big(M(t)-X_M S(t)\big)\Big\} \\
&\times (\sigma_M^2)^{-(a_M+1)}\exp(-b_M/\sigma_M^2) \\
&\propto (\sigma_M^2)^{-(a_M+\frac{TN_M}{2}+1)}\exp\Big\{\sum_{t=1}^{T}-\frac{1}{2\sigma_M^2}\big(M(t)-X_M S(t)\big)^T H_M^{-1}\big(M(t)-X_M S(t)\big)-\frac{b_M}{\sigma_M^2}\Big\} \\
&\propto (\sigma_M^2)^{-(a_M+\frac{TN_M}{2}+1)}\exp\Big\{\frac{-1}{\sigma_M^2}\Big(\sum_{t=1}^{T}\frac{1}{2}\big(M(t)-X_M S(t)\big)^T H_M^{-1}\big(M(t)-X_M S(t)\big)+b_M\Big)\Big\}
\end{aligned}
$$

Therefore, the full conditional for σ_M^2 is an Inverse-Gamma distribution with

$$
\begin{aligned}
a_M^* &= a_M + \frac{TN_M}{2} \\
b_M^* &= \sum_{t=1}^{T}\frac{1}{2}\big(M(t)-X_M S(t)\big)^T H_M^{-1}\big(M(t)-X_M S(t)\big)+b_M
\end{aligned}
$$

Similarly, we can get the full conditional for σ_E^2 as:

$$
\begin{aligned}
P(\sigma_E^2|Rest) &\propto \prod_{t=1}^{T}[P(E(t)|S(t),\sigma_E^2)] \times P(\sigma_E^2) \\
&\propto \prod_{t=1}^{T}\Big[|\sigma_E^2 H_E|^{-1/2}\exp\Big\{-\frac{1}{2}\big(E(t)-X_E S(t)\big)^T(\sigma_E^2 H_E)^{-1}\big(E(t)-X_E S(t)\big)\Big\}\Big] \\
&\times (\sigma_E^2)^{-(a_E+1)}\exp(-b_E/\sigma_E^2) \\
&\propto (\sigma_E^2)^{-TN_E/2}\exp\Big\{\sum_{t=1}^{T}-\frac{1}{2}\big(E(t)-X_E S(t)\big)^T(\sigma_E^2 H_E)^{-1}\big(E(t)-X_E S(t)\big)\Big\} \\
&\times (\sigma_E^2)^{-(a_E+1)}\exp(-b_E/\sigma_E^2) \\
&\propto (\sigma_E^2)^{-(a_E+\frac{TN_E}{2}+1)}\exp\Big\{\sum_{t=1}^{T}-\frac{1}{2\sigma_E^2}\big(E(t)-X_E S(t)\big)^T H_E^{-1}\big(E(t)-X_E S(t)\big)-\frac{b_E}{\sigma_E^2}\Big\} \\
&\propto (\sigma_E^2)^{-(a_E+\frac{TN_E}{2}+1)}\exp\Big\{\frac{-1}{\sigma_E^2}\Big(\sum_{t=1}^{T}\frac{1}{2}\big(E(t)-X_E S(t)\big)^T H_E^{-1}\big(E(t)-X_E S(t)\big)+b_E\Big)\Big\}
\end{aligned}
$$

Therefore, the full conditional for σ_E^2 is an Inverse-Gamma distribution with

$$
\begin{aligned}
a_E^* &= a_E + \frac{TN_E}{2} \\
b_E^* &= \sum_{t=1}^{T}\frac{1}{2}\big(E(t)-X_E S(t)\big)^T H_E^{-1}\big(E(t)-X_E S(t)\big)+b_E
\end{aligned}
$$

The full conditional for σ_a^2 is obtained as follows:

$$
\begin{aligned}
P(\sigma_a^2|Rest) &\propto \prod_{t=2}^{T}\Big[P(\boldsymbol{\mu}^A(t)|\boldsymbol{\mu}^A(t-1),\boldsymbol{A},\sigma_a^2)\Big]\times P(\sigma_a^2) \\
&\propto \prod_{t=2}^{T}\Big[|\sigma_a^2\boldsymbol{I}|^{-1/2}\exp\big\{-\frac{1}{2}(\boldsymbol{\mu}^A(t)-\boldsymbol{A}\boldsymbol{\mu}^A(t-1))^T(\sigma_a^2\boldsymbol{I})^{-1}(\boldsymbol{\mu}^A(t)-\boldsymbol{A}\boldsymbol{\mu}^A(t-1))\big\}\Big] \\
&\times (\sigma_a^2)^{-(a_a+1)}\exp(-b_a/\sigma_a^2) \\
&\propto (\sigma_a^2)^{-\frac{(T-1)(K-1)}{2}}\exp\Big\{\sum_{t=2}^{T}-\frac{1}{2}(\boldsymbol{\mu}^A(t)-\boldsymbol{A}\boldsymbol{\mu}^A(t-1))^T(\sigma_a^2\boldsymbol{I})^{-1}(\boldsymbol{\mu}^A(t)-\boldsymbol{A}\boldsymbol{\mu}^A(t-1))\Big\} \\
&\times (\sigma_a^2)^{-(a_a+1)}\exp(-b_a/\sigma_a^2) \\
&\propto (\sigma_a^2)^{-\frac{(T-1)(K-1)}{2}}\exp\Big\{\frac{1}{\sigma_a^2}\Big(-\frac{1}{2}\sum_{t=2}^{T}(\boldsymbol{\mu}^A(t)-\boldsymbol{A}\boldsymbol{\mu}^A(t-1))^T(\boldsymbol{\mu}^A(t)-\boldsymbol{A}\boldsymbol{\mu}^A(t-1))\Big)\Big\} \\
&\times (\sigma_a^2)^{-(a_a+1)}\exp(-b_a/\sigma_a^2) \\
&\propto (\sigma_a^2)^{-(a_a+\frac{(T-1)(K-1)}{2}+1)} \\
&\times \exp\Big\{\frac{-1}{\sigma_a^2}\Big(\sum_{t=2}^{T}\frac{1}{2}(\boldsymbol{\mu}^A(t)-\boldsymbol{A}\boldsymbol{\mu}^A(t-1))^T\boldsymbol{I}^{-1}(\boldsymbol{\mu}^A(t)-\boldsymbol{A}\boldsymbol{\mu}^A(t-1))+b_a\Big)\Big\}
\end{aligned}
$$

Therefore, the full conditional distribution for σ_a^2 is still Inverse-Gamma distribution with new parameters as:

$$a_a^* = a_a + \frac{(T-1)(K-1)}{2}$$
$$b_a^* = \sum_{t=2}^{T} \frac{1}{2}(\boldsymbol{\mu}^A(t) - \boldsymbol{A}\boldsymbol{\mu}^A(t-1))^T(\boldsymbol{\mu}^A(t) - \boldsymbol{A}\boldsymbol{\mu}^A(t-1)) + b_a$$

For the matrix $\boldsymbol{A}$, which describes the connectivity between states, we will transform it into a vector via $vec(A)$, we have:

$$\begin{aligned}\boldsymbol{A}\boldsymbol{\mu}^A(t-1) &= vec(\boldsymbol{A}\boldsymbol{\mu}^A(t-1)) \\ &= \Big(\boldsymbol{\mu}^A(t-1)^T \otimes \boldsymbol{I}_{k-1}\Big) vec(\boldsymbol{A}) \\ &= \boldsymbol{K}\boldsymbol{r}_{\boldsymbol{t}} \times vec(\boldsymbol{A}) \\ &\text{where } \boldsymbol{K}\boldsymbol{r}_{\boldsymbol{t}} = \Big(\boldsymbol{\mu}^A(t-1)^T \otimes \boldsymbol{I}_{k-1}\Big)\end{aligned}$$

Then the full conditional could be obtained as:

$$\begin{aligned}P(vec(\boldsymbol{A})|Rest) &\propto \prod_{i=1}^{k-1}\prod_{j=1}^{k-1} P(A_{ij}|\sigma_A^2) \times \prod_{t=2}^{T}\Big[P(\boldsymbol{\mu}^A(t)|\boldsymbol{\mu}^A(t-1), \boldsymbol{A}, \sigma_a^2)\Big] \\ &\propto MVN_{(k-1)^2}(vec(\boldsymbol{A}); \boldsymbol{0}, \sigma_A^2 \boldsymbol{I}_{(k-1)^2}) \\ &\times \prod_{t=2}^{T} \exp\Bigg(-\frac{1}{2\sigma_a^2}\Big(\boldsymbol{\mu}^A(t) - \big(\boldsymbol{\mu}^A(t-1)^T \otimes \boldsymbol{I}_{k-1}\big) vec(\boldsymbol{A})\Big)^T \\ &\Big(\boldsymbol{\mu}^A(t) - \big(\boldsymbol{\mu}^A(t-1)^T \otimes \boldsymbol{I}_{k-1}\big) vec(\boldsymbol{A})\Big)\Bigg) \\ &\propto MVN_{(k-1)^2}(vec(\boldsymbol{A}); \boldsymbol{0}, \sigma_A^2 \boldsymbol{I}_{(k-1)^2}) \times \prod_{t=2}^{T} \exp\Bigg(-\frac{1}{2\sigma_a^2}\Big(\boldsymbol{\mu}^A(t) - \boldsymbol{K}\boldsymbol{r}_{\boldsymbol{t}} \times vec(\boldsymbol{A})\Big)^T \\ &\Big(\boldsymbol{\mu}^A(t) - \boldsymbol{K}\boldsymbol{r}_{\boldsymbol{t}} \times vec(\boldsymbol{A})\Big)\Bigg) \\ &\propto \exp\Bigg(-\frac{1}{2\sigma_A^2} vec(\boldsymbol{A})^T vec(\boldsymbol{A}) + \frac{1}{\sigma_a^2}\Big(\sum_{t=2}^{T} \boldsymbol{\mu}^A(t)^T \boldsymbol{K}\boldsymbol{r}_{\boldsymbol{t}}\Big) \times vec(\boldsymbol{A}) \\ &- \frac{1}{2\sigma_a^2} vec(\boldsymbol{A})^T \Big(\sum_{t=2}^{T} \boldsymbol{K}\boldsymbol{r}_{\boldsymbol{t}}{}^T \boldsymbol{K}\boldsymbol{r}_{\boldsymbol{t}}\Big) vec(\boldsymbol{A})\Bigg) \\ &\propto \exp\Bigg(-\frac{1}{2}\big(vec(\boldsymbol{A}) - \boldsymbol{V}_1\big)^T \boldsymbol{C}_1 \big(vec(\boldsymbol{A}) - \boldsymbol{V}_1\big)\Bigg)\end{aligned}$$

Therefore full conditional distribution for $vec(\boldsymbol{A})$ is $MVN_{(K-1)^2}(\boldsymbol{V}_1, \boldsymbol{C}_1^{-1})$, where

$$\begin{aligned}
\boldsymbol{C}_1 &= \frac{1}{\sigma_A^2}\boldsymbol{I}_{(k-1)^2} + \frac{1}{\sigma_a^2}\Big(\sum_{t=2}^{T} \boldsymbol{Kr_t}^T \boldsymbol{Kr_t}\Big) \\
\boldsymbol{V}_1^T\boldsymbol{C}_1 &= \frac{1}{\sigma_a^2}\Big(\sum_{t=2}^{T} \boldsymbol{\mu}^A(t)^T \boldsymbol{Kr_t}\Big) \\
\boldsymbol{V}_1 &= \Big(\frac{1}{\sigma_a^2}\Big(\sum_{t=2}^{T} \boldsymbol{\mu}^A(t)^T \boldsymbol{Kr_t}\Big) \times \boldsymbol{C}_1^{-1}\Big)^T
\end{aligned}$$

For all the variance components α_l, the full conditional could be obtained together as:

$$\begin{aligned}
P(\alpha_1, \alpha_2, ..., \alpha_k|\text{Rest}) &\propto [\prod_{j=1}^{p}\prod_{t=1}^{T} P(S_j(t))|\boldsymbol{\mu}(t), \boldsymbol{\alpha}, \boldsymbol{Z_{v(j)}}] \times \prod_{l=1}^{k} P(\alpha_l|a_\alpha, b_\alpha) \\
&\propto [\prod_{j=1}^{p}\prod_{t=1}^{T}\prod_{l=1}^{k} N(S_j(t); \mu_l(t), \alpha_l)^{Z_{v(j)l}}] \times \prod_{l=1}^{k} IG(\alpha_l; a_\alpha, b_\alpha) \\
&\propto \prod_{l=1}^{k}\Big[\prod_{j=1}^{p}\prod_{t=1}^{T}\big(N(S_j(t); \mu_l(t), \alpha_l)^{Z_{v(j)l}}\big)\Big] \times IG(\alpha_l; a_\alpha, b_\alpha) \\
&\propto \prod_{l=1}^{k}\Bigg[\prod_{j=1}^{p}\prod_{t=1}^{T}\Big(\alpha_l^{-\frac{1}{2}}\exp(-\frac{(S_j(t)-\mu_l(t))^2}{2\alpha_l})\Big)^{Z_{v(j)l}} \times (\alpha_l)^{-(a_\alpha+1)}\exp(-b_\alpha/\alpha_l)\Bigg] \\
&\propto \prod_{l=1}^{k}\Bigg[\prod_{j=1}^{p}\prod_{t=1}^{T}\Big(\alpha_l^{-\frac{1}{2}}\exp(-\frac{(S_j(t)-\mu_l(t))^2}{2\alpha_l})\Big)^{Z_{v(j)l}} \times (\alpha_l)^{-(a_\alpha+1)}\exp(-b_\alpha/\alpha_l)\Bigg] \\
&\propto \prod_{l=1}^{k}\Bigg[\prod_{j=1}^{p}\prod_{t=1}^{T}\Big(\alpha_l^{-\frac{Z_{v(j)l}}{2}}\exp(-\frac{Z_{v(j)l}(S_j(t)-\mu_l(t))^2}{2\alpha_l})\Big) \\
&\quad \times (\alpha_l)^{-(a_\alpha+1)}\exp(-b_\alpha/\alpha_l)\Bigg] \\
&\propto \prod_{l=1}^{k}\Bigg[\Big(\alpha_l^{-\frac{T\sum_{j=1}^{p} Z_{v(j)l}}{2}}\exp(-\frac{\sum_{j=1}^{p}\sum_{t=1}^{T} Z_{v(j)l}(S_j(t)-\mu_l(t))^2}{2\alpha_l})\Big) \\
&\quad \times (\alpha_l)^{-(a_\alpha+1)}\exp(-b_\alpha/\alpha_l)\Bigg] \\
&\propto \prod_{l=1}^{k}\Bigg[\alpha_l^{-\big(\frac{T\sum_{j=1}^{p} Z_{v(j)l}}{2}+a_\alpha+1\big)}\exp\Big(-\frac{1}{\alpha_l}\Big(\frac{\sum_{j=1}^{p}\sum_{t=1}^{T} Z_{v(j)l}(S_j(t)-\mu_l(t))^2}{2}+b_\alpha\Big)\Big)\Bigg]
\end{aligned}$$

Therefore, we can see that for each α_l, the individual full conditional is still a Inverse-

Gamma distribution with new parameters as:

$$a^*_{\alpha_l} = \frac{T\sum_{j=1}^{p} Z_{v(j)l}}{2} + a_\alpha$$
$$b^*_{\alpha_l} = \frac{\sum_{j=1}^{p}\sum_{t=1}^{T} Z_{v(j)l}(S_j(t)-\mu_l(t))^2}{2} + b_\alpha$$

Also, full conditional for $\boldsymbol{\mu}(t)$ for when $t = 1$:

$$\begin{aligned}
P(\boldsymbol{\mu}(1)|\text{Rest}) &\propto \prod_{j=1}^{p} P(S_j(1))|\boldsymbol{\mu}^A(1), \boldsymbol{\alpha}, \boldsymbol{Z_{v(j)}}) \times P(\boldsymbol{\mu}^A(2)|\boldsymbol{\mu}^A(1), \boldsymbol{A}, \sigma_a^2) \times P(\boldsymbol{\mu}^A(1)|\sigma^2_{\mu_1}) \\
&\propto [\prod_{j=1}^{p}\prod_{l=1}^{k} N(S_j(1); \mu_l(1), \alpha_l)^{Z_{v(j)l}}] \times [MVN_{k-1}(\boldsymbol{\mu}^A(2); \boldsymbol{A}\mu^A(1), \sigma_a^2\boldsymbol{I})] \\
&\times MVN_{k-1}(\boldsymbol{\mu}^A(1); \boldsymbol{0}, \sigma^2_{\mu_1}\boldsymbol{I}) \\
&\propto \Big[\prod_{j=1}^{p}\prod_{l=1}^{k} \exp\Big(-\frac{Z_{v(j)l}(S_j(1)-\mu_l(1))^2}{2\alpha_l}\Big)\Big] \times \exp\Big(-\frac{1}{2\sigma_a^2}(\boldsymbol{\mu}^A(2) - \boldsymbol{A\mu}^A(1))^T \\
&(\boldsymbol{\mu}^A(2) - \boldsymbol{A\mu}^A(1))\Big) \times \exp\Big(-\frac{1}{2\sigma^2_{\mu_1}}(\boldsymbol{\mu}^A(1))^T(\boldsymbol{\mu}^A(1))\Big) \\
&\propto \exp\Bigg(-\frac{1}{2}\sum_{j=1}^{p}\sum_{l=1}^{k}\frac{Z_{v(j)l}}{\alpha_l}(S_j(1)-\mu_l(1))^2 - \frac{1}{2\sigma_a^2}(\boldsymbol{\mu}^A(2) - \boldsymbol{A\mu}^A(1))^T(\boldsymbol{\mu}^A(2) - \boldsymbol{A\mu}^A(1)) \\
&- \frac{1}{2\sigma^2_{\mu_1}}(\boldsymbol{\mu}^A(1))^T(\boldsymbol{\mu}^A(1))\Bigg) \\
&\propto \exp\Bigg(-\frac{1}{2}\sum_{j=1}^{p}\big(S_j(1)\vec{I}_{k-1} - \boldsymbol{\mu}^A(1)\big)^T \text{Diag}(\frac{Z_{v(j)l}}{\alpha_l}, l = 2, ..., k)\big(S_j(1)\vec{I}_{k-1} - \boldsymbol{\mu}^A(1)\big) \\
&- \frac{1}{2\sigma_a^2}(\boldsymbol{\mu}^A(2) - \boldsymbol{A\mu}^A(1))^T(\boldsymbol{\mu}^A(2) - \boldsymbol{A\mu}^A(1)) - \frac{1}{2\sigma^2_{\mu_1}}(\boldsymbol{\mu}^A(1))^T(\boldsymbol{\mu}^A(1))\Bigg) \\
&\text{where } \vec{I}_{k-1} \text{ is a all ones vector with length k-1. } \boldsymbol{D}_j = \text{Diag}(\frac{Z_{v(j)l}}{\alpha_l}, l = 2, ..., k) \\
P(\boldsymbol{\mu}(1)|\text{Rest}) &\propto \exp\Bigg(\sum_{j=1}^{p}\Big((S_j(1)\vec{I}_{k-1})^T\boldsymbol{D}_j\boldsymbol{\mu}^A(1) - \frac{1}{2}(\boldsymbol{\mu}^A(1))^T\boldsymbol{D}_j\boldsymbol{\mu}^A(1)\Big) + \frac{1}{\sigma_a^2}(\boldsymbol{\mu}^A(2))^T\boldsymbol{A\mu}^A(1)) \\
&- \frac{1}{2\sigma_a^2}(\boldsymbol{\mu}^A(1))^T\boldsymbol{A}^T\boldsymbol{A}(\boldsymbol{\mu}^A(1)) - \frac{1}{2\sigma^2_{\mu_1}}(\boldsymbol{\mu}^A(1))^T(\boldsymbol{\mu}^A(1))\Bigg) \\
&\propto \exp\Bigg(\Big(\sum_{j=1}^{p}(S_j(1)\vec{I}_{k-1})^T\boldsymbol{D}_j + \frac{1}{\sigma_a^2}(\boldsymbol{\mu}^A(2))^T\boldsymbol{A}\Big)\boldsymbol{\mu}^A(1) \\
&- \frac{1}{2}\boldsymbol{\mu}^A(1)^T\{\sum_{j=1}^{p} D_j + \frac{1}{\sigma_a^2}\boldsymbol{A^TA} + \frac{1}{\sigma^2_{\mu_1}}\boldsymbol{I}_{k-1}\}\boldsymbol{\mu}^A(1)\Bigg) \\
&\propto \exp\Bigg(-\frac{1}{2}\Big(\boldsymbol{\mu}^A(1) - \boldsymbol{M}_1\Big)^T\boldsymbol{B}_1\Big(\boldsymbol{\mu}^A(1) - \boldsymbol{M}_1\Big)\Bigg)
\end{aligned}$$

