Scalable Bayesian Methods for the Analysis of Neuroimaging Data

Anna Menacher

University College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Hilary 2024

Abstract

The recent surge in large-scale population health datasets, such as the UK Biobank or the Adolescent Brain Cognitive Development (ABCD) study, requires the development of scalable statistical methods that are capable of analysing the rich multitude of data sources. This thesis focuses on the scalable analysis of Magnetic Resonance Imaging (MRI) neuroimaging data, such as binary lesion masks and task-based functional Magnetic Resonance Imaging (fMRI). In particular, we introduce two Bayesian spatial models with sparsity priors on the spatially varying coefficients and extend our work to suit the large sample sizes found in population health studies.

Firstly, we propose a scalable hierarchical Bayesian image-on-scalar regression model, called BLESS, capable of handling binary responses and of placing continuous spike-and-slab mixture priors on spatially varying parameters. Thereby, enforcing spatial dependency on the parameter dictating the amount of sparsity within the probability of inclusion. The use of mean-field variational inference with dynamic posterior exploration, which is an annealing-like strategy that improves optimisation, allows our method to scale to large sample sizes. We validate our results via simulation studies and an application to binary lesion masks from the UK Biobank.

Secondly, we extend our method to account for underestimation of posterior variance due to variational inference by providing an approximate posterior sampling approach inspired by Bayesian bootstrap ideas and spike-and-slab priors with random shrinkage targets. Besides accurate uncertainty quantification, this approach is capable of producing novel cluster size-based imaging statistics, such as credible intervals of cluster size, and measures of reliability of cluster occurrence.

Thirdly, we develop a Bayesian nonparametric scalar-on-image regression model with a relaxed-thresholded Gaussian process prior on the spatially varying coefficients in order to introduce sparsity and smoothness into the model. Our main contribution is the improved scalability, allowing for larger sample sizes and bigger image dimensions, which is made possible by replacing posterior sampling with a variational approximation. We validate our results via simulation studies and an application to cortical surface task-based fMRI data from the ABCD study.

Scalable Bayesian Methods for the Analysis of Neuroimaging Data

Anna Menacher

University College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2024

Acknowledgements

Personal

First and foremost, I would like to thank Professor Thomas E. Nichols and Professor Chris Holmes for their time and dedication in the supervision of my thesis. To Tom, thank you for teaching me the complexities and beauty of neuroimaging data. To Chris, I extend my thanks for always inspiring me with your passion and ideas for combining medical applications with sound statistical methods. Besides my supervisors, I am grateful for my collaborators, Professor Jian Kang and Professor Timothy Johnson, from the University of Michigan for letting me extend their work and dedicating their time to the last project of my PhD. I am also thankful for my time spent with Kajsa Kvist and her team at Novo Nordisk in Denmark which has shown me that good research can be pursued in many forms. Thanks also to my undergraduate lecturer, Professor Ralf Kellner, who I have seen as a first inspiration to pursue a PhD in Statistics and who has been a mentor ever since.

To my family, Mama, Papa, and Lisa, I am ever so grateful to have you as my constant support. May that be through our weekly Skype calls or just by knowing that I can always come home to the comfort of being surrounded by family. This gratitude of course also extends to the rest of my family who always welcome me back home.

To all the friends across the world, who bring so much joy into my life, I am so grateful to be your friend and thankful that you are part of my life and have supported me through the ups and downs of my PhD pursuit. To Samvida, we became flatmates by surprise but I would not want to imagine it any other way. I am so grateful that we have shared so many breakfast chats along the years, found our shared love for dumplings, and have a friendship that allows us to chat about science and everything else alike. To Laura, thank you for always entertaining an adventure, cheering me on, being an excellent researcher and friend alike, and to both Laura and Alex, for always hosting me in your Cotswold cottage, enjoying good food and music with me, and being the best climbing buddies to share my time in Oxford with. To Cori and Alice, I am so grateful for your friendship across our years in Oxford and will always remember all the lockdown walks, tea chats, and board game afternoons. On that note, I want to also thank all past flatmates, the remainder of the Edinburgh Werewolves, and all other friends that have supported me or have taken me on incredible travel adventures.