Therefore, the full conditional for $\boldsymbol{\mu}(1)$ is multivariate normal distribution $MVN_{k-1}(\boldsymbol{M}_1, \boldsymbol{B}_1^{-1})$ with parameters as:

$$\boldsymbol{B}_1 = \sum_{j=1}^{p} \boldsymbol{D}_j + \frac{1}{\sigma_a^2}\boldsymbol{A^T A} + \frac{1}{\sigma_{\mu_1}^2}\boldsymbol{I}_{k-1}$$

$$\boldsymbol{M}_1^T \boldsymbol{B}_1 = \sum_{j=1}^{p} (S_j(1)\vec{I}_{k-1})^T \boldsymbol{D}_j + \frac{1}{\sigma_a^2}(\boldsymbol{\mu}^A(2))^T \boldsymbol{A}$$

$$\boldsymbol{M}_1 = \left(\left(\sum_{j=1}^{p} (S_j(1)\vec{I}_{k-1})^T \boldsymbol{D}_j + \frac{1}{\sigma_a^2}(\boldsymbol{\mu}^A(2))^T \boldsymbol{A} \right) \times \boldsymbol{B}_1^{-1} \right)^T$$

When $1 < t < T$, the full condition is:

$$
\begin{aligned}
P(\boldsymbol{\mu}(t)|\text{Rest}) &\propto \prod_{j=1}^{p} P(S_j(t))|0, \boldsymbol{\mu}^A(t), \boldsymbol{\alpha}, \boldsymbol{Z_{v(j)}}) \times P(\boldsymbol{\mu}^A(t+1)|\boldsymbol{\mu}^A(t), \boldsymbol{A}, \sigma_a^2) \\
&\times P(\boldsymbol{\mu}^A(t)|\boldsymbol{\mu}^A(t-1), \boldsymbol{A}, \sigma_a^2) \\
&\propto [\prod_{j=1}^{p}\prod_{l=1}^{k} N(S_j(t); \mu_l(t), \alpha_l)^{Z_{v(j)l}}] \times [MVN_{k-1}(\boldsymbol{\mu}^A(t+1); \boldsymbol{A}\mu^A(t), \sigma_a^2\boldsymbol{I})] \\
&\times [MVN_{k-1}(\boldsymbol{\mu}^A(t); \boldsymbol{A}\mu^A(t-1), \sigma_a^2\boldsymbol{I})] \\
&\propto \Big[\prod_{j=1}^{p}\prod_{l=1}^{k} \exp\Big(-\frac{Z_{v(j)l}(S_j(t)-\mu_l(t))^2}{2\alpha_l}\Big)\Big] \times \exp\Big(-\frac{1}{2\sigma_a^2}(\boldsymbol{\mu}^A(t+1) - \boldsymbol{A}\boldsymbol{\mu}^A(t))^T \\
&(\boldsymbol{\mu}^A(t+1) - \boldsymbol{A}\boldsymbol{\mu}^A(t))\Big) \times \exp\Big(-\frac{1}{2\sigma_a^2}(\boldsymbol{\mu}^A(t) - \boldsymbol{A}\boldsymbol{\mu}^A(t-1))^T(\boldsymbol{\mu}^A(t) - \boldsymbol{A}\boldsymbol{\mu}^A(t-1))\Big) \\
&\propto \exp\Bigg(-\frac{1}{2}\sum_{j=1}^{p}\sum_{l=1}^{k}\frac{Z_{v(j)l}}{\alpha_l}(S_j(t)-\mu_l(t))^2 - \frac{1}{2\sigma_a^2}\big(\boldsymbol{\mu}^A(t+1) - \boldsymbol{A}\boldsymbol{\mu}^A(t)\big)^T\big(\boldsymbol{\mu}^A(t+1) \\
&- \boldsymbol{A}\boldsymbol{\mu}^A(t)\big) - \frac{1}{2\sigma_a^2}\big(\boldsymbol{\mu}^A(t) - \boldsymbol{A}\boldsymbol{\mu}^A(t-1)\big)^T\big(\boldsymbol{\mu}^A(t) - \boldsymbol{A}\boldsymbol{\mu}^A(t-1)\big)\Bigg) \\
&\propto \exp\Bigg(-\frac{1}{2}\sum_{j=1}^{p}\big(S_j(t)\vec{I}_{k-1} - \boldsymbol{\mu}^A(t)\big)^T \text{Diag}(\frac{Z_{v(j)l}}{\alpha_l}, l=2,...,k)\big(S_j(t)\vec{I}_{k-1} - \boldsymbol{\mu}^A(t)\big) \\
&- \frac{1}{2\sigma_a^2}\big(\boldsymbol{\mu}^A(t+1) - \boldsymbol{A}\boldsymbol{\mu}^A(t)\big)^T\big(\boldsymbol{\mu}^A(t+1) - \boldsymbol{A}\boldsymbol{\mu}^A(t)\big) \\
&- \frac{1}{2\sigma_a^2}\big(\boldsymbol{\mu}^A(t) - \boldsymbol{A}\boldsymbol{\mu}^A(t-1)\big)^T\big(\boldsymbol{\mu}^A(t) - \boldsymbol{A}\boldsymbol{\mu}^A(t-1)\big)\Bigg) \\
&\text{where } \vec{I}_{k-1} \text{ is a all ones vector with length k-1. } \boldsymbol{D}_j = \text{Diag}(\frac{Z_{v(j)l}}{\alpha_l}, l=2,...,k) \\
P(\boldsymbol{\mu}(t)|\text{Rest}) &\propto \exp\Bigg(\sum_{j=1}^{p}\Big((S_j(t)\vec{I}_{k-1})^T\boldsymbol{D}_j\boldsymbol{\mu}^A(t) - \frac{1}{2}(\boldsymbol{\mu}^A(t))^T\boldsymbol{D}_j\boldsymbol{\mu}^A(t)\Big) + \frac{1}{\sigma_a^2}(\boldsymbol{\mu}^A(t+1)^T\boldsymbol{A})\boldsymbol{\mu}^A(t) \\
&- \frac{1}{2\sigma_a^2}\boldsymbol{\mu}^A(t)^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\mu}^A(t) - \frac{1}{2\sigma_a^2}\boldsymbol{\mu}^A(t)^T\boldsymbol{\mu}^A(t) + \frac{1}{\sigma_a^2}(\boldsymbol{\mu}^A(t-1)^T\boldsymbol{A}^T)\boldsymbol{\mu}^A(t)\Bigg) \\
&\propto \exp\Bigg(-\frac{1}{2}\Big(\boldsymbol{\mu}^A(t) - \boldsymbol{M}_2\Big)^T\boldsymbol{B}_2\Big(\boldsymbol{\mu}^A(t) - \boldsymbol{M}_2\Big)\Bigg)
\end{aligned}
$$

Then, the full conditional distribution is a $MVN_{k-1}(\boldsymbol{M}_2, \boldsymbol{B}_2^{-1})$ as:

$$\begin{aligned}
\boldsymbol{B}_2 &= \sum_{j=1}^{p} \boldsymbol{D}_j + \frac{1}{\sigma_a^2}(\boldsymbol{A^T A} + \boldsymbol{I}_{k-1}) \\
\boldsymbol{M}_2^T \boldsymbol{B}_2 &= \sum_{j=1}^{p} (S_j(t)\vec{I}_{k-1})^T \boldsymbol{D}_j + \frac{1}{\sigma_a^2}(\boldsymbol{\mu}^A(t+1))^T \boldsymbol{A} + \frac{1}{\sigma_a^2}(\boldsymbol{\mu}^A(t-1)^T \boldsymbol{A}^T) \\
\boldsymbol{M}_2 &= \left(\Big(\sum_{j=1}^{p} (S_j(t)\vec{I}_{k-1})^T \boldsymbol{D}_j + \frac{1}{\sigma_a^2}(\boldsymbol{\mu}^A(t+1))^T \boldsymbol{A} + \frac{1}{\sigma_a^2}(\boldsymbol{\mu}^A(t-1)^T \boldsymbol{A}^T) \Big) \times \boldsymbol{B}_2^{-1} \right)^T
\end{aligned}$$

When $t = T$, the full conditional is:

$$\begin{aligned}
P(\boldsymbol{\mu}(T)|\text{Rest}) &\propto \prod_{j=1}^{p} P(S_j(T))|\boldsymbol{\mu}^A(T), \boldsymbol{\alpha}, \boldsymbol{Z_{v(j)}}) \times P(\boldsymbol{\mu}^A(T)|\boldsymbol{\mu}^A(T-1), \boldsymbol{A}, \sigma_a^2) \\
&\propto [\prod_{j=1}^{p}\prod_{l=1}^{k} N(S_j(T); \mu_l(T), \alpha_l)^{Z_{v(j)l}}] \times [MVN_{k-1}(\boldsymbol{\mu}^A(T); \boldsymbol{A}\mu^A(T-1), \sigma_a^2 \boldsymbol{I})] \\
&\propto \Big[\prod_{j=1}^{p}\prod_{l=1}^{k} \exp\Big(- \frac{Z_{v(j)l}(S_j(T) - \mu_l(T))^2}{2\alpha_l} \Big) \Big] \times \exp\Big(- \frac{1}{2\sigma_a^2}(\boldsymbol{\mu}^A(T) - \boldsymbol{A\mu}^A(T-1))^T \\
&\quad (\boldsymbol{\mu}^A(T) - \boldsymbol{A\mu}^A(T-1)) \Big) \\
&\propto \exp\left(- \frac{1}{2} \sum_{j=1}^{p}\sum_{l=1}^{k} \frac{Z_{v(j)l}}{\alpha_l}(S_j(T) - \mu_l(T))^2 - \frac{1}{2\sigma_a^2}\big(\boldsymbol{\mu}^A(T) - \boldsymbol{A\mu}^A(T-1)\big)^T \right. \\
&\quad \left. \big(\boldsymbol{\mu}^A(T) - \boldsymbol{A\mu}^A(T-1)\big) \right) \\
&\propto \exp\left(- \frac{1}{2} \sum_{j=1}^{p} \big(S_j(T)\vec{I}_{k-1} - \boldsymbol{\mu}^A(T)\big)^T \text{Diag}(\frac{Z_{v(j)l}}{\alpha_l}, l = 2, ..., k) \big(S_j(T)\vec{I}_{k-1} - \boldsymbol{\mu}^A(T)\big) \right. \\
&\quad \left. - \frac{1}{2\sigma_a^2}\big(\boldsymbol{\mu}^A(T) - \boldsymbol{A\mu}^A(T-1)\big)^T \big(\boldsymbol{\mu}^A(T) - \boldsymbol{A\mu}^A(T-1)\big) \right) \\
&\quad \text{where } \boldsymbol{D}_j = \text{Diag}(\frac{Z_{v(j)l}}{\alpha_l}, l = 2, ..., k) \\
P(\boldsymbol{\mu}(T)|\text{Rest}) &\propto \exp\left(\sum_{j=1}^{p} \Big((S_j(T)\vec{I}_{k-1})^T \boldsymbol{D}_j \boldsymbol{\mu}^A(T) - \frac{1}{2}(\boldsymbol{\mu}^A(T))^T \boldsymbol{D}_j \boldsymbol{\mu}^A(T) \Big) \right. \\
&\quad \left. + \frac{1}{\sigma_a^2}(\boldsymbol{\mu}^A(T-1)^T \boldsymbol{A}^T)\boldsymbol{\mu}^A(T) - \frac{1}{2\sigma_a^2}\boldsymbol{\mu}^A(T)^T \boldsymbol{\mu}^A(T) \right) \\
&\propto \exp\Big(- \frac{1}{2}\Big(\boldsymbol{\mu}^A(T) - \boldsymbol{M}_3\Big)^T \boldsymbol{B}_3 \Big(\boldsymbol{\mu}^A(T) - \boldsymbol{M}_3\Big) \Big)
\end{aligned}$$

Therefore, full condition distribution when $t = T$ is a $MVN_{K-1}(\boldsymbol{M}_3, \boldsymbol{B}_3^{-1})$ with:

$$\begin{aligned}
\boldsymbol{B}_3 &= \sum_{j=1}^{p} \boldsymbol{D}_j + \frac{1}{\sigma_a^2}\boldsymbol{I}_{k-1} \\
\boldsymbol{M}_3^T\boldsymbol{B}_3 &= \sum_{j=1}^{p} (S_j(T)\vec{I}_{k-1})^T\boldsymbol{D}_j + \frac{1}{\sigma_a^2}(\boldsymbol{\mu}^A(T-1)^T\boldsymbol{A}^T) \\
\boldsymbol{M}_3 &= \left(\Big(\sum_{j=1}^{p} (S_j(T)\vec{I}_{k-1})^T\boldsymbol{D}_j + \frac{1}{\sigma_a^2}(\boldsymbol{\mu}^A(T-1)^T\boldsymbol{A}^T)\Big) \times \boldsymbol{B}_3^{-1}\right)^T
\end{aligned}$$

Regarding to $S_j(t)$ for $t = 1, 2, ..., T$, the full conditional distribution could be obtained as:

$$\begin{aligned}
P(S_j(t)|\text{Rest}) &\propto \prod_{t=1}^{T}[P(E(t)|S(t), \sigma_E^2)P(M(t)|S(t), \sigma_M^2)] \times [\prod_{t=1}^{T} P(S_j(t))|\boldsymbol{\mu}^A(t), \boldsymbol{\alpha}, \boldsymbol{Z_{v(j)}}] \\
&\propto \prod_{t=1}^{T}\Bigg[\exp\Big\{-\frac{1}{2}\big(M(t) - X_M S(t)\big)^T(\sigma_M^2 H_M)^{-1}\big(M(t) - X_M S(t)\big)\Big\} \\
&\times \exp\Big\{-\frac{1}{2}\big(E(t) - X_E S(t)\big)^T(\sigma_E^2 H_E)^{-1}\big(E(t) - X_E S(t)\big)\Big\} \\
&\times \prod_{l=1}^{k}\exp\Big(-\frac{Z_{v(j)l}(S_j(t) - \mu_l(t))^2}{2\alpha_l}\Big)\Bigg] \\
&\propto \prod_{t=1}^{T}\exp\Bigg[-\frac{1}{2}\big(M(t) - X_M S(t)\big)^T(\sigma_M^2 H_M)^{-1}\big(M(t) - X_M S(t)\big) \\
&- \frac{1}{2}\big(E(t) - X_E S(t)\big)^T(\sigma_E^2 H_E)^{-1}\big(E(t) - X_E S(t)\big) - \frac{1}{2}\sum_{l=1}^{k}\frac{Z_{v(j)l}(S_j(t) - \mu_l(t))^2}{\alpha_l}\Bigg] \\
&\propto \prod_{t=1}^{T}\exp\Bigg[-\frac{1}{2\sigma_M^2}\big(M(t)^T H_M^{-1} M(t) - 2M(t)^T H_M^{-1} X_M S(t) + (X_M S(t))^T H_M^{-1} X_M S(t)\big) \\
&- \frac{1}{2\sigma_E^2}\big(E(t)^T H_E^{-1} E(t) - 2E(t)^T H_E^{-1} X_E S(t) + (X_E S(t))^T H_E^{-1} X_E S(t)\big) \\
&- \frac{1}{2}\sum_{l=1}^{k}\frac{Z_{v(j)l}(S_j(t)^2 - 2\mu_l(t)S_j(t) + \mu_l(t)^2)}{\alpha_l}\Bigg]
\end{aligned}$$

Let $X_M[, v]$ denote the vth column in the matrix Then, we can rewrite $X_M S(t)$ as:

$X_M S(t) = \sum_{v=1}^{p} X_M[, v]S_v(t)$. Then:

$$\begin{aligned}
P(S_j(t)|\text{Rest}) \propto \prod_{t=1}^{T}\exp\Bigg[&-\frac{1}{2\sigma_M^2}\Big(-2M(t)^T H_M^{-1}(\sum_{v=1}^{p} X_M[, v]S_v(t)) + \\
&(\sum_{v=1}^{p} X_M[, v]S_v(t))^T H_M^{-1}(\sum_{v=1}^{p} X_M[, v]S_v(t))\Big) - \frac{1}{2\sigma_E^2}\Big(-2E(t)^T H_E^{-1}(\sum_{v=1}^{p} X_E[, v]S_v(t)) \\
&+ ((\sum_{v=1}^{p} X_E[, v]S_v(t)))^T H_E^{-1}(\sum_{v=1}^{p} X_E[, v]S_v(t))\Big) - \frac{1}{2}\sum_{l=1}^{k}\frac{Z_{v(j)l}(S_j(t)^2 - 2\mu_l(t)S_j(t))}{\alpha_l}\Bigg]
\end{aligned}$$

We want to keep the terms that have $S_j(t)$:

$$
\begin{aligned}
P(S_j(t)|\text{Rest}) &\propto \prod_{t=1}^{T} \exp\Bigg[-\frac{1}{2\sigma_M^2}\Big(-2M(t)^T H_M^{-1} X_M[,j] S_j(t)) \\
&+ (\sum_{v\neq j} X_M[,v] S_v(t))^T H_M^{-1} (X_M[,j] S_j(t)) + (X_M[,j] S_j(t))^T H_M^{-1} (\sum_{v\neq j} X_M[,v] S_v(t)) \\
&+ (X_M[,j] S_j(t))^T H_M^{-1} (X_M[,j] S_j(t))\Big) - \frac{1}{2\sigma_E^2}\Big(-2E(t)^T H_E^{-1} X_E[,j] S_j(t)) \\
&+ (\sum_{v\neq j} X_E[,v] S_v(t))^T H_E^{-1} (X_E[,j] S_j(t)) + (X_E[,j] S_j(t))^T H_E^{-1} (\sum_{v\neq j} X_E[,v] S_v(t)) \\
&+ (X_E[,j] S_j(t))^T H_E^{-1} (X_E[,j] S_j(t))\Big) - \frac{1}{2}\sum_{l=1}^{k} \frac{Z_{v(j)l}(S_j(t)^2 - 2\mu_l(t) S_j(t))}{\alpha_l}\Bigg] \\
&\propto \prod_{t=1}^{T} \exp\Bigg[-\frac{1}{2\sigma_M^2}\Big(-2M(t)^T H_M^{-1} X_M[,j] S_j(t) \\
&+ 2(\sum_{v\neq j} X_M[,v] S_v(t))^T H_M^{-1} (X_M[,j] S_j(t)) + (X_M[,j] S_j(t))^T H_M^{-1} (X_M[,j] S_j(t))\Big) \\
&- \frac{1}{2\sigma_E^2}\Big(-2E(t)^T H_E^{-1} X_E[,j] S_j(t) + 2(\sum_{v\neq j} X_E[,v] S_v(t))^T H_E^{-1} (X_E[,j] S_j(t)) \\
&+ (X_E[,j] S_j(t))^T H_E^{-1} (X_E[,j] S_j(t))\Big) - \frac{1}{2}\sum_{l=1}^{k} \frac{Z_{v(j)l}(S_j(t)^2 - 2\mu_l(t) S_j(t))}{\alpha_l}\Bigg] \\
&\propto \exp \sum_{t=1}^{T} \Bigg[-\frac{1}{2\sigma_M^2}\Big(S_j(t)^2 \big(X_M[,j]^T H_M^{-1} X_M[,j]\big) + S_j(t)\big(-2M(t)^T H_M^{-1} X_M[,j] \\
&+ 2(\sum_{v\neq j} X_M[,v] S_v(t))^T H_M^{-1} X_M[,j]\big)\Big) - \frac{1}{2\sigma_E^2}\Big(S_j(t)^2 \big(X_E[,j]^T H_E^{-1} X_E[,j]\big) \\
&+ S_j(t)\big(-2E(t)^T H_E^{-1} X_E[,j] + 2(\sum_{v\neq j} X_E[,v] S_v(t))^T H_E^{-1} X_E[,j]\big)\Big) \\
&- \frac{1}{2}\sum_{l=1}^{k} \frac{Z_{v(j)l}(S_j(t)^2 - 2\mu_l(t) S_j(t))}{\alpha_l}\Bigg] \\
&\propto \exp -\frac{1}{2}\sum_{t=1}^{T} \Bigg\{ S_j(t)^2 \Bigg[\frac{1}{\sigma_M^2}\Big(X_M[,j]^T H_M^{-1} X_M[,j]\Big) + \frac{1}{\sigma_E^2}\Big(X_E[,j]^T H_E^{-1} X_E[,j]\Big) \\
&+ \sum_{l=1}^{k} \frac{Z_{v(j)l}}{\alpha_l}\Bigg] + S_j(t)\Bigg[\frac{1}{\sigma_M^2}\Big(-2M(t)^T H_M^{-1} X_M[,j] + 2(\sum_{v\neq j} X_M[,v] S_v(t))^T H_M^{-1} X_M[,j]\Big) \\
&+ \frac{1}{\sigma_E^2}\Big(-2E(t)^T H_E^{-1} X_E[,j] + 2(\sum_{v\neq j} X_E[,v] S_v(t))^T H_E^{-1} X_E[,j]\Big) - 2\sum_{l=1}^{k} \frac{\mu_l(t)}{\alpha_l}\Bigg]\Bigg\} \\
&\propto \exp -\frac{1}{2}\sum_{t=1}^{T} \Big\{ S_j(t)^2 W_{1j} + S_j(t) W_{1j}\Big\}
\end{aligned}
$$

Where:

$$W_{1j} = \frac{1}{\sigma_M^2}\Big(X_M[,j]^T H_M^{-1} X_M[,j]\Big) + \frac{1}{\sigma_E^2}\Big(X_E[,j]^T H_E^{-1} X_E[,j]\Big) + \sum_{l=1}^{k} \frac{Z_{v(j)l}}{\alpha_l}$$

$$W_{2j}(t) = \frac{1}{\sigma_M^2}\Big(- 2M(t)^T H_M^{-1} X_M[,j] + 2(\sum_{v\neq j} X_M[,v] S_v(t))^T H_M^{-1} X_M[,j]\Big)$$

$$+ \frac{1}{\sigma_E^2}\Big(- 2E(t)^T H_E^{-1} X_E[,j] + 2(\sum_{v\neq j} X_E[,v] S_v(t))^T H_E^{-1} X_E[,j]\Big) - 2\sum_{l=1}^{k} \frac{\mu_l(t)}{\alpha_l}$$

Since we are interested in the full conditional distribution for $S_j(t)$ over all $t = 1, 2, ..., T$, Then we can write as follows:

$$P(\boldsymbol{S}_j|\text{Rest}) \propto \exp\Big\{ - \frac{1}{2}(\boldsymbol{S}_j - \boldsymbol{\mu}_{S_j})^T \boldsymbol{\Sigma}_{S_j}^{-1} (\boldsymbol{S}_j - \boldsymbol{\mu}_{S_j})\Big\}$$

The precision matrix $\boldsymbol{\Sigma}_{S_j}^{-1}$ is a $T \times T$ matrix where:

$$\begin{aligned}
\boldsymbol{S}_j^T \boldsymbol{\Sigma}_{S_j}^{-1} \boldsymbol{S}_j &= \sum_{t=1}^{T} S_j(t)^2 W_{1j} \\
\Rightarrow \boldsymbol{\Sigma}_{S_j}^{-1} &= Diag(W_{1j}) \\
\text{Also: } \ - 2\boldsymbol{\mu}_{S_j}^T \boldsymbol{\Sigma}_{S_j}^{-1} \boldsymbol{S}_j(t) &= \sum_{t=1}^{T} S_j(t) W_{2j}(t) \\
\Rightarrow -2\boldsymbol{\mu}_{S_j}^T \boldsymbol{\Sigma}_{S_j}^{-1} \boldsymbol{S}_j(t) &= \boldsymbol{W}_{2j}^T \boldsymbol{S}_j(t) \\
\text{where: } \boldsymbol{W}_{2j}^T &= (W_{2j}(1), W_{2j}(2), ..., W_{2j}(T)) \\
-2\boldsymbol{\mu}_{S_j}^T \boldsymbol{\Sigma}_{S_j}^{-1} &= \boldsymbol{W}_{2j}^T \\
\boldsymbol{\mu}_{S_j} &= \big(- \frac{1}{2} \boldsymbol{W}_2^T \boldsymbol{\Sigma}_{S_j}\big)^T \\
\boldsymbol{\mu}_{S_j} &= -\frac{1}{2} \boldsymbol{\Sigma}_{S_j} \boldsymbol{W}_{2j}
\end{aligned}$$

For $t = 1, ..., T$

$$\boldsymbol{\Sigma}_{S_j}^{-1}[t,t] = \frac{1}{\sigma_M^2}\Big(X_M[,j]^T H_M^{-1} X_M[,j]\Big) + \frac{1}{\sigma_E^2}\Big(X_E[,j]^T H_E^{-1} X_E[,j]\Big) + \sum_{l=1}^{k} \frac{Z_{v(j)l}}{\alpha_l}$$

For $t = 1, 2, ..., T,$ for $w = 1, 2, ..., T,$ and $t \neq w$

$$\boldsymbol{\Sigma}_{S_j}^{-1}[t,w] = \frac{1}{\sigma_M^2}(\sum_{v\neq j} X_M[,v] S_v(t))^T H_M^{-1} X_M[,j] + \frac{1}{\sigma_E^2}(\sum_{v\neq j} X_E[,v] S_v(t))^T H_E^{-1} X_E[,j]$$

Now, for β, which is the spatial cohesion parameter for Potts model, the full

condition distribution is :

$$\begin{aligned}
P(\beta|\text{Rest}) &\propto P(\boldsymbol{Z}|\beta) \times P(\beta) \\
&\propto \frac{\exp\{\beta \sum_{h\sim j} \delta(\boldsymbol{Z_v}, \boldsymbol{Z_h})\}}{G(\beta)} \times \frac{1}{\beta_u} \\
&\text{where } \delta(\boldsymbol{Z_v}, \boldsymbol{Z_h}) = 2\boldsymbol{Z_v'}\boldsymbol{Z_h} - 1 \\
&\text{or the approximation based on pseudolikelihood:} \\
&\propto P_{PL}(\boldsymbol{Z}|\beta) \times P(\beta)
\end{aligned}$$

Noticing that the parameterization we used for Potts model, which is the same parameterization as McGrory et al (2009), is:

$$\delta(\boldsymbol{Z_v}, \boldsymbol{Z_h}) = 2\boldsymbol{Z_v'}\boldsymbol{Z_h} - 1 = \begin{cases} -1, & \boldsymbol{Z_v}, \boldsymbol{Z_h} \text{ are not in the same state} \\ 1, & \boldsymbol{Z_v}, \boldsymbol{Z_h} \text{ are in the same state} \end{cases}$$

Let N_{np} denote the total number of neighbours, N_{ss} denote the total number of neighbours share the same state and N_{ns} is the total number of neighbours that do not share the same state. Then we can rewrite the Potts model as:

$$\begin{aligned}
P(\boldsymbol{Z}|\beta_1) &= \frac{\exp\{\beta_1 \sum_{h\sim j} \delta(\boldsymbol{Z_v}, \boldsymbol{Z_h})\}}{G(\beta_1)} \\
&= \frac{1}{G(\beta_1)} \exp\left(\beta_1(N_{ss} - N_{ns})\right) \\
&= \frac{1}{G(\beta_1)} \exp\left(\beta_1(N_{ss} - (N_{np} - N_{ss}))\right) \\
&= \frac{1}{G(\beta_1)} \exp\left(2\beta_1 N_{ss} - \beta_1 N_{np}\right) \\
&= \frac{\exp(-\beta_1 N_{np})}{G(\beta_1)} \exp\left(2\beta_1 N_{ss}\right)
\end{aligned}$$

However, the parameterization used in Moores et al (2015) is that, $\sum_{h\sim j} \delta(\boldsymbol{Z_v}, \boldsymbol{Z_h})$ counts the neighbours that share the same states. which could be expressed as:

$$\begin{aligned}
P(\boldsymbol{Z}|\beta_2) &= \frac{\exp\{\beta_2 \sum_{h\sim j} \delta(\boldsymbol{Z_v}, \boldsymbol{Z_h})\}}{C(\beta_2)} \\
P(\boldsymbol{Z}|\beta_2) &= \frac{1}{C(\beta_2)} \exp(\beta_2 N_{ss})
\end{aligned}$$

The reparameterization could be:

$$\begin{aligned}
2\beta_1 &= \beta_2 \\
\frac{\exp(-\beta_1 N_{np})}{G(\beta_1)} &= \frac{1}{C(\beta_2)} \\
\Rightarrow C(\beta_2) \exp(-\beta_1 N_{np}) &= G(\beta_1)
\end{aligned}$$

The Pseudolikelihood for $\boldsymbol{Z}$ is defined as:

$$\begin{aligned} P_{PL}(\boldsymbol{Z}|\beta) &= \prod_{i=1}^{N_v} P(\boldsymbol{Z}_i|\boldsymbol{Z}_{-i},\beta) \\ &= \prod_{i=1}^{N_v} \frac{\exp\left(2\beta \sum_{j=1}^{k} Z_{ij} \sum_{l\in\delta_i} Z_{lj}\right)}{\sum_{q=1}^{k} \exp(2\beta \sum_{l\in\delta_i} Z_{lq})} \end{aligned}$$

For ICM update, and assuming a pseudolikelihood approximation, we want the value $\hat{\beta}$ that maximizes the following function over $[0, \beta_u]$

$$f(\beta) = P_{PL}(Z|\beta)P(\beta).$$

For $f(\beta)$,

$$\begin{aligned} f(\beta) &= \prod_{i=1}^{N_v} P(\boldsymbol{Z}_i|\boldsymbol{Z}_{-i},\beta) \times P(\beta) \\ &= \prod_{i=1}^{N_v} \frac{\exp\left(2\beta \sum_{j=1}^{k} Z_{ij} \sum_{l\in\delta_i} Z_{lj}\right)}{\sum_{q=1}^{k} \exp(2\beta \sum_{l\in\delta_i} Z_{lq})} \times \boldsymbol{I}(0 \le \beta \le \beta_u) \\ H(\beta) = \log(f(\beta)) &= \log(I(0 \le \beta \le \beta_u)) + \sum_{v=1}^{N_v} \Big[2\beta \sum_{j=1}^{k} Z_{ij} \sum_{l\in\delta_i} Z_{lj} - \log\big\{ \sum_{q=1}^{k} \exp(2\beta \sum_{l\in\delta_i} Z_{lq}) \big\} \Big] \\ &= 2\beta \sum_{i=1}^{N_v} \sum_{j=1}^{k} Z_{ij} \sum_{l\in\delta_i} Z_{lj} - \sum_{i=1}^{N_v} \log\{ \sum_{q=1}^{k} \exp(2\beta \sum_{l\in\delta_i} Z_{lq}) \} \end{aligned}$$

Taking the derivative of $H(\beta)$ and assuming β is in $[0, \beta_u]$, we have:

$$H'(\beta) = 2 \sum_{i=1}^{N_v} \sum_{j=1}^{k} Z_{ij} \sum_{l\in\delta_i} Z_{lj} - \sum_{i=1}^{N_v} \{ \sum_{q=1}^{k} \exp(2\beta \sum_{l\in\delta_i} Z_{lq}) \}^{-1} (\sum_{q=1}^{k} \exp(2\beta \sum_{l\in\delta_i} Z_{lq})) (2 \sum_{l\in\delta_i} Z_{lq})$$

We will use the forms of $H(\beta)$ and $H'(\beta)$ to numerically maximize $H(\beta)$ at each ICM iteration. This constrained 1-dimensional maximization is easily carried out using numerical routines in the R programming language.

The full conditional distribution of $\boldsymbol{Z}$ is:

$$P(\boldsymbol{Z}|Rest) \propto [\prod_{j=1}^{p} \prod_{t=1}^{T} P(S_j(t))|\boldsymbol{\mu}^A(t), \boldsymbol{\alpha}, \boldsymbol{Z_{v(j)}}] \times P(\boldsymbol{Z}|\beta)$$

For now, let's focus on the label for rth voxel, which is $\boldsymbol{Z}_r$. Let N_{jr} denote the number

of points such that being mapped into rth voxel as $j|v(j)=r$. Then:

$$
\begin{aligned}
P(\boldsymbol{Z}_r|Rest) &\propto \Big[\prod_{j|v(j)=r}\prod_{l=1}^{k}\alpha_l^{-TZ_{v(j)l}/2}\exp\big(-\sum_{t=1}^{T}\frac{Z_{v(j)l}(S_j(t)-\mu_l(t))^2}{2\alpha_l}\big)\Big]\times P_{PL}(\boldsymbol{Z}_r|\beta)\\
&\propto \Big[\prod_{j|v(j)=r}\prod_{l=1}^{k}\alpha_l^{-TZ_{v(j)l}/2}\exp\big(-\sum_{t=1}^{T}\frac{Z_{v(j)l}(S_j(t)-\mu_l(t))^2}{2\alpha_l}\big)\Big]\\
&\times P(\boldsymbol{Z}_r|Z_{\delta_r},\beta)\prod_{i=\delta_r}P(Z_i|Z_{\delta_i},\beta)\\
&\propto \Big[\prod_{j|v(j)=r}\prod_{l=1}^{k}\alpha_l^{-TZ_{v(j)l}/2}\exp\big(-\sum_{t=1}^{T}\frac{Z_{v(j)l}(S_j(t)-\mu_l(t))^2}{2\alpha_l}\big)\Big]\times\exp(2\beta\sum_{j=1}^{k}\sum_{l\in\delta_r}Z_{rj}Z_{lj})\\
&\times\prod_{i\in\delta_r}\frac{\exp(2\beta\sum_{j=1}^{k}Z_{ij}Z_{rj})}{\sum_{q=1}^{k}\exp(2\beta\sum_{l\in\delta_i}Z_{lq})}\\
&\propto \Big[\prod_{j|v(j)=r}\prod_{l=1}^{k}\alpha_l^{-TZ_{v(j)l}/2}\exp\big(-\sum_{t=1}^{T}\frac{Z_{v(j)l}(S_j(t)-\mu_l(t))^2}{2\alpha_l}\big)\Big]\times\exp(2\beta\sum_{j=1}^{k}\sum_{l\in\delta_r}Z_{rj}Z_{lj})\\
&\times\frac{\exp(2\beta\sum_{i\in\delta_i}\sum_{j=1}^{k}Z_{ij}Z_{rj})}{\prod_{i\in\delta_i}(\sum_{q=1}^{k}\exp(2\beta\sum_{l\in\delta_i}Z_{lq}))}\\
&\propto \Big[\prod_{j|v(j)=r}\prod_{l=1}^{k}\alpha_l^{-TZ_{v(j)l}/2}\exp\big(-\sum_{t=1}^{T}\frac{Z_{v(j)l}(S_j(t)-\mu_l(t))^2}{2\alpha_l}\big)\Big]\\
&\times\frac{\exp\big(2\beta\sum_{j=1}^{k}\sum_{l\in\delta_r}Z_{rj}Z_{lj}+2\beta\sum_{i\in\delta_i}\sum_{j=1}^{k}Z_{ij}Z_{rj})+\big)}{\prod_{i\in\delta_i}(\sum_{q=1}^{k}\exp(2\beta\sum_{l\in\delta_i}Z_{lq}))}\\
&\propto \Big[\prod_{j|v(j)=r}\prod_{l=1}^{k}\alpha_l^{-TZ_{v(j)l}/2}\exp\big(-\sum_{t=1}^{T}\frac{Z_{v(j)l}(S_j(t)-\mu_l(t))^2}{2\alpha_l}\big)\Big]\\
&\times\frac{\exp\big(4\beta\sum_{j=1}^{k}\sum_{l\in\delta_r}Z_{rj}Z_{lj}\big)}{\prod_{i\in\delta_r}(\sum_{q=1}^{k}\exp(2\beta\sum_{l\in\delta_i}Z_{lq}))}\\
&\propto \Big[\prod_{l=1}^{k}\alpha^{-TN_{jr}Z_{rl}/2}\Big]\times\exp\Big[-\sum_{l=1}^{k}Z_{rl}\sum_{t=1}^{T}\sum_{j|v(j)=r}\frac{(S_j(t)-\mu_l(t))^2}{2\alpha_l}+4\beta\sum_{j=1}^{k}\sum_{l\in\delta_r}Z_{rj}Z_{lj}\Big]\\
&\times\frac{1}{\prod_{i\in\delta_r}(\sum_{q=1}^{k}\exp(2\beta\sum_{l\in\delta_i}Z_{lq}))}
\end{aligned}
$$