Last but not least, I am thankful for the wonderful people at the Statistics Department and at the Big Data Institute who have enriched my academic experience and have always understood the emotional roller coaster of pursuing a PhD. A special thanks to my office mate Sahra who has always supported me with her kindness. I am also grateful for Bernd, Sahra, Edwin, and Lorenzo who have provided me with valuable feedback for various parts of this thesis.

Institutional

I would also like to thank all my funding sources that have helped me to pursue my further education, present my research at conferences, and learn more about how to use statistics for medical applications. Therefore, I am grateful to the financial support of the EPSRC Research Council, Novartis, and the Oxford Radcliffe Graduate Scholarship for fully funding my postgraduate research position in the StatML CDT and for University College at the University of Oxford for continuously supporting my international travel to conferences.

Previously Published Material

I am grateful to have collaborated with my supervisors and other more senior researchers from the University of Oxford and the University of Michigan on many projects that are now included in this thesis. I therefore want to acknowledge that research is collaborative in nature and hence my usage of “I” and “we” throughout this thesis can be seen as interchangeable. I am responsible for writing all the code to create tables and figures of this thesis as well as producing the first draft of derivations and manuscripts. Nevertheless, my collaborators have enhanced my work immensely through their ideas, guidance, and feedback.

In this thesis, Chapter 2 and 3 are adapted from my first-authored journal publication *Bayesian Lesion Estimation with a Structured Spike-and-Slab Prior* published in the *Journal of the American Statistical Association* (Menacher et al., 2023).

Abstract

The recent surge in large-scale population health datasets, such as the UK Biobank or the Adolescent Brain Cognitive Development (ABCD) study, requires the development of scalable statistical methods that are capable of analysing the rich multitude of data sources. This thesis focuses on the scalable analysis of Magnetic Resonance Imaging (MRI) neuroimaging data, such as binary lesion masks and task-based functional Magnetic Resonance Imaging (fMRI). In particular, we introduce two Bayesian spatial models with sparsity priors on the spatially varying coefficients and extend our work to suit the large sample sizes found in population health studies.

Firstly, we propose a scalable hierarchical Bayesian image-on-scalar regression model, called BLESS, capable of handling binary responses and of placing continuous spike-and-slab mixture priors on spatially varying parameters. Thereby, enforcing spatial dependency on the parameter dictating the amount of sparsity within the probability of inclusion. The use of mean-field variational inference with dynamic posterior exploration, which is an annealing-like strategy that improves optimisation, allows our method to scale to large sample sizes. We validate our results via simulation studies and an application to binary lesion masks from the UK Biobank.

Secondly, we extend our method to account for underestimation of posterior variance due to variational inference by providing an approximate posterior sampling approach inspired by Bayesian bootstrap ideas and spike-and-slab priors with random shrinkage targets. Besides accurate uncertainty quantification, this approach is capable of producing novel cluster size-based imaging statistics, such as credible intervals of cluster size, and measures of reliability of cluster occurrence.

Thirdly, we develop a Bayesian nonparametric scalar-on-image regression model with a relaxed-thresholded Gaussian process prior on the spatially varying coefficients in order to introduce sparsity and smoothness into the model. Our main contribution is the improved scalability, allowing for larger sample sizes and bigger image dimensions, which is made possible by replacing posterior sampling with a variational approximation. We validate our results via simulation studies and an application to cortical surface task-based fMRI data from the ABCD study.

Contents

Appendices

List of Figures

List of Abbreviations

1D, 2D, 3D	One-, two- or three-dimensional, referring in this thesis to spatial dimensions in an image.
ABCD	Adolescent Brain Cognitive Development.
AIC	Akaike information criterion.
ARI	Adjusted Rand Index.
BB	Bayesian bootstrap.
BB-SSL	Bayesian Bootstrap Spike-and-Slab LASSO.
BIC	Bayesian information criterion.
BLESS	Bayesian Lesion Estimation with a Structured Spike-and-slab Prior.
BNL	Bayesian nonparametric learning.
BOLD	Blood oxygenation level dependent.
BR	Bayesian regression.
BSGLMM	Bayesian spatial generalised linear mixed model.
CAVI	Coordinate ascent variational inference.
COPE	Contrast of parameter estimate.
CV	Cross validation.
ELBO	Evidence lower bound.
DPE	Dynamic posterior exploration.
FLAIR	Fluid attenuated inversion recovery.
fMRI	Functional magnetic resonance imaging.
FDR	False Discovery Rate.
FN	False Negative.
FP	False Positive.
FPR	False Positive Rate.
GLM	Generalised linear model.