When $Z_{rh}=1$, We could find that:

$$
\begin{aligned}
P(\boldsymbol{Z}_{rh}=1|Rest) &\propto \alpha^{-TN_{jr}/2}\times\exp\Big[-\sum_{t=1}^{T}\sum_{j|v(j)=r}\frac{(S_j(t)-\mu_l(t))^2}{2\alpha_l}+4\beta\sum_{l\in\delta_r}Z_{rj}Z_{lj}\Big]\\
&\times\frac{1}{\prod_{i\in\delta_r}\sum_{q=1}^{k}\exp\big(2\beta(\boldsymbol{I}(q=h)+\sum_{l\in\delta_i,l\neq r}Z_{lq})\big)}
\end{aligned}
$$

Then, the probability of $Z_{rh} = 1$ give the rest could be computed as:

$$P(Z_{rh} = 1|Rest) = \frac{\frac{\alpha_h^{-TN_{jr}/2} \times \exp\left[-\sum_{t=1}^{T}\sum_{j|v(j)=r}\frac{(S_j(t)-\mu_h(t))^2}{2\alpha_h} + 4\beta\sum_{l\in\delta_r} Z_{rh}Z_{lh}\right]}{\prod_{i\in\delta_r}\sum_{q=1}^{k}\exp\left(2\beta(\boldsymbol{I}(q=h)+\sum_{l\in\delta_i, l\neq r} Z_{lq})\right)}}{\sum_{h=1}^{k}\frac{\alpha_h^{-TN_{jr}/2} \times \exp\left[-\sum_{t=1}^{T}\sum_{j|v(j)=r}\frac{(S_j(t)-\mu_h(t))^2}{2\alpha_h} + 4\beta\sum_{l\in\delta_r} Z_{rh}Z_{lh}\right]}{\prod_{i\in\delta_r}\sum_{q=1}^{k}\exp\left(2\beta(\boldsymbol{I}(q=h)+\sum_{l\in\delta_i, l\neq r} Z_{lq})\right)}}$$

For block updating, let's define the block setting as:

$$B = \{\text{Black Square Index}\}$$
$$W = \{\text{White Square Index}\}$$

Then, the labelling for white square or black square could be expressed as:

$$\boldsymbol{Z}_B = \{Z_i|\ i \in B\}$$
$$\boldsymbol{Z}_W = \{Z_i|\ i \in W\}$$

We know that (Moores et al., 2015):

$$\begin{aligned}
P(\boldsymbol{Z}_B|\boldsymbol{Z}_W, \beta) &\propto P(\boldsymbol{Z}|\beta_2) \\
&\propto \prod_{l\in\boldsymbol{B}} f_l(\boldsymbol{Z}_l|\boldsymbol{Z}_W, \beta_2) \\
&\propto \prod_{l\in B}\prod_{j\in\delta_l} \exp\left(\beta_2 \times \delta(\boldsymbol{Z}_l, \boldsymbol{Z}_j)\right) \\
&\propto \prod_{l\in B} \exp\left(\beta_2 \sum_{j\in\delta_l} \delta(\boldsymbol{Z}_l, \boldsymbol{Z}_j)\right) \\
&\propto \prod_{l\in B} \exp\left(2\beta \boldsymbol{Z}_l^T \sum_{j\in\delta_l} \boldsymbol{Z}_j\right)
\end{aligned}$$

By symmetry, we could also get the white square as:

$$P(\boldsymbol{Z}_W|\boldsymbol{Z}_B, \beta) \propto \prod_{l\in W} \exp\left(2\beta \boldsymbol{Z}_l^T \sum_{j\in\delta_l} \boldsymbol{Z}_j\right)$$

With normalizing constant, we can get $P(\boldsymbol{Z}_B|\boldsymbol{Z}_W, \beta)$ as:

$$P(\boldsymbol{Z}_B|\boldsymbol{Z}_W, \beta) = \prod_{l\in B} \frac{\exp\left(2\beta \boldsymbol{Z}_l^T \sum_{j\in\delta_l} \boldsymbol{Z}_j\right)}{\sum_{q=1} k \exp\left(2\beta \boldsymbol{e}_q^T \sum_{j\in\delta_l} \boldsymbol{Z}_j\right)}$$

where $\boldsymbol{e}_q^T = (0, 0, 0, 0, 0, 0, 1, 0, 0, ...)$ is a coordinate vector that qth location is 1 and

0 otherwise. Now, the full conditional distribution for Z_B is:

$$
\begin{aligned}
P(\boldsymbol{Z}_B|Rest) &\propto \Big[\prod_{j=1}^{p}\prod_{t=1}^{T} P(S_j(t))|\boldsymbol{\mu}^A(t), \boldsymbol{\alpha}, \boldsymbol{Z_{v(j)}}\Big] \times P(\boldsymbol{Z}|\beta) \\
&\propto \Big[\prod_{j=1}^{p}\prod_{t=1}^{T} P(S_j(t))|\boldsymbol{\mu}^A(t), \boldsymbol{\alpha}, \boldsymbol{Z_{v(j)}}\Big] \times P(\boldsymbol{Z}_B|\boldsymbol{Z}_W, \beta) \\
&\propto \Big[\prod_{l\in B}\prod_{j|v(j)=l}\prod_{t=1}^{T} P(S_j(t))|\boldsymbol{\mu}^A(t), \boldsymbol{\alpha}, \boldsymbol{Z_{v(j)}}\Big] \times P(\boldsymbol{Z}_B|\boldsymbol{Z}_W, \beta) \\
&\propto \prod_{l\in B}\Big([\prod_{j|v(j)=l}\prod_{t=1}^{T} P(S_j(t)|\boldsymbol{\mu}^A(t), \boldsymbol{\alpha}, \boldsymbol{Z_{v(j)}})] \times \exp(2\beta \boldsymbol{Z}_l^T \sum_{j\in\delta_l} \boldsymbol{Z}_j)\Big) \\
&\propto \prod_{l\in B}\Big([\prod_{j|v(j)=l}\prod_{t=1}^{T}\prod_{\kappa=1}^{k} \alpha_\kappa^{-Z_{v(j)\kappa}/2} \exp(-\frac{Z_{v(j)\kappa}(S_j(t)-\mu_\kappa(t))^2}{2\alpha_\kappa})] \times \exp(2\beta \boldsymbol{Z}_l^T \sum_{j\in\delta_l} \boldsymbol{Z}_j)\Big) \\
&\propto \prod_{l\in B}[\prod_{\kappa=1}^{k}\prod_{j|v(j)=l} \alpha_\kappa^{-TZ_{v(j)\kappa}/2} \times \exp(-\sum_{t=1}^{T}\frac{Z_{v(j)\kappa}(S_j(t)-\mu_\kappa(t))^2}{2\alpha_\kappa})] \times \exp(2\beta \boldsymbol{Z}_l^T \sum_{j\in\delta_l} \boldsymbol{Z}_j) \\
&\propto \prod_{l\in B}[\prod_{\kappa=1}^{k} \alpha_\kappa^{-TZ_{lk}}]^{N_{jr}/2} \times \exp\Big(-\frac{1}{2}\sum_{j|vj=l}\sum_{\kappa=1}^{k}\frac{Z_{l\kappa}}{\alpha_\kappa}\sum_{t=1}^{T}(S_j(t)-\mu_\kappa(t))^2 \\
&+ 2\beta\sum_{\kappa=1}^{k}\sum_{j\in\delta_l} Z_{lk}Z_{jk}\Big)
\end{aligned}
$$

Therefore, we can have that (κ is the voxel and h is the mixture component):

$$
P(Z_{\kappa h}=1|Rest) = \frac{\alpha_h^{-TN_{j\kappa}/2} \times \exp\big(-\frac{1}{2}\sum_{j|v(j)=\kappa}\alpha_h^{-1}\sum_{t=1}^{T}(S_j(t)-\mu_h(t))^2 + 2\beta\sum_{v\in\delta_\kappa} Z_{vh}\big)}{\sum_{l=1}^{K}\alpha_l^{-TN_{j\kappa}/2} \times \exp\big(-\frac{1}{2}\sum_{j|v(j)=\kappa}\alpha_l^{-1}\sum_{t=1}^{T}(S_j(t)-\mu_l(t))^2 + 2\beta\sum_{v\in\delta_\kappa} Z_{vl}\big)}
$$

$$
\sum_{l=1}^{K}\alpha_l^{-TN_{j\kappa}/2} \times \exp\big(-\frac{1}{2}\sum_{j|v(j)=\kappa}\alpha_l^{-1}\sum_{t=1}^{T}(S_j(t)-\mu_l(t))^2 + 2\beta\sum_{v\in\delta_\kappa} Z_{vl}\big)
$$

The full conditional for $\boldsymbol{Z}_B$ is:

$$
P(\boldsymbol{Z}_B|Rest) = \prod_{l\in B}\left[\frac{[\prod_{\kappa=1}^{k}\alpha_\kappa^{-TZ_{lk}}]^{N_{jr}} \times \exp\Big(-\frac{1}{2}\sum_{j|vj=l}\sum_{\kappa=1}^{k}\frac{Z_{l\kappa}}{\alpha_\kappa}\sum_{t=1}^{T}(S_j(t)-\mu_\kappa(t))^2 + 2\beta\sum_{\kappa=1}^{k}\sum_{j\in\delta_l} Z_{lk}Z_{jk}\Big)}{\sum_{h=1}^{k}\alpha_h^{-TN_{jl}/2} \times \exp\big(-\frac{1}{2}\sum_{j|v(j)=l}\alpha_h^{-1}\sum_{t=1}^{T}(S_j(t)-\mu_\kappa(t))^2 + 2\beta\sum_{v\in\delta_l} Z_v h\big)}\right]
$$

The full conditional for $\boldsymbol{Z}_W$ is obtained in the same way and has the same form modulo minor change in notation.

A.3 Analysis of Synthetic Data

In this section we evaluate our methodology using synthetic data. We consider four cases each with different levels of complexity in the true scene. For simplicity, in each of our four examples we fix K in our algorithm to be the true number of mixture components, though, in Section 2.4 of Chapter 2 we present a simulation study designed to evaluate our estimator of the number of mixture components $\hat{K}_{ICM}$.

Two Sources with Gaussian Signals

We simulate brain activity on 8,196 locations on the cortex with two active subregions, the first containing 250 locations and the second containing 150 locations. The locations including those comprising the active regions are depicted in Figure A.3, panel (a). The neural activity $S_j(t)$ for locations j in each of the active subregions is based on the two Gaussian curves depicted in Figure A.4, panel (a) while inactive locations have $S_j(t) = 0$. The neural activity $\boldsymbol{S}(t)$ is projected onto the MEG and EEG sensor arrays using the forward operators $\boldsymbol{X}_M$ and $\boldsymbol{X}_E$. The simulated data are then obtained by adding Gaussian noise at each sensor, where the variance of the noise at each sensor is set to be 5% of the temporal variance of the signal at that sensor.

We run the ICM algorithm with $K = 3$ with the cortical locations divided into $J = 250$ clusters over a 3D grid of $N_v = 450$ voxels. The required computing time is roughly 400 seconds on a MacBook Pro with a 2.7 GHz Intel Core i5 processor and 8 GB memory. The estimated allocation variables $\hat{\boldsymbol{Z}}$ are depicted in Figure A.3, panel (b). The two regions of neural activity appear to be correctly localized spatially, while the estimated source time series are depicted in Figure A.4, panel (b). The temporal patterns of the latter match the true signals depicted in Figure A.4, panel (a) reasonably well with both temporal peaks being correctly identified, though there appears to be some over-estimation of the amplitudes in the first component. The overestimation in amplitude can be up to a factor of 2 and then, there is also some underestimation by as much as a factor of 4. Zooming in suggests that there are a few cases where the signal would not have been detected because it would be lost in the noise. In summary, both Figure A.3 panel (b) and Figure A.4, panel (b) indicate that the spatial and temporal localizations are adequate, in the sense that the broad scale features of the signal are recovered, and the required computing time is also reasonable given the complexity of the model and the dimension of parameter space.

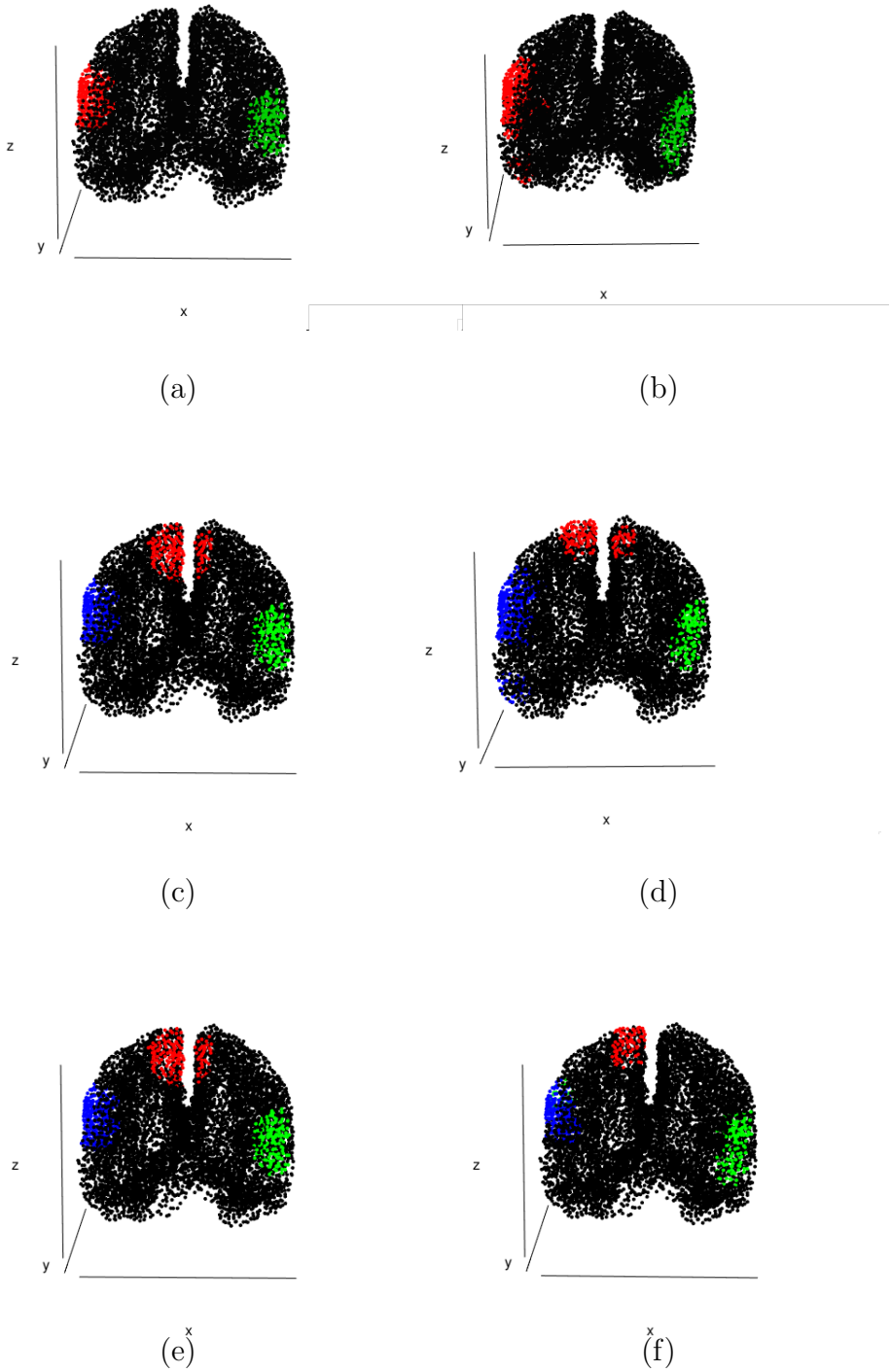


Figure A.3: The true partition of the cortex into active and inactive states for examples of Appendix Section A.3 for synthetic data analysis (for $K = 3$ and $K = 4$) and the simulation studies of Section 2.4 (for $K = 3$ and $K = 4$) are depicted in the left column. The right column presents the corresponding estimated mixture allocation variables ($\hat{\boldsymbol{Z}}$) for the examples considered in Section 2.4 (for $K = 3$ and $K = 4$).

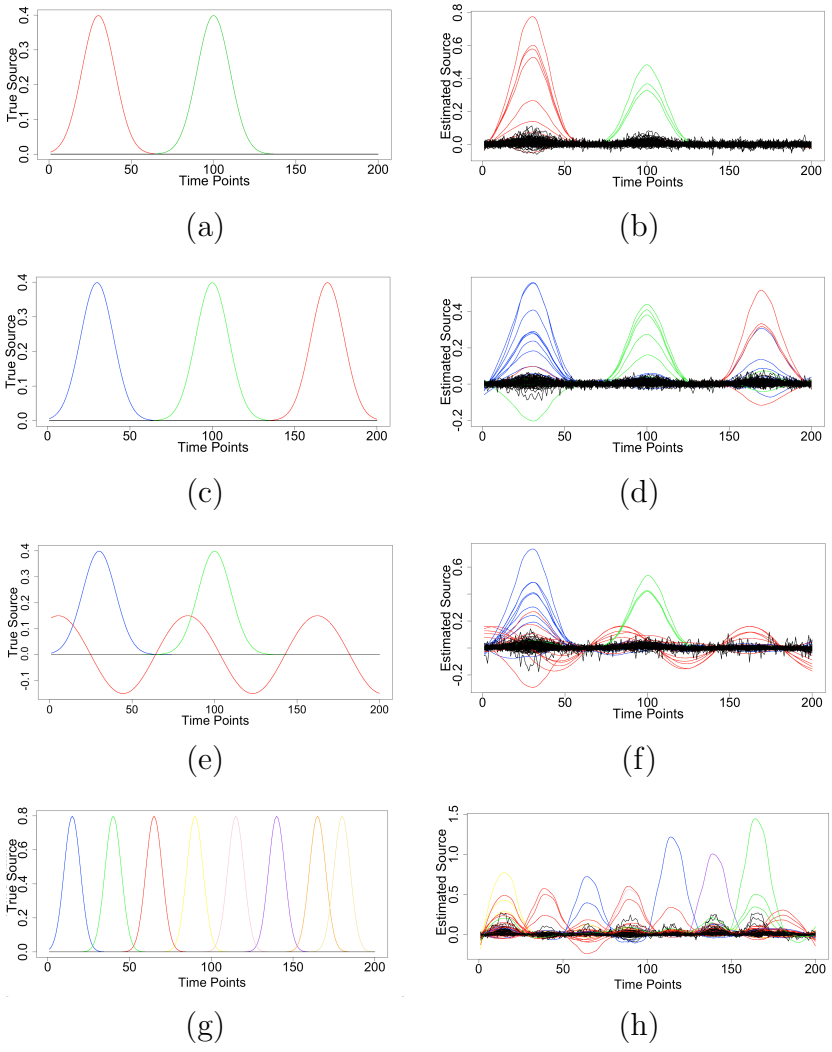


Figure A.4: The true signal $S_j(t)$ used in each of the distinct active and inactive regions in the four examples considered in the Appendix Section A.3 for synthetic data analysis and the simulation studies of Section 2.4 (for $K = 3$, $K = 4$ and $K = 9$) are depicted in the left column. The right column presents the corresponding estimated sources $\hat{S}_j(t)$ at each location of the cortex in the examples of Section 2.4.

Three Sources with Gaussian or Sinusoidal Signals

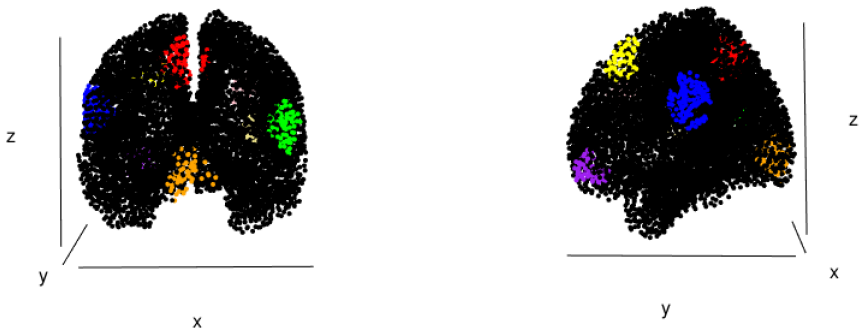
In our second example we consider three subregions of activity as depicted in Figure A.3, panel (c). The neural activity $S_j(t)$ is based on the three Gaussian curves depicted in Figure A.4, panel (c). The data are otherwise simulated in the same manner as described in the previous section. We run the ICM algorithm with $K = 4$ and other settings as previously described and the running time is again approximately 400 seconds. The estimated allocation variables are depicted in Figure A.3, panel (d), and when compared to the true allocations in Figure A.3, panel (c), we see that the three regions of neural activity have been identified reasonably well spatially. Comparing the true signals in Figure A.4, panel (c), to the estimated sources in Figure A.4, panel (d), we see that the three temporal peaks of activity have been correctly identified, though we note that the curves $(S_j(t))$ for a few locations have been estimated incorrectly with respect to their shape and some of these have estimated amplitudes that are negative. Still, the overall reconstruction of the neural activity appears reasonably accurate in this case.

In our third example we again consider the three subregions of activity where the Gaussian signal in the third region is replaced with a sinusoid as depicted in Figure A.4, panel (e). In this case the signal from the third activated region overlaps with both signals from the other active regions. The ICM algorithm requires the same computing time as in the previous examples. The estimated mixture allocation variables are depicted in Figure A.3, panel (f), and comparing to the true scene our algorithm appears to have correctly spatially localized the three regions of neural activity, though part of the third region (the red coloured region in the third row of Figure A.3) appears to have been incorrectly classified as 'inactive'. Examining the estimated sources $\hat{S}_j(t)$ in Figure A.4, panel (f), in comparison to the true signals in Figure A.4, panel (e), we see that the patterns of the temporal signals including the sinusoid appear to be mostly well estimated.

Eight Sources with Gaussian Signals

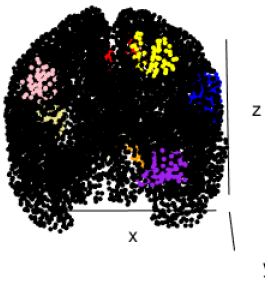
In our fourth and final example we consider a much more difficult setting where there are eight active regions, each with spatial extent roughly one-fourth that of the active regions considered in the previous three examples. The true spatial configuration of the nine states is depicted from various different angles in Figure A.5. The temporal profile of the brain activation is represented with eight Gaussian signals depicted

in Figure A.4, panel (g). The running time for the algorithm is approximately 800 seconds and we obtain, in this case, $\hat{K}_{ICM} = 7$. The estimated signals are shown in Figure A.4, panel (h), while the estimated state allocation variables are shown in Figure A.6, with the panels of this figure corresponding to the panels of Figure A.5 where the true states are depicted. Overall, the algorithm is able to capture roughly the broad spatiotemporal pattern of brain activation, though the quality of the estimates relative to the simpler examples is not as high. Nevertheless, it appears as though the approach can be applied successfully to reconstruct general patterns of brain activity in this more difficult case, while the true number of states in the brain is under-estimated by 2.

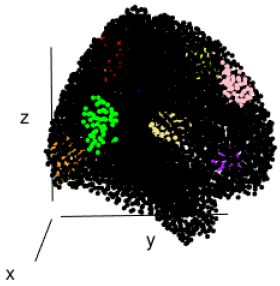


(a)

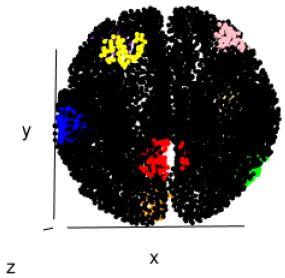
(b)



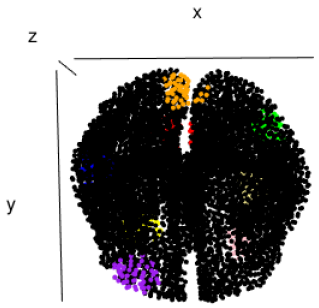
(c)



(d)



(e)



(f)

Figure A.5: The true partition of the cortex into active and inactive states for the case of $K = 9$ states.

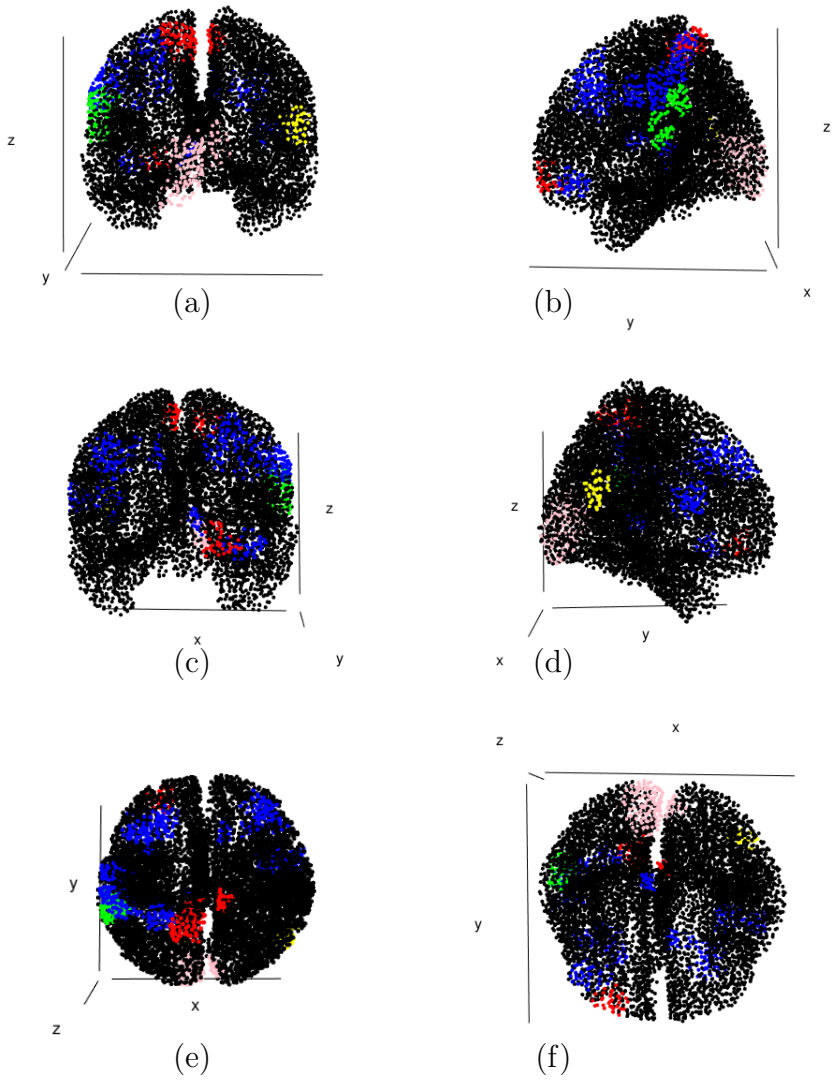


Figure A.6: The estimated allocation of the cortex into active and inactive states for the case of $K = 9$ true states. In this case $\hat{K}_{ICM} = 7$. The panels in this figure correspond to the panels in Figure A.5.

Appendix B

Appendix for Chapter 3

B.1 Selected SNPs and the Corresponding Regions of Interest for the ADNI-1 Application

Table B.1: Application to ADNI-1 data: The 75 SNPs and corresponding phenotypes selected from the proposed Bayesian spatial group lasso regression model with Gibbs Sampling combined with Bayesian FDR at $\alpha = 0.05$. These same SNP-ROI pairs are also selected by variational Bayes combined with Bayesian FDR at $\alpha = 0.05$. SNPs and phenotypes in bold correspond to those also chosen using 95% credible intervals and the model of Greenlaw et al. (2017).

SNP	Gene	Phenotype ID (hemisphere)
rs4305	ACE	SupTemporal(L), Supramarg(L)
rs4311	ACE	AmygVol(R), **CerebCtx(R)**, HippVol(R), **InfParietal(R)**, Parahipp(R), Precentral(L), **SupFrontal(R)**,**SupParietal(R)**, Supramarg(R), TemporalPole(R), MeanCing(R), MeanMedTemp(R)
rs4353	ACE	Supramarg(R)
rs405509	APOE	**MidTemporal(R)**, **Supramarg(R)**, **MeanFront(R)**, **MeanLatTemp(R)**
rs11191692	CALHM1	SupTemporal(L)
rs3811450	CHRNB2	SupParietal(R)
rs2025935	CR1	Postcentral(L), Supramarg(L)
rs10780849	DAPK1	InfParietal(R)
rs1105384	DAPK1	TemporalPole(R), MeanCing(L)
rs1473180	DAPK1	CerebWM(L)
rs17399090	DAPK1	MeanCing(L)
rs3095747	DAPK1	Postcentral(L)
rs3118853	DAPK1	SupTemporal(L)
rs3124237	DAPK1	InfParietal(R)
rs3124238	DAPK1	SupTemporal(L)
rs4877368	DAPK1	Parahipp(R)
rs4878117	DAPK1	Parahipp(R)
rs913782	DAPK1	InfParietal(R)

Continued on next page

Table B.1 – *Continued from previous page*

SNP	Gene	Phenotype ID (hemisphere)
rs10916959	ECE1	Supramarg(L)
rs212539	ECE1	SupTemporal(L), Supramarg(L)
rs4654916	ECE1	SupTemporal(L), Supramarg(L)
rs6584307	ENTPD7	InfParietal(R)
rs11601726	GAB2	SupTemporal(L), Supramarg(L)
rs7927923	GAB2	SupTemporal(L)
rs17561	IL1A	InfTemporal(R)
rs16924159	IL33	TemporalPole(L), **MeanCing(L)**
rs928413	IL33	Postcentral(L)
rs1433099	LDLR	CerebWM(L), SupFrontal(R)
rs2228671	LDLR	MidTemporal(R)
rs2569537	LDLR	MeanCing(R)
rs6511720	LDLR	Postcentral(L), Supramarg(L)
rs688	LDLR	Supramarg(L)
rs2184226	MTHFR	SupTemporal(L), Supramarg(L)
rs3737964	MTHFR	MeanSensMotor(R)
rs4846048	MTHFR	Supramarg(L)
rs12209631	NEDD9	CerebWM(L)
rs1475345	NEDD9	InfParietal(R)
rs16871157	NEDD9	SupTemporal(L), Supramarg(L)
rs17496723	NEDD9	MeanFront(R)
rs2072834	NEDD9	Supramarg(L)
rs2182335	NEDD9	Precuneus(R), MeanTemp(R)
rs2182337	NEDD9	SupTemporal(L)
rs2950	NEDD9	SupTemporal(L), Supramarg(L)
rs4713379	NEDD9	InfParietal(R)
rs744970	NEDD9	Supramarg(L)
rs760680	NEDD9	PostCing(L), Postcentral(L), SupTemporal(L), Supramarg(L)
rs10501604	PICALM	Supramarg(L)
rs7938033	PICALM	Supramarg(L)
rs6084833	PRNP	PostCing(L), SupTemporal(L)
rs10748924	SORCS1	InfTemporal(R)
rs10786972	SORCS1	MeanCing(L), MeanTemp(R)
rs10787010	SORCS1	PostCing(L), MeanSensMotor(L)
rs10787011	SORCS1	Supramarg(L)
rs10884399	SORCS1	Supramarg(L)
rs11193198	SORCS1	SupTemporal(L)
rs12240854	SORCS1	Postcentral(L), SupTemporal(L)
rs1269918	SORCS1	**CerebWM(L)**
rs1887635	SORCS1	SupTemporal(L)
rs2149196	SORCS1	MidTemporal(R), Parahipp(R)
rs2243581	SORCS1	SupTemporal(L)
rs2418811	SORCS1	PostCing(L), SupTemporal(L)
rs596577	SORCS1	Supramarg(L)
rs7903481	SORCS1	InfParietal(R), InfTemporal(R)
rs10502262	SORL1	Postcentral(L)
rs1699102	SORL1	PostCing(L), Postcentral(L), SupTemporal(L), Supramarg(L)
rs1699105	SORL1	SupTemporal(L)
rs2276346	SORL1	Supramarg(L)
rs3781832	SORL1	Supramarg(L)

Continued on next page

Table B.1 – *Continued from previous page*

SNP	Gene	Phenotype ID (hemisphere)
rs4936632	SORL1	SupTemporal(L)
rs661057	SORL1	SupTemporal(L)
rs726601	SORL1	Supramarg(L)
rs762484	TF	MeanCing(L)
rs1568400	THRA	MeanTemp(R)
rs3744805	THRA	HippVol(R), Parahipp(R), Precuneus(R)
rs7219773	TNK1	Parahipp(R)

B.2 Derivations for the Gibbs Sampling and Mean Field Variational Bayes Algorithm

Based on the hierarchical prior setting, the joint posterior distribution can be expressed up to a normalizing constant as

$$
\begin{aligned}
p(\boldsymbol{W}\,\omega_1^2,\cdots,\omega_{\mathrm{d}}^2,\Sigma,|\boldsymbol{Y}) &\propto p(\boldsymbol{Y}|\boldsymbol{W},\Sigma)p(\boldsymbol{W}|\Sigma,\boldsymbol{\omega}^2)p(\boldsymbol{\omega}^2)p(\Sigma)\\
&\propto |(D_A-\rho A)^{-1}\otimes\Sigma|^{-\frac{n}{2}}\exp\Bigg\{-\frac{1}{2}\sum_{\ell=1}^{n}(\boldsymbol{y}_\ell-\boldsymbol{W}^T\boldsymbol{x}_\ell)^T[(D_A-\rho A)^{-1}\otimes\Sigma]^{-1}\\
&\quad(\boldsymbol{y}_\ell-\boldsymbol{W}^T\boldsymbol{x}_\ell)\Bigg\}\\
&\quad\times\prod_{\mathrm{i}=1}^{\mathrm{d}}(\omega_i^2)^{-\frac{c}{2}}|\Sigma|^{-\frac{c}{4}}\exp\{-\frac{1}{2}\sum_{j=1}^{\frac{c}{2}}\tilde{W}_{ij}^T(\omega_i^2\Sigma)^{-1}\tilde{W}_{ij}\}\\
&\quad\times\prod_{\mathrm{i}=1}^{\mathrm{d}}\frac{(\frac{\lambda^2}{2})^{\frac{c+1}{2}}}{\Gamma(\frac{c+1}{2})}(\omega_i^2)^{\frac{c}{2}-\frac{1}{2}}\exp\{-\frac{\lambda^2}{2}\omega_i^2\}\\
&\quad\times\frac{|S|^{\frac{v}{2}}}{2^v\Gamma_2(\frac{v}{2})}|\Sigma|^{-\frac{v+3}{2}}\exp\Big\{-\frac{1}{2}tr(S\Sigma^{-1})\Big\}
\end{aligned}
$$

The full conditional distribution of $\boldsymbol{W^{(k)}}$

The full conditional distribution of $\boldsymbol{W^{(k)}},\ k=1,\ldots,d$ is expressed as

$$(\boldsymbol{W}^{(k)T})\big|\boldsymbol{Y},\boldsymbol{W}^{(-k)},\boldsymbol{\omega},\Sigma\sim MVN_{m_kc}(\underset{\sim}{\mu}_{\mathrm{k}},\ \boldsymbol{\Sigma}_k),$$

where

$$
\begin{aligned}
\underset{\sim}{\mu}_{\mathrm{k}} = \boldsymbol{\Sigma}_k\Bigg(&-\sum_{\ell=1}^{\mathrm{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)[(D_A-\rho A)\otimes\Sigma^{-1}](\boldsymbol{x}_\ell^{(-k)T}\otimes I_c)(\boldsymbol{W}^{(-k)T})\\
&+\sum_{\ell=1}^{\mathrm{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)[(D_A-\rho A)\otimes\Sigma^{-1}]\boldsymbol{y}_\ell\Bigg)
\end{aligned}
$$

$$\boldsymbol{\Sigma}_k = \left(\boldsymbol{H}_k + \sum_{\ell=1}^{\mathrm{n}} (\boldsymbol{x}_\ell^{(k)} \otimes I_c)[(D_A - \rho A) \otimes \Sigma^{-1}](\boldsymbol{x}_\ell^{(k)T} \otimes I_c) \right)^{-1}.$$

$$\boldsymbol{H}_k = \left[\left\{ \frac{1}{\omega_k^2} \right\} \otimes I_{\frac{c}{2}} \otimes \Sigma^{-1} \right].$$