GP	Gaussian process.
GPR	Gaussian process regression.
HCP	Human Connectome Project.
HPDI	Highest posterior density interval.
HRF	Hemodynamic response function.
iid	Independent and identically distributed.
KL	Kullback Leibler.
MAP	Maximum-a-posteriori.
MCAR	Multivariate conditional autoregressive.
MCMC	Markov chain Monte Carlo.
MFVI	Mean-field variational inference.
MNI	Montreal neurological institute.
MR	Magnetic resonance.
MRI	Magnetic resonance imaging.
MSE	Mean squared error.
NNGP	Nearest-neighbour Gaussian process.
PPC	Posterior predictive check.
RI	Rand Index.
SS-LASSO	Spike-and-slab LASSO.
TDR	True Discovery Rate.
TN	True Negative.
TP	True Positives.
TPR	True Positive Rate.
UKBB	UK Biobank.
VI	Variational inference.
WBB	Weighted Bayesian Bootstrap.
WLB	Weighted Likelihood Bootstrap.

1
Introduction

Contents

Medical imaging of the human body has greatly advanced the diagnosis of diseases, the management of disease progression, and the research of biomarkers for diseases. Most imaging procedures are entirely non-invasive, meaning that no injections are required; however, they are capable of providing clinicians or researchers with a view of internal body structures and/or organs. Common imaging types include X-ray, computed tomography (CT), ultrasound, positron emission tomography (PET) and magnetic resonance imaging (MRI) where each type has its own unique advantages and disadvantages with respect to its usage for providing medical care in a clinical setting or answering research questions in an academic setting. Moreover, while some methods share similarities with respect to their structure, each method requires the development of statistical methods and preprocessing techniques that consider the complexities of each imaging modality itself. In our work, we focus on developing statistical methods relevant towards advancing the analysis of MRI data. However, we note that while most of our statistical advancements have been inspired by MRI applications, specifically MRI of the brain, their usage is not entirely exclusive to MRI.

In the neuroimaging community, large population health-based studies, containing a breadth of imaging, omics, electronic health records, and further data sources depending on the study, have become available to researchers in the recent years. The UK Biobank (Miller et al., 2016), the ABCD study (Casey et al., 2018), or the Human Connectome Project (Van Essen et al., 2012) are some examples of studies containing hundreds or thousands of subjects with various MRI modalities, such as structural, functional and/or diffusion-weighted MRI. Population health-based studies pose a huge increase in the sample size of studies, as previously studies with MRI data were in the size of tens of subjects rather than thousands. Hence, one of the main aims of our work is creating statistical models and algorithms for parameter estimation, prediction, and inference that scale to thousands of subjects.

Besides the large sample sizes, imaging data usually consists of thousands to millions of variables making up each image of a subject, for example a 3D volumetric MRI scan with a resolution determined by the voxel size $2 \times 2 \times 2$ mm^3, where a voxel is defined as a volume in the 3D image, and the dimensions $91 \times 109 \times 91$ contains a total of 902,629 spatial locations. Moreover, modern MRI scanners are able to acquire images at an even better resolution; hence, decreasing the voxel size and increasing the number of spatial locations within an image. Therefore, suggesting that we not only require more scalable methods for the increase in sample size but also methods that can scale to the increase in spatial locations within an image.

However, signal in an image is also inherently sparse and only few spatial areas are usually significant in a statistical analysis. Another main aim of our work is therefore to build statistical models which allow for sparsity and smoothness. Hence, respecting the scarcity of signal and the spatial dependence between neighbouring locations within the brain. Bayesian methods are capable of integrating both sparsity and smoothness through prior specifications on our model parameters, specifically we advance the areas of Bayesian structured shrinkage priors and thresholded Gaussian process priors which allow us to utilise Bayesian variable selection as a proxy for inference.

The rest of this chapter reviews the core topics that motivate this thesis, which are MRI, large-scale population health data, the fundamental statistical theory we build on, image-based regression problems, Bayesian variable selection, approximate posterior inference, and finally we describe the contributions of this thesis.