Since we already have the full conditional distribution, the mean field approximation for variational bayes can be derived as:

$$\begin{aligned}
q((\boldsymbol{W}^{(k)T})) &\propto exp\Big\{ \mathbf{E}_{-k}(logP((\boldsymbol{W}^{(k)T})|rest)) \Big\} \\
&\propto exp\Big\{ \mathbf{E}_{-k}\Big(-\frac{c}{2} log(2\pi) - \frac{1}{2} log(det|\Sigma_k|) - \frac{1}{2}((\boldsymbol{W}^{(k)T}) - \underset{\sim}{\mu}_{\mathrm{k}})^T \boldsymbol{\Sigma}_k^{-1}((\boldsymbol{W}^{(k)T}) - \underset{\sim}{\mu}_{\mathrm{k}}) \Big) \Big\} \\
&\propto exp\Big\{ \mathbf{E}_{-k}\Big(-const - \frac{1}{2}(\boldsymbol{W}^{(k)T})^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{W}^{(k)T}) + (\boldsymbol{W}^{(k)T})^T \boldsymbol{\Sigma}_k^{-1} \underset{\sim}{\mu}_{\mathrm{k}} \Big) \Big\} \\
&\propto exp\Big\{ const - \frac{1}{2}(\boldsymbol{W}^{(k)T})^T \mathbf{E}_{-k}(\boldsymbol{\Sigma}_k^{-1})(\boldsymbol{W}^{(k)T}) + (\boldsymbol{W}^{(k)T})^T \mathbf{E}_{-k}\big(\boldsymbol{\Sigma}_k^{-1} \underset{\sim}{\mu}_{\mathrm{k}}\big) \Big\}
\end{aligned}$$

We still can see that $q((\boldsymbol{W}^{(k)T}))$ is still MVN with

$$\begin{aligned}
\boldsymbol{\Sigma}_{q(W^k)}^{-1} &= \mathbf{E}_{-k}(\boldsymbol{\Sigma}_k^{-1}) \\
&= E_{-k}\left(\left[\left\{ \frac{1}{\omega_k^2} \right\} \otimes I_{\frac{c}{2}} \otimes \Sigma^{-1} \right] + \sum_{\ell=1}^{\mathrm{n}} (\boldsymbol{x}_\ell^{(k)} \otimes I_c)[(D_A - \rho A) \otimes \Sigma^{-1}](\boldsymbol{x}_\ell^{(k)T} \otimes I_c) \right). \\
&= \left(\left[\left\{ \frac{1}{\omega_k^2} \right\} \otimes I_{\frac{c}{2}} \otimes E_{q(\Sigma)}(\Sigma^{-1}) \right] + \sum_{\ell=1}^{\mathrm{n}} (\boldsymbol{x}_\ell^{(k)} \otimes I_c)[(D_A - E_{q(\rho)}(\rho A)) \otimes E_{q(\Sigma)}(\Sigma^{-1})](\boldsymbol{x}_\ell^{(k)T} \otimes I_c) \right)
\end{aligned}$$

Also, we can find that:

$$\begin{aligned}
\mathbf{E}_{-k}\big(\boldsymbol{\Sigma}_k^{-1} \underset{\sim}{\mu}_{\mathrm{k}}\big) = \boldsymbol{\Sigma}_{q(W^k)}^{-1} \boldsymbol{\mu}_{q(W_k)} &= \mathbf{E}_{-k}\Big(-\sum_{\ell=1}^{\mathrm{n}} (\boldsymbol{x}_\ell^{(k)} \otimes I_c)[(D_A - \rho A) \otimes \Sigma^{-1}](\boldsymbol{x}_\ell^{(-k)T} \otimes I_c)(\boldsymbol{W}^{(-k)T}) \\
&\quad + \sum_{\ell=1}^{\mathrm{n}} (\boldsymbol{x}_\ell^{(k)} \otimes I_c)[(D_A - \rho A) \otimes \Sigma^{-1}]\boldsymbol{y}_\ell \Big) \\
\Rightarrow \boldsymbol{\Sigma}_{q(W^k)}^{-1} \boldsymbol{\mu}_{q(W_k)} &= \Big(-\sum_{\ell=1}^{\mathrm{n}} (\boldsymbol{x}_\ell^{(k)} \otimes I_c)[(D_A - E_{q(\rho)}(\rho) A) \otimes E_{q(\sigma)}(\Sigma^{-1})](\boldsymbol{x}_\ell^{(-k)T} \otimes I_c)\big(\boldsymbol{\mu}_{q(w^{-k})}\big) \\
&\quad + \sum_{\ell=1}^{\mathrm{n}} (\boldsymbol{x}_\ell^{(k)} \otimes I_c)[(D_A - E_{q(\rho)}(\rho) A) \otimes E_{q(\Sigma)}(\Sigma^{-1})]\boldsymbol{y}_\ell \Big) \\
\Rightarrow \boldsymbol{\mu}_{q(w^k)} &= \boldsymbol{\Sigma}_{q(W^k)} \Big(-\sum_{\ell=1}^{\mathrm{n}} (\boldsymbol{x}_\ell^{(k)} \otimes I_c)[(D_A - E_{q(\rho)}(\rho) A) \otimes E_{q(\sigma)}(\Sigma^{-1})](\boldsymbol{x}_\ell^{(-k)T} \otimes I_c)\big(\boldsymbol{\mu}_{q(w^{-k})}\big) \\
&\quad + \sum_{\ell=1}^{\mathrm{n}} (\boldsymbol{x}_\ell^{(k)} \otimes I_c)[(D_A - E_{q(\rho)}(\rho) A) \otimes E_{q(\Sigma)}(\Sigma^{-1})]\boldsymbol{y}_\ell \Big)
\end{aligned}$$

Then, we can also compute:

$$
\begin{aligned}
E_q[log(q(vec\boldsymbol{W}^{(k)T}))] &= E_q\Bigg[-\frac{1}{2}log|2\pi\Sigma_{q(W^k)}| - \frac{1}{2}\Big(vec(\boldsymbol{W}^{(k)T}) - \boldsymbol{\mu}_{q(W^k)}\Big)^T \Sigma^{-1}_{q(W^k)} \\
&\quad \Big(vec(\boldsymbol{W}^{(k)T}) - \boldsymbol{\mu}_{q(W^k)}\Big)\Bigg] \\
&= -\frac{1}{2}log|2\pi\Sigma_{q(W^k)}| - \frac{1}{2}\boldsymbol{\mu}^T_{q(W^k)}\Sigma^{-1}_{q(W^k)}\boldsymbol{\mu}_{q(W^k)} \\
&+ E_q\Bigg[-\frac{1}{2}(\boldsymbol{W}^{(k)T})^T\boldsymbol{\Sigma}^{-1}_{q(w^k)}(\boldsymbol{W}^{(k)T}) + (\boldsymbol{W}^{(k)T})^T\boldsymbol{\Sigma}^{-1}_{q(w^k)}\boldsymbol{\mu}_{q(w^k)} \Bigg] \\
&= -\frac{1}{2}log|2\pi\Sigma_{q(W^k)}| - \frac{1}{2}\boldsymbol{\mu}^T_{q(W^k)}\Sigma^{-1}_{q(W^k)}\boldsymbol{\mu}_{q(W^k)} \\
&- \frac{1}{2}\Big(\boldsymbol{\mu}^T_{q(W^k)}\boldsymbol{\Sigma}^{-1}_{q(w^k)}\boldsymbol{\mu}_{q(W^k)} + tr(\boldsymbol{\Sigma}^{-1}_{q(W^k)}\boldsymbol{\Sigma}_{q(W^k)})\Big) + \boldsymbol{\mu}^T_{q(W^k)}\boldsymbol{\Sigma}^{-1}_{q(w^k)}\boldsymbol{\mu}_{q(w^k)} \\
&= -\frac{1}{2}log|2\pi\Sigma_{q(W^k)}| - \frac{1}{2}tr(\boldsymbol{\Sigma}^{-1}_{q(W^k)}\boldsymbol{\Sigma}_{q(W^k)})
\end{aligned}
$$

Full conditional distribution of Σ**:**

$$
\begin{aligned}
p(\Sigma\big|\boldsymbol{Y},\boldsymbol{W},\boldsymbol{\omega}^2) &\propto p(\boldsymbol{Y}|\boldsymbol{W},\Sigma)p(\boldsymbol{W}|\Sigma,\boldsymbol{\omega}^2)p(\Sigma) \\
&\propto |(D_A-\rho A)^{-1}\otimes\Sigma|^{-\frac{n}{2}}\exp\Bigg\{ -\frac{1}{2}\sum_{\ell=1}^{n}(\boldsymbol{y}_\ell - \boldsymbol{W}^T\boldsymbol{x}_\ell)^T[(D_A-\rho A)\otimes\Sigma^{-1}] \\
&(\boldsymbol{y}_\ell - \boldsymbol{W}^T\boldsymbol{x}_\ell)\Bigg\} \\
&\times\prod_{\mathrm{i}=1}^{\mathrm{d}}|\omega_i^2\Sigma|^{-\frac{c}{4}}\exp\left\{-\frac{1}{2}\sum_{j^*=1}^{\frac{c}{2}}\tilde{W}^T_{ij^*}\ (\omega_i^2\Sigma)^{-1}\ \tilde{W}_{ij^*}\right\} \\
&\times\frac{|S|^{\frac{v}{2}}}{2^v\Gamma_2(\frac{v}{2})}|\Sigma|^{-\frac{v+3}{2}}\exp\left\{-\frac{1}{2}tr(S\Sigma^{-1})\right\}
\end{aligned}
$$

Denote $y^*_l = y_l - \boldsymbol{W}^T x_l$, $\tilde{y^*}^T_{lj^*} = (y^*_{lj}, y^*_{lj+1})$, $j = 2j^*-1$, $j^* = 1,...,\frac{c}{2}$, $l = 1,...,n$, $B = D_A - \rho A$, then $\dim(B)=\frac{c}{2}\times\frac{c}{2}$. Let $b_{ij} = B[i,j]$, b_{ij} is a scalar,where $i = 1,...,\frac{c}{2}, j = 1,...,\frac{c}{2}$. Using $|E\otimes F| = |E|^n|F|^m$, where $dim(E) = n\times n$ and $dim(F) = m\times m$. $tr(G)+tr(Q) = tr(G+Q)$ where dim(G)=dim(Q) and $tr(JK) = tr(KJ)$ where dim(J)=dim(K^T). This can be simplified as:

$$
\begin{aligned}
p(\Sigma\big|\boldsymbol{Y},\boldsymbol{W},\boldsymbol{\omega}^2) &\propto |D_A-\rho A|^{\frac{nc}{4}}|\Sigma|^{-n}\exp\left\{-\frac{1}{2}tr\left(\sum_{l=1}^{n}\sum_{i=1}^{\frac{c}{2}}\sum_{j=1}^{\frac{c}{2}}b_{ij}\tilde{y^*_{li}}\tilde{y^*_{li}}^T\Sigma^{-1}\right)\right\} \\
&\times\prod_{\mathrm{i}=1}^{\mathrm{d}}|\omega_i^2\Sigma|^{-\frac{c}{4}}\exp\left\{-\frac{1}{2}\sum_{j^*=1}^{\frac{c}{2}}\tilde{W}^T_{ij^*}\ (\omega_i^2\Sigma)^{-1}\ \tilde{W}_{ij^*}\right\} \\
&\times\frac{|S|^{\frac{v}{2}}}{2^v\Gamma_2(\frac{v}{2})}|\Sigma|^{-\frac{v+3}{2}}\exp\left\{-\frac{1}{2}tr(S\Sigma^{-1})\right\}.
\end{aligned}
$$

Since $|D_A - \rho A|$, $\prod_{\mathrm{i}=1}^{\mathrm{d}} |\omega_i^2|^{-\frac{c}{2}}$ and $\frac{|S|^{\frac{v}{2}}}{2^v \Gamma_2(\frac{v}{2})}$ do not depend on Σ, they can be factored out of the expression. This leaves,

$$\begin{aligned} p(\Sigma|\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\omega}^2) \propto |\Sigma|^{-n} \exp\left\{ -\frac{1}{2} tr\left(\sum_{l=1}^{n} \sum_{i=1}^{\frac{c}{2}} \sum_{j=1}^{\frac{c}{2}} b_{ij} \tilde{y_{li}^*} \tilde{y_{li}^*}^T \Sigma^{-1} \right) \right\} \\ \times |\Sigma|^{-\frac{cd}{4}} \exp\left\{ -\frac{1}{2} tr\left(\sum_{i=1}^{d} \sum_{j^*=1}^{\frac{c}{2}} \frac{\tilde{W_{ij^*}} \tilde{W_{ij^*}}^T}{\omega_i^2} \Sigma^{-1} \right) \right\} \\ \times |\Sigma|^{-\frac{v+3}{2}} \exp\left\{ -\frac{1}{2} tr(S\Sigma^{-1}) \right\}. \end{aligned}$$

$$\propto |\Sigma|^{-\frac{2n+\frac{cd}{2}+v+3}{2}} \exp\left\{ -\frac{1}{2} tr\left[\left(\sum_{l=1}^{n} \sum_{i=1}^{\frac{c}{2}} \sum_{j=1}^{\frac{c}{2}} b_{ij} \tilde{y_{li}^*} \tilde{y_{li}^*}^T + \sum_{i=1}^{d} \sum_{j^*=1}^{\frac{c}{2}} \frac{\tilde{W}_{ij^*} \tilde{W}_{ij^*}^T}{\omega_i^2} + S \right) \Sigma^{-1} \right] \right\}$$

Therefore

$$\Sigma \sim Inverse - Wishart(S^*, v^*)$$

Where

$$S^* = \sum_{l=1}^{n} \sum_{i=1}^{\frac{c}{2}} \sum_{j=1}^{\frac{c}{2}} b_{ij} \tilde{y_{li}^*} \tilde{y_{li}^*}^T + \sum_{i=1}^{d} \sum_{j^*=1}^{\frac{c}{2}} \frac{\tilde{W}_{ij^*} \tilde{W}_{ij^*}^T}{\omega_i^2} + S$$

$$v^* = 2n + \frac{cd}{2} + v$$

S^* is a 2×2 matrix and v^* is a scalar. Similarly, we can derive the mean field approximation for Σ based on the full conditional distribution as follows:

$$\begin{aligned} q(\Sigma) &\propto exp\Big\{ \mathbf{E}_{rest}\Big(logP(\Sigma|rest) \Big) \Big\} \\ &\propto exp\Big\{ \mathbf{E}_{rest}\Big(-\frac{2n + \frac{cd}{2} + v + 3}{2} log(|\Sigma|) - \frac{1}{2} tr\Big[\Big(\sum_{l=1}^{n} \sum_{i=1}^{\frac{c}{2}} \sum_{j=1}^{\frac{c}{2}} b_{ij} \tilde{y_{li}^*} \tilde{y_{li}^*}^T \\ &+ \sum_{i=1}^{d} \sum_{j^*=1}^{\frac{c}{2}} \frac{\tilde{W_{ij^*}} \tilde{W_{ij^*}}^T}{\omega_i^2} + S \Big) \Sigma^{-1} \Big] \Big) \Big\} \\ &\propto exp\Big\{ -\frac{2n + \frac{cd}{2} + v + 3}{2} log(|\Sigma|) - \mathbf{E}_{rest}\Big(\frac{1}{2} tr\Big[\Big(\sum_{l=1}^{n} \sum_{i=1}^{\frac{c}{2}} \sum_{j=1}^{\frac{c}{2}} b_{ij} \tilde{y_{li}^*} \tilde{y_{li}^*}^T \\ &+ \sum_{i=1}^{d} \sum_{j^*=1}^{\frac{c}{2}} \frac{\tilde{W_{ij^*}} \tilde{W_{ij^*}}^T}{\omega_i^2} + S \Big) \Sigma^{-1} \Big] \Big) \Big\} \end{aligned}$$

This is still a Inverse-Wishart distribution. That is

$$q(\Sigma) \sim Inverse - Wishart(S_{q(\Sigma)}, v_{q(\Sigma)})$$

Since $B = D_A - \rho A$, then $\dim(B)=\frac{c}{2} \times \frac{c}{2}$. Let $b_{ij} = B[i,j]$, b_{ij} is a scalar,where $i = 1, ..., \frac{c}{2}, j = 1, ..., \frac{c}{2}$. W_{ij} be i^{th} row and j^{th} column of matrix $\mathbf{W}$, $\tilde{W}_{ij^*} = (W_{ij}, W_{ij+1})$, $j = 2j^* - 1$, $j^* = 1, ..., \frac{c}{2}$. Then we can have:

$$E_{rest}(b_{ij}) = [D_A - (\rho A)]_{ij}$$

For $\tilde{w}_{ij^*} = (W_{ij}, W_{ij+1})$, $j = 2j^* - 1$, $j^* = 1, ..., \frac{c}{2}$.

$$E_{q(W)}(\tilde{W}_{ij^*}\tilde{W}_{ij^*}^T) = \begin{bmatrix} E_{q(W)}W_{ij}^2 & E_{q(W)}(W_{ij}W_{ij+1}) \\ E_{q(W)}(W_{ij}W_{ij+1}) & E_{q(W)}W_{ij+1}^2 \end{bmatrix}$$

Now, for $E_{q(W)}W_{ij}^2$ and $E_{q(W)}W_{ij}W_{ij+1}$, we can get that:

$$\begin{aligned} E_{q(W)}W_{ij}^2 &= (\boldsymbol{\mu}_{q(w^i)_j})^2 + \boldsymbol{\Sigma}_{q(w^i)(j,j)} \\ E_{q(W)}W_{ij+1}^2 &= (\boldsymbol{\mu}_{q(w^i)_{j+1}})^2 + \boldsymbol{\Sigma}_{q(w^i)(j+1,j+1)} \\ E_{q(W)}W_{ij}W_{ij+1} &= \boldsymbol{\mu}_{q(w^i)_j}\boldsymbol{\mu}_{q(w^i)_{j+1}} + \boldsymbol{\Sigma}_{q(w^i)(j,j+1)} \end{aligned}$$

$$\begin{aligned} S_{q(\Sigma)} &= E_{rest}\left(\sum_{l=1}^{n}\sum_{i=1}^{\frac{c}{2}}\sum_{j=1}^{\frac{c}{2}} b_{ij}\tilde{y_{li}^*}\tilde{y_{li}^*}^T + \sum_{i=1}^{d}\sum_{j^*=1}^{\frac{c}{2}} \frac{\tilde{W_{ij^*}}\tilde{W_{ij^*}}^T}{\omega_i^2} + S\right) \\ &= \left(\sum_{l=1}^{n}\sum_{i=1}^{\frac{c}{2}}\sum_{j=1}^{\frac{c}{2}} E_q(b_{ij})\tilde{y_{li}^*}\tilde{y_{li}^*}^T + \sum_{i=1}^{d}\sum_{j^*=1}^{\frac{c}{2}} E_q\Big(\tilde{W_{ij^*}}\tilde{W_{ij^*}}^T\Big)E_q\Big(\frac{1}{\omega_i^2}\Big) + S\right) \end{aligned}$$

$$v_{q(\Sigma)} = 2n + \frac{cd}{2} + v$$

Now for $log(q(\Sigma))$:

$$\begin{aligned} E_q(log(q(\Sigma))) &= E_q\left[\frac{v_{q(\Sigma)}}{2}log|S_{q(\Sigma)}| - log(2^{v_{q(\Sigma)}}) - log\Gamma_2(\frac{v_{q(\Sigma)}}{2}) - \frac{v+3}{2}log(|\Sigma|) - \frac{1}{2}tr(S_{q(\Sigma)}\Sigma^{-1})\right] \\ &= \frac{v_{q(\Sigma)}}{2}log|S_{q(\Sigma)}| - log(2^{v_{q(\Sigma)}}) - log\Gamma_2(\frac{v_{q(\Sigma)}}{2}) - E_q\left[\frac{v+3}{2}log(|\Sigma|) - \frac{1}{2}tr(S_{q(\Sigma)}\Sigma^{-1})\right] \end{aligned}$$

Full Conditional of $\boldsymbol{\omega}^2$:

We consider a joint update of the scale mixing variable based on the corresponding full conditional distribution. We have:

$$\begin{aligned} p(\boldsymbol{\omega}^2|\boldsymbol{Y},\boldsymbol{W},\Sigma) &\propto p(\boldsymbol{W}|\Sigma,\boldsymbol{\omega}^2)p(\boldsymbol{\omega}^2|\lambda^2) \\ &\propto \prod_{\mathrm{i}=1}^{\mathrm{d}}(\omega_i^2)^{-\frac{c}{2}}|\Sigma|^{-\frac{c}{4}}\exp\left\{-\frac{1}{2}\sum_{j^*=1}^{\frac{c}{2}}\tilde{w}_{ij^*}^T\ (\omega_i^2\Sigma)^{-1}\ \tilde{w}_{ij^*}\right\} \\ &\times \prod_{\mathrm{i}=1}^{\mathrm{d}}\frac{(\frac{\lambda^2}{2})^{\frac{c+1}{2}}}{\Gamma(\frac{c+1}{2})}(\omega_i^2)^{\frac{c+1}{2}-1}\exp\{-\frac{\lambda^2}{2}\omega_i^2\} \\ &\propto \prod_{\mathrm{i}=1}^{\mathrm{d}}(\omega_i^2)^{-\frac{1}{2}}\exp\left\{-\left(\frac{\lambda^2}{2}\right)\omega_i^2-\frac{c_i^*}{2\omega_i^2}\right\} \end{aligned}$$

where:

$$c_i^* = \sum_{j^*=1}^{\frac{c}{2}}\tilde{W}_{ij^*}^T\Sigma^{-1}\tilde{W}_{ij^*} = tr(\sum_{j^*=1}^{\frac{c}{2}}\tilde{W}_{ij^*}\tilde{W}_{ij^*}^T\Sigma^{-1})$$

The above expression shows that the scale mixing variables are conditionally independent given $\boldsymbol{Y},\boldsymbol{W},\Sigma,\rho,\lambda^2$. We next apply a transformation of variable $\eta_i=(\omega_i^2)^{-1}$, Jacobian $=\left|\frac{d}{d\eta_i}\omega_i^2\right|=\eta_i^{-2}$ which yields:

$$p(\boldsymbol{\omega}^2|\boldsymbol{Y},\boldsymbol{W},\Sigma,\rho,\lambda^2)\propto\prod_{\mathrm{i}=1}^{\mathrm{d}}(\eta_i)^{-\frac{3}{2}}\exp\left\{-\left(\frac{\lambda^2}{2\eta_i}\right)-\frac{\eta_i c_i^*}{2}\right\}$$

and from this we see that the conditional distributions lie within the Inverse Gaussian family. More specifically we have

$$\eta_i=\frac{1}{\omega_{\mathrm{i}}^2}\ \Big|\ \boldsymbol{Y},\boldsymbol{W},\Sigma,\rho,\lambda^2\sim \textit{Inverse-Gaussian}\left(\sqrt{\frac{\lambda^2}{c_i^*}}\ ,\ \ \lambda^2\right),\ \ i=1,\ldots,d.$$

Now, since we already know the full conditional distribution of ω_i, we have:

$$\begin{aligned}
q(\eta_i) &\propto exp\Big\{\mathbf{E}_q\Big(logP(\eta_i|rest)\Big)\Big\} \\
&\propto exp\Big\{\mathbf{E}_q\Big(-\frac{3}{2}log(\eta_i) - \Big(\frac{\lambda^2}{2\omega_i^2}\Big) - \frac{\omega_i^2 c_i^*}{2}\Big)\Big\} \\
&\propto exp\left\{\Big(-\frac{3}{2}log(\eta_i) - \mathbf{E}_{rest}\left(\frac{\lambda^2}{2\eta_i}\right) - \mathbf{E}_q\left(\frac{\eta_i c_i^*}{2}\right)\Big)\right\} \\
&\propto exp\left\{\Big(-\frac{3}{2}log(\eta_i) - \mathbf{E}_{q(\lambda^2)}\left(\frac{\lambda^2}{2}\right)\frac{1}{\eta_i} - \mathbf{E}_{c_i^*}(c_i^*)\left(\frac{\eta_i}{2}\right)\Big)\right\} \\
&\propto exp\left\{\Big(-\frac{3}{2}log(\omega_i^2) - \big(\mu_{q(\lambda^2)}\big)\frac{1}{2\eta_i} - \mathbf{E}_{c_i^*}(c_i^*)\left(\frac{\eta_i}{2}\right)\Big)\right\}
\end{aligned}$$

Therefore, $q(\eta_i)$ is still an Inverse-Gaussian distribution with

$$\mu_{q(\eta_i)} = \sqrt{\frac{\mu_{q(\lambda^2)}}{E_{c_i^*}(c_i^*)}}$$

$$\lambda_{q(\eta_i)} = E_{q(\lambda^2)}(\lambda^2) = \mu_{q(\lambda^2)}$$

Now, since η_i is an Inverse Gaussian, then $\omega_i^2 = 1/\eta_i$ will be a reciprocal of inverse gaussian. where we have:

$$\begin{aligned}
\mu_{q(\omega_i^2)} &= \frac{1}{\mu_{q(\eta_i)}} + \frac{1}{\lambda_{q(\eta_i)}} \\
Var_{q(\omega_i^2)} &= \frac{1}{\mu_{q(\eta_i)}\lambda_{q(\eta_i)}} + \frac{2}{\lambda_{q(\eta_i)}^2}
\end{aligned}$$

For $E_q(log(q(\omega_i^2)))$, we can compute as:

$$\begin{aligned}
E_q(log(q(\omega_i^2))) &= E_q\left[\frac{1}{2}\Big(log(\lambda_{q_{(\eta_i)}}) - log(2\pi) - log(\omega_i^2)\Big) - \frac{\lambda_{q(\eta_i)}(1-\omega_i^2\mu_{q(\eta_i)})^2}{2\mu_{q(\eta_i)}^2\omega_i^2}\right] \\
&= \frac{1}{2}\Big(log(\lambda_{q_{(\eta_i)}}) - log(2\pi)\Big) - E_q\left[log(\omega_i^2)\right] - \lambda_{q_{(\eta_i)}}E_q\left[\frac{1}{2\mu_{q_{(\eta_i)}}^2}\left(\frac{1}{\omega_i^2}\right) - \frac{1}{\mu_{q_{(\eta_i)}}^2} + \frac{\omega_i^2}{2}\right]
\end{aligned}$$

For $E_q(log(\omega_i^2))$, we can use Taylor series to approximate as:

$$E_q(log(\omega_i^2)) = log(\mu_{q(\omega_i^2)}) - \frac{1}{2\mu_{\omega_i^2}}Var_q[\omega_i^2]$$

Then, we can have:

$$\begin{aligned}
E_q(log(q(\omega_i^2))) &= \frac{1}{2}\Big(log(\lambda_{q(\eta_i)}) - log(2\pi)\Big) - log(\mu_{q(\omega_i^2)}) - \frac{1}{2\mu_{\omega_i^2}} Var_q[\omega_i^2] \\
&\quad - \lambda_{q(\eta_i)} E_q\left[\frac{1}{2\mu_{q(\eta_i)}^2}\left(\frac{1}{\omega_i^2}\right) - \frac{1}{\mu_{q(\eta_i)}^2} + \frac{\omega_i^2}{2}\right] \\
&= \frac{1}{2}\Big(log(\lambda_{q(\eta_i)}) - log(2\pi)\Big) - log(\mu_{q(\omega_i^2)}) - \frac{1}{2\mu_{\omega_i^2}} Var_q[\omega_i^2] \\
&\quad - \lambda_{q(\eta_i)}\left[\frac{1}{2\mu_{q(\eta_i)}^2}\mu_{q(\eta_i)} - \frac{1}{\mu_{q(\eta_i)}^2} + \frac{\mu_{q(\omega_i^2)}}{2}\right]
\end{aligned}$$

Therefore, the posterior distribution can be approximated by mean field variational bayes as:

$$P(\boldsymbol{\Theta}|Y) \approx \prod_{k=1}^{d}\Big[q_{\boldsymbol{W}^{(k)}}(\boldsymbol{W}^{(k)})\Big]\prod_{i=1}^{d}\Big[q_{\omega_i^2}(\omega_i^2)\Big]q_{\Sigma}(\Sigma)$$

where:

$$\begin{aligned}
q_{\boldsymbol{W}^{(k)}}(\boldsymbol{W}^{(k)}) &\equiv MVN(\boldsymbol{\mu}_{q_{\boldsymbol{W}^{(k)}}}, \boldsymbol{\Sigma}_{q_{\boldsymbol{W}^{(k)}}}) \\
q_{\omega_i^2}(\omega_i^2) &\equiv \text{reciprocal of Inverse Gaussian}(\mu_{q(\eta_i)}, \lambda_{q(\eta_i)}) \\
q_{\Sigma}(\Sigma) &\equiv Inverse - Wishart(S_{q(\Sigma)}, v_{q(\Sigma)})
\end{aligned}$$

Lower Bound $\mathcal{L}(q)$ for Variational bayes.

We now have derived the optimal q distributions. The logarithm lower bound takes following explicit form:

$$\begin{aligned}
\mathcal{L}(q) &= E_q[log(P(\boldsymbol{Y}, \boldsymbol{\theta}))] - E_q[log(q(\boldsymbol{\theta}))] \\
&= E_q\big[log(p(\boldsymbol{Y}|\boldsymbol{W}, \Sigma, \rho)) + log(p(\boldsymbol{W}|\Sigma, \boldsymbol{\omega}^2)) + log(p(\boldsymbol{\omega}^2|\lambda^2)) + log(p(\Sigma))\big] - E_q[log(q(\boldsymbol{\theta}))]
\end{aligned}$$

Now, taking the expectation with respect to q for each component in above, we have:

$$\begin{aligned}
E_q(log(p(\boldsymbol{Y}|\boldsymbol{W},\Sigma,\rho))) &= E_q\bigg[-\frac{n}{2}log|(D_A-\rho A)^{-1}\otimes\Sigma| - \frac{1}{2}\sum_{\ell=1}^{n}(\boldsymbol{y}_\ell-\boldsymbol{W}^T\boldsymbol{x}_\ell)^T[(D_A-\rho A)\otimes\Sigma^{-1}] \\
&\quad (\boldsymbol{y}_\ell-\boldsymbol{W}^T\boldsymbol{x}_\ell)\bigg] \\
&= -\frac{n}{2}log|(D_A-\mu_{q(\rho)}A)^{-1}\otimes\mu_{q(\Sigma)}| - E_q\bigg[\frac{1}{2}\sum_{\ell=1}^{n}(\boldsymbol{y}_\ell-\boldsymbol{W}^T\boldsymbol{x}_\ell)^T[(D_A-\rho A) \\
&\quad \otimes\Sigma^{-1}](\boldsymbol{y}_\ell-\boldsymbol{W}^T\boldsymbol{x}_\ell)\bigg] \\
E_q(log(p(\boldsymbol{W}|\Sigma,\boldsymbol{\omega}^2))) &= E_q\bigg[\sum_{\mathrm{i}=1}^{\mathrm{d}}\sum_{j=1}^{\frac{c}{2}}\bigg(-\frac{1}{2}log|\omega_i^2\Sigma| - \frac{1}{2}\tilde{w}_{ij}^T(\omega_i^2\Sigma)^{-1}\tilde{w}_{ij}\bigg)\bigg] \\
&= \bigg[\sum_{\mathrm{i}=1}^{\mathrm{d}}\sum_{j=1}^{\frac{c}{2}} -\frac{1}{2}log|\mu_{q(\omega_i^2)}\mu_{q(\Sigma)}| - E_q\bigg(\frac{1}{2}\tilde{w}_{ij}^T(\omega_i^2\Sigma)^{-1}\tilde{w}_{ij}\bigg)\bigg]
\end{aligned}$$

$$\begin{aligned}
E_q(log(p(\boldsymbol{\omega}^2|\lambda^2))) &= E_q\bigg[\sum_{\mathrm{i}=1}^{\mathrm{d}}\Big(\frac{c+1}{2}log(\frac{\lambda^2}{)} - log(\Gamma(\frac{c+1}{2})) + (\frac{c+1}{2}-1)log(\omega_i^2) - \frac{\lambda^2}{2}\omega_i^2\Big)\bigg] \\
&= \bigg[\sum_{\mathrm{i}=1}^{\mathrm{d}}\Big(\frac{c+1}{2}E_q(log(\lambda^2)-log(2)) - log(\Gamma(\frac{c+1}{2})) + (\frac{c+1}{2}-1)E_q(log(\omega_i^2)) \\
&\quad -\frac{1}{2}E_q(\lambda^2)E_q(\omega_i^2)\Big)\bigg]
\end{aligned}$$

$$\begin{aligned}
E_q(log(p(\Sigma))) &= E_q\bigg[const - (\frac{v+3}{2})log|\Sigma| - \frac{1}{2}tr(S\Sigma^{-1})\bigg] \\
&= \bigg[const - (\frac{v+3}{2})log|\mu_{q(\Sigma)}| - \frac{1}{2}tr(S\mu_{q(\Sigma^{-1})})\bigg]
\end{aligned}$$

Now, let's take a look at the $E_q[log(q(\boldsymbol{\theta}))]$, which could be written as:

$$\begin{aligned}
E_q[log(q(\boldsymbol{\theta}))] &= E_q[log(q(\boldsymbol{W}))] + E_q[log(q(\boldsymbol{\omega}^2)] + E_q[log(q(\Sigma))] \\
&= \sum_{k=1}^{K}E_q[log(q(vec(\boldsymbol{W}^{(k)T})))] + \sum_{i=1}^{d}E_q[log(q(\omega_i^2)] + E_q[log(q(\Sigma))]
\end{aligned}$$

Bibliography

[1] Akaike, H. Information theory and the maximum likelihood principle in 2nd international symposium on information theory (bn petrov and f. csäki, eds.). *Akademiai Kiàdo, Budapest* (1973).

[2] Arndt, J., and Reder, L. M. The effect of distinctive visual information on false recognition. *Journal of Memory and Language 48*, 1 (2003), 1–15.

[3] Aydin, Ü., Rampp, S., Wollbrink, A., Kugel, H., Cho, J.-H., Knösche, T. R., Grova, C., Wellmer, J., and Wolters, C. Zoomed MRI guided by combined EEG/MEG source analysis: a multimodal approach for optimizing presurgical epilepsy work-up and its application in a multi-focal epilepsy patient case study. *Brain Topography 30*, 4 (2017), 417–433.

[4] Baars, B., and Gage, N. M. *Fundamentals of cognitive neuroscience: a beginner's guide*. Academic Press, (2012).

[5] Baillet, S. Forward and inverse problems of MEG/EEG. *Encyclopedia of Computational Neuroscience* (2015), 1226–1233.

[6] Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language 68*, 3 (2013), 255–278.

[7] Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. Parsimonious mixed models. *arXiv preprint arXiv:1506.04967* (2015).

[8] Bates, D., Mächler, M., Bolker, B., and Walker, S. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823* (2014).

[9] Belaoucha, B., and Papadopoulo, T. Spatial regularization based on dMRI to solve EEG/MEG inverse problem. In *EMBC 2017-39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2017).

[10] Berger, J., Bayarri, M., and Pericchi, L. The effective sample size. *Econometric Reviews 33*, 1-4 (2014), 197–217.

[11] Berger, J. O., and Berry, D. A. The relevance of stopping rules in statistical inference. *Statistical Decision Theory and Related Topics IV 1* (1988), 29–47.

[12] Bertram, L., McQueen, M. B., Mullin, K., Blacker, D., and Tanzi, R. E. Systematic meta-analyses of alzheimer disease genetic association studies: the alzgene database. *Nature Genetics 39*, 1 (2007), 17.

[13] Besag, J. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)* (1986), 259–302.

[14] Bowman, F. D. Spatio-temporal modeling of localized brain activity. *Biostatistics 6*, 4 (2005), 558–575.

[15] Bowman, F. D., Caffo, B., Bassett, S. S., and Kilts, C. A bayesian hierarchical framework for spatial modeling of fMRI data. *Neuroimage 39*, 1 (2008), 146–156.

[16] Burnham, K., and Anderson, D. *Model Selection and Inference: A Practical Informationtheoretic Approach*. Springer, Berlin, Heidelberg, New York, (1998).

[17] Calvetti, D., Pascarella, A., Pitolli, F., Somersalo, E., and Vantaggi, B. A hierarchical Krylov–Bayes iterative inverse solver for MEG with physiological preconditioning. *Inverse Problems 31*, 12 (2015), 125005.

[18] Carlin, B. P., and Louis, T. A. *Bayesian Methods for Data Analysis*. CRC Press, (2008).

[19] Chateau, D., and Jared, D. Spelling–sound consistency effects in disyllabic word naming. *Journal of Memory and Language 48*, 2 (2003), 255–280.

[20] Chen, M.-H. Computing marginal likelihoods from a single mcmc output. *Statistica Neerlandica 59*, 1 (2005), 16–29.

[21] Chib, S., and Jeliazkov, I. Marginal likelihood from the metropolis–hastings output. *Journal of the American Statistical Association 96*, 453 (2001), 270–281.

[22] Clark, H. H. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behaviour 12*, 4 (1973), 335–359.

[23] Cochran, W. G. The analysis of variance when experimental errors follow the Poisson or binomial laws. *The Annals of Mathematical Statistics 11*, 3 (1940), 335–347.

[24] Cohen, J. The earth is round ($p < 0.5$). *American Psychologist 49*, 12 (1994), 997–1003.

[25] Corder, E. H., Saunders, A. M., Strittmatter, W. J., Schmechel, D. E., Gaskell, P. C., Small, G., Roses, A. D., Haines, J., and Pericak-Vance, M. A. Gene dose of apolipoprotein e type 4 allele and the risk of alzheimer's disease in late onset families. *Science 261*, 5123 (1993), 921–923.

[26] Cumming, G. Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science 3*, 4 (2008), 286–300.

[27] Daunizeau, J., and Friston, K. J. A mesostate-space model for EEG and MEG. *Neuroimage 38*, 1 (2007), 67–81.

[28] Derado, G., Bowman, F. D., Zhang, L., and Initiative, A. D. N. Predicting brain activity using a Bayesian spatial model. *Statistical Methods in Medical Research 22*, 4 (2013), 382–397.

[29] Dixon, P. The p-value fallacy and how to avoid it. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale 57*, 3 (2003), 189.

[30] Dixon, P. Models of accuracy in repeated-measures designs. *Journal of Memory and Language 59*, 4 (2008), 447–456.

[31] Efron, B. Why isn't everyone a Bayesian? *The American Statistician 40*, 1 (1986), 1–5.

[32] Gatz, M., Reynolds, C. A., Fratiglioni, L., Johansson, B., Mortimer, J. A., Berg, S., Fiske, A., and Pedersen, N. L. Role of genes and environments for explaining alzheimer disease. *Archives of General Psychiatry 63*, 2 (2006), 168–174.

[33] Ge, T., Feng, J., Hibar, D. P., Thompson, P. M., and Nichols, T. E. Increasing power for voxel-wise genome-wide association studies: the random field theory, least square kernel machines and fast permutation procedures. *Neuroimage 63*, 2 (2012), 858–873.

[34] Ge, T., Müller-Lenke, N., Bendfeldt, K., Nichols, T. E., and Johnson, T. D. Analysis of multiple sclerosis lesions via spatially varying coefficients. *The Annals of Applied Statistics 8*, 2 (2014), 1095.

[35] Gelfand, A. E., and Banerjee, S. Multivariate spatial process models. *Handbook of Spatial Statistics* (2010), 495–515.

[36] Gelfand, A. E., and Vounatsou, P. Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics 4*, 1 (2003), 11–15.

[37] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. *Bayesian Data Analysis*, vol. 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

[38] Gelman, A., Hwang, J., and Vehtari, A. Understanding predictive information criteria for bayesian models. *Statistics and Computing 24*, 6 (2014), 997–1016.

[39] Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* (2008), 1360–1383.

[40] Gelman, A., and Meng, X.-L. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science* (1998), 163–185.

[41] Greenlaw, K., Szefer, E., Graham, J., Lesperance, M., Nathoo, F. S., and Initiative, A. D. N. A Bayesian group sparse multi-task regression model for imaging genetics. *Bioinformatics 33*, 16 (2017), 2513–2522.

[42] Hadamard, J. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin* (1902), 49–52.

[43] Haller, H., and Krauss, S. Misinterpretations of significance: A problem students share with their teachers. *Methods of Psychological Research 7*, 1 (2002), 1–20.

[44] Helmholtz, H. v. Ueber einige gesetze der vertheilung elektrischer ströme in körperlichen leitern, mit anwendung auf die thierisch-elektrischen versuche (schluss.). *Annalen der Physik 165*, 7 (1853), 353–377.

[45] Henson, R., Flandin, G., Friston, K. J., and Mattout, J. A parametric empirical Bayesian framework for fMRI-constrained MEG/EEG source reconstruction. *Human Brain Mapping 31*, 10 (2010), 1512–1531.

[46] Henson, R., Goshen-Gottstein, Y., Ganel, T., Otten, L., Quayle, A., and Rugg, M. Electrophysiological and haemodynamic correlates of face perception, recognition and priming. *Cerebral Cortex 13*, 7 (2003), 793–805.

[47] Henson, R., Mattout, J., Phillips, C., and Friston, K. J. Selecting forward models for MEG source-reconstruction using model-evidence. *Neuroimage 46*, 1 (2009a), 168–176.

[48] Henson, R., Mattout, J., Singh, K., Barnes, G., Hillebrand, A., and Friston, K. Population-level inferences for distributed MEG source localization under multiple constraints: application to face-evoked fields. *Neuroimage 38*, 3 (2007), 422–438.

[49] Henson, R., Mouchlianitis, E., and Friston, K. J. MEG and EEG data fusion: simultaneous localisation of face-evoked responses. *Neuroimage 47*, 2 (2009b), 581–589.

[50] Hibar, D. P., Stein, J. L., Kohannim, O., Jahanshad, N., Saykin, A. J., Shen, L., Kim, S., Pankratz, N., Foroud, T., Huentelman, M. J., et al. Voxelwise gene-wide association study (vgenewas): multivariate gene-based association testing in 731 elderly subjects. *Neuroimage 56*, 4 (2011), 1875–1891.

[51] Hibar, D. P., Stein, J. L., Renteria, M. E., Arias-Vasquez, A., Desrivières, S., Jahanshad, N., Toro, R., Wittfeld, K., Abramovic, L., Andersson, M., et al. Common genetic variants influence human subcortical brain structures. *Nature 520*, 7546 (2015), 224.

[52] Huang, C., Thompson, P., Wang, Y., Yu, Y., Zhang, J., Kong, D., Colen, R. R., Knickmeyer, R. C., Zhu, H., Initiative, A. D. N., et al. Fgwas: Functional genome wide association analysis. *Neuroimage 159* (2017), 107–121.

[53] Huang, M., Nichols, T., Huang, C., Yu, Y., Lu, Z., Knickmeyer, R. C., Feng, Q., Zhu, H., Initiative, A. D. N., et al. Fvgwas: Fast voxelwise genome wide association analysis of large-scale imaging genetic data. *Neuroimage 118* (2015), 613–627.

[54] Inkster, B., Nichols, T. E., Saemann, P. G., Auer, D. P., Holsboer, F., Muglia, P., and Matthews, P. M. Pathway-based approaches to imaging genetics association studies: Wnt signaling, gsk3beta substrates and major depression. *Neuroimage 53*, 3 (2010), 908–917.

[55] Ismail, S., Sun, W., Nathoo, F. S., Babul, A., Moiseev, A., Beg, M. F., and Virji-Babul, N. A skew-t space-varying regression model for the spectral analysis of resting state brain activity. *Statistical Methods in Medical Research 22*, 4 (2013), 424–438.

[56] Jacoby, L. L., Wahlheim, C. N., and Kelley, C. M. Memory consequences of looking back to notice change: Retroactive and proactive facilitation. *Journal of Experimental Psychology: Learning, Memory, and Cognition 41*, 5 (2015), 1282.

[57] Jaeger, T. F. Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of Memory and Language 59*, 4 (2008), 434–446.

[58] Jin, X., Carlin, B. P., and Banerjee, S. Generalized hierarchical multivariate car models for areal data. *Biometrics 61*, 4 (2005), 950–961.

[59] John, B., and Lewis, K. R. Chromosome variability and geographic distribution in insects. *Science 152*, 3723 (1966), 711–721.

[60] Johnson, T. D., Liu, Z., Bartsch, A. J., and Nichols, T. E. A Bayesian non-parametric Potts model with application to pre-surgical fMRI data. *Statistical Methods in Medical Research 22*, 4 (2013), 364–381.

[61] Jones, R. H. Bayesian information criterion for longitudinal and clustered data. *Statistics in Medicine 30*, 25 (2011), 3050–3056.

[62] Kass, R. E., and Raftery, A. E. Bayes factors. *Journal of the American Statistical Association 90*, 430 (1995), 773–795.

[63] Kruschke, J. K. Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General 142*, 2 (2013), 573.

[64] Kyung, M., Gill, J., Ghosh, M., Casella, G., et al. Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis 5*, 2 (2010), 369–411.

[65] Lim, M., Ales, J. M., Cottereau, B. R., Hastie, T., and Norcia, A. M. Sparse EEG/MEG source estimation via a group lasso. *PLoS One 12*, 6 (2017), e0176835.

[66] Lin, X. Variance component testing in generalised linear models with random effects. *Biometrika* (1997), 309–326.

[67] Liu, J., and Calhoun, V. D. A review of multivariate analyses in imaging genetics. *Frontiers in Neuroinformatics 8* (2014), 29.

[68] Long, C. J., Purdon, P. L., Temereanca, S., Desai, N. U., Hämäläinen, M. S., and Brown, E. N. State-space solutions to the dynamic magnetoencephalography inverse problem using high performance computing. *The Annals of Applied Statistics 5*, 2B (2011), 1207.

[69] Lu, Z.-H., Khondker, Z., Ibrahim, J. G., Wang, Y., Zhu, H., Initiative, A. D. N., et al. Bayesian longitudinal low-rank regression models for imaging genetic data from longitudinal studies. *Neuroimage 149* (2017), 305–322.

[70] Lucotte, G., Loirat, F., and Hazout, S. Pattern of gradient of apolipoprotein e allele* 4 frequencies in western europe. *Human Biology* (1997), 253–262.

[71] MacNab, Y. C. Linear models of coregionalization for multivariate lattice data: a general framework for coregionalized multivariate car models. *Statistics in Medicine 35*, 21 (2016), 3827–3850.

[72] Masson, M. E. A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods 43*, 3 (2011), 679–690.

[73] Matsuura, K., and Okabe, Y. Selective minimum-norm solution of the biomagnetic inverse problem. *IEEE Transactions on Biomedical Engineering 42*, 6 (1995), 608–615.

[74] McCulloch, C. E., and Neuhaus, J. M. Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Statistical Science* (2011), 388–402.

[75] Meng, X.-L., and Schilling, S. Warp bridge sampling. *Journal of Computational and Graphical Statistics 11*, 3 (2003), 552–586.

[76] Meng, X.-L., and Wong, W. H. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica* (1996), 831–860.

[77] Moores, M. T., Pettitt, A. N., and Mengersen, K. Scalable Bayesian inference for the inverse temperature of a hidden potts model. *arXiv preprint arXiv:1503.08066* (2015).

[78] Morris, J. S., Brown, P. J., Herrick, R. C., Baggerly, K. A., and Coombes, K. R. Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics 64*, 2 (2008), 479–489.

[79] Nathoo, F. S., Babul, A., Moiseev, A., Virji-Babul, N., and Beg, M. A variational bayes spatiotemporal model for electromagnetic brain mapping. *Biometrics 70*, 1 (2014), 132–143.

[80] Nathoo, F. S., and Ghosh, P. Skew-elliptical spatial random effect modeling for areal data with application to mapping health utilization rates. *Statistics in Medicine 32*, 2 (2013), 290–306.

[81] Nathoo, F. S., Greenlaw, K., and Lesperance, M. Regularization parameter selection for a bayesian group sparse multi-task regression model with application to imaging genomics. In *Pattern Recognition in Neuroimaging (PRNI), 2016 International Workshop on* (2016), IEEE, pp. 1–4.

[82] Nathoo, F. S., Kilshaw, R. E., and Masson, M. A better (Bayesian) interval estimate for within-subject designs. *Journal of Mathematical Psychology 86* (2018), 1–9.

[83] Nathoo, F. S., Kong, L., and Zhu, H. A review of statistical methods in imaging genetics. *Canadian Journal of Statistics* (2018), DOI: 10.1002/cjs.11487.

[84] Nathoo, F. S., Lesperance, M. L., Lawson, A. B., and Dean, C. B. Comparing variational Bayes with Markov chain Monte Carlo for Bayesian computation in neuroimaging. *Statistical Methods in Medical Research 22*, 4 (2013), 398–423.

[85] Nathoo, F. S., and Masson, M. E. Bayesian alternatives to null-hypothesis significance testing for repeated-measures designs. *Journal of Mathematical Psychology 72* (2016), 144–157.

[86] Ormerod, J. T., and Wand, M. P. Explaining variational approximations. *The American Statistician 64*, 2 (2010), 140–153.

[87] Park, T., and Casella, G. The Bayesian lasso. *Journal of the American Statistical Association 103*, 482 (2008), 681–686.

[88] Pascual-Marqui, R. D., Michel, C. M., and Lehmann, D. Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. *International Journal of Psychophysiology 18*, 1 (1994), 49–65.

[89] Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., and Nichols, T. E. *Statistical parametric mapping: The analysis of functional brain images*. Academic Press, (2011).

[90] Penny, W. D., Trujillo-Barreto, N. J., and Friston, K. J. Bayesian fMRI time series analysis with spatial priors. *Neuroimage 24*, 2 (2005), 350–362.

[91] Plummer, M. Jags: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (2003), vol. 124, Technische Universit at Wien Wien, Austria, p. 125.

[92] Plummer, M. rjags: Bayesian graphical models using mcmc. *R package version 3* (2013).

[93] Potts, R. B. Some generalized order-disorder transformations. In *Mathematical Proceedings of the Cambridge Philosophical Society* (1952), vol. 48, Cambridge University Press, pp. 106–109.

[94] Quené, H., and Van den Bergh, H. Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language 59*, 4 (2008), 413–425.

[95] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. ISBN 3-900051-07-0.

[96] Raftery, A. E., Newton, M. A., Satagopan, J. M., and Krivitsky, P. N. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian Statistics 8* (2007), 1–45.

[97] Ramírez, R. R. Source localization. *Scholarpedia 3*, 11 (2008), 1733.

[98] Ren, Q., Banerjee, S., Finley, A. O., and Hodges, J. S. Variational bayesian methods for spatial data analysis. *Computational Statistics & Data Analysis 55*, 12 (2011), 3197–3217.

[99] Rouder, J. N., Morey, R. D., Speckman, P. L., and Province, J. M. Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology 56*, 5 (2012), 356–374.

[100] Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review 16*, 2 (2009), 225–237.

[101] Sarvas, J. Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem. *Physics in Medicine & Biology 32*, 1 (1987), 11.

[102] Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., and McCulloch, R. E. Bayes and big data: The consensus monte carlo algorithm. *International Journal of Management Science and Engineering Management 11*, 2 (2016), 78–88.

[103] Shen, X., Papademetris, X., and Constable, R. T. Graph-theory based parcellation of functional subunits in the brain from resting-state fmri data. *Neuroimage 50*, 3 (2010), 1027–1035.

[104] Silver, M., Montana, G., Nichols, T. E., and Initiative, A. D. N. False positives in neuroimaging genetics using voxel-based morphometry data. *Neuroimage 54*, 2 (2011), 992–1000.

[105] Sorrentino, A., and Piana, M. Inverse modeling for MEG/EEG data. In *Mathematical and Theoretical Neuroscience*. Springer, (2017), pp. 239–253.

[106] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*, 4 (2002), 583–639.

[107] Stein, J. L., Hua, X., Lee, S., Ho, A. J., Leow, A. D., Toga, A. W., Saykin, A. J., Shen, L., Foroud, T., Pankratz, N., et al. Voxelwise genome-wide association study (vgwas). *Neuroimage 53*, 3 (2010), 1160–1174.

[108] Stingo, F. C., Guindani, M., Vannucci, M., and Calhoun, V. D. An integrative bayesian modeling approach to imaging genetics. *Journal of the American Statistical Association 108*, 503 (2013), 876–891.

[109] Szefer, E., Lu, D., Nathoo, F. S., Beg, M. F., Graham, J., and Initiative, A. D. N. Multivariate association between single-nucleotide polymorphisms in alzgene linkage regions and structural changes in the brain: discovery, refinement and validation. *Statistical Applications in Genetics and Molecular Biology 16*, 5-6 (2017), 367–386.

[110] Teng, M., Johnson, T. D., and Nathoo, F. S. Time series analysis of fMRI data: Spatial modelling and Bayesian computation. *Statistics in Medicine* (2018a).

[111] Teng, M., Nathoo, F. S., and Johnson, T. D. Bayesian analysis of functional magnetic resonance imaging data with spatially varying auto-regressive orders. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* (2018b).

[112] Thompson, P. M., Ge, T., Glahn, D. C., Jahanshad, N., and Nichols, T. E. Genetics of the connectome. *Neuroimage 80* (2013), 475–488.

[113] Tian, T. S., and Li, Z. A spatio-temporal solution for the eeg/meg inverse problem using group penalization methods. *Statistics and its Interface 4*, 4 (2011), 521–533.

[114] Vehtari, A., Gelman, A., and Gabry, J. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing 27*, 5 (2017), 1413–1432.

[115] Vivaldi, V., and Sorrentino, A. Bayesian smoothing of dipoles in magneto-/electroencephalography. *Inverse Problems 32*, 4 (2016), 045007.

[116] Vounou, M., Nichols, T. E., Montana, G., Initiative, A. D. N., et al. Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *Neuroimage 53*, 3 (2010), 1147–1159.

[117] Wagenmakers, E.-J. A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review 14*, 5 (2007), 779–804.

[118] Wang, H., Nie, F., Huang, H., Risacher, S. L., Saykin, A. J., Shen, L., and Initiative, A. D. N. Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics 28*, 12 (2012), 127–136.

[119] Watanabe, S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research 11*, Dec (2010), 3571–3594.

[120] Wipf, D., and Nagarajan, S. A unified Bayesian framework for MEG/EEG source imaging. *Neuroimage 44*, 3 (2009), 947–966.

[121] Yap, M. J., Balota, D. A., Tse, C.-S., and Besner, D. On the additive effects of stimulus quality and word frequency in lexical decision: Evidence for opposing interactive influences revealed by rt distributional analyses. *Journal of Experimental Psychology: Learning, Memory, and Cognition 34*, 3 (2008), 495.

[122] Zakharova, T., Karpov, P., and Bugaevskii, V. Localization of the activity source in the inverse problem of magnetoencephalography. *Computational Mathematics and Modeling 28*, 2 (2017), 148–157.

[123] Zhu, H., Khondker, Z., Lu, Z., and Ibrahim, J. G. Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. *Journal of the American Statistical Association 109*, 507 (2014), 977–990.