1.1 Magnetic Resonance Imaging

MRI is a non-invasive imaging tool used in clinical and research settings to diagnose and monitor diseases and to study brain anatomy and activity. Magnetic resonance (MR) images are created by using the magnetic properties of atomic nuclei, where electromagnetic pulses induce infinitesimal magnetic moments in tissue in a way that can be measured by a MRI scanner (Jenkinson and Chappell, 2018). Most commonly in MRI, hydrogen protons found within water or fat molecules are targeted as they are abundant in the human body and possess the required spin property (Brown and Semelka, 2011). The MRI scanner consists of a large superconducting coil that creates a strong magnetic field which forces the hydrogen nuclei to align in the direction of the magnetic field. Without an external magnetic field the net magnetisation, defined as the sum of magnetisation from all nuclei, measured amongst the hydrogen nuclei within the tissue is zero and no magnetisation can be detected. In the MRI scanner the tissue has a net magnetisation but is incredibly weak relative to the field of the magnet.

Since the tissue magnetisation is so weak, it can only be detected in a plane orthogonal to the external magnetic field. Hence, in a process called “excitation” a second magnetic field is introduced with radio-frequency pulses which changes the orientation of the magnetisation of the hydrogen nuclei to be orthogonal with the initial magnetic field. The process of “relaxation” occurs when the radio-frequency pulses stop, and orthogonal magnetisation is measured with receiver coils as the hydrogen nuclei return to their original state, in alignment with the static external magnetic field.

As described so far, the net magnetisation measured would be made up from contributions of all hydrogen nuclei excited, e.g. the whole head. To enable image formation, gradient coils are employed to induce spatial gradients in the magnetic field (Jenkinson and Chappell, 2018). The application of these changing gradient fields introduces spatially dependent frequency and phase variations that can then be transformed into a MR image by using Fourier transforms. For a more detailed explanation on MRI physics, we refer the reader to Brown and Semelka (2011).

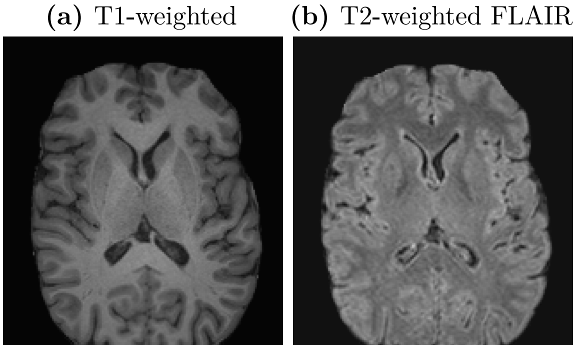

(a) T1-weighted **(b)** T2-weighted FLAIR

Figure 1.1: Examples of different MRI modalities: (a) T1-weighted and (b) T2-weighted FLAIR MRI scan in native space of a random subject from the UK Biobank.

MRI is capable of producing different types of images that are called “modalities”; each modality has a different tissue contrast or captures a different part of the brain anatomy or metabolism. Two fundamental properties in MR are the T1 and T2 relaxation rates, and their different scanning sequences can emphasise differences in tissue T1 or differences in tissue T2; these kind of scans are called T1-weighted and T2-weighted, respectively. Figure 1.1 shows an example of a 2D axial slice of a T1-weighted and a T2-weighted fluid attenuated inversion recovery (FLAIR) MRI scan. In a T1-weighted MRI scan, white matter appears as brightest, grey matter as darker and cerebral spinal fluid as darkest. On the other hand, in T2-weighted images cerebral spinal fluid shows as most intense, grey matter as darker and white matter as darkest. Different contrasts enhance different tissue types and signal and therefore, T1-weighted scans are predominately used to discriminate between tissue types and T2-weighted scans are used to identify brain structure (Jenkinson and Chappell, 2018).

1.1.1 Functional MRI

Functional MRI (fMRI) enables the study of brain activity at rest or during a specific task and requires the collection of a series of MRI scans over time. As neuronal activity cannot be measured directly, the Blood Oxygenation Level Dependent (BOLD) signal is recorded as a reasonable proxy as it measures changes in metabolic demand (Buxton et al., 2004). These metabolic changes lead to changes in localised blood volume, blood flow, and oxygen extraction and interact with the main magnetic field due to the magnetic properties of deoxygenated haemoglobin found in blood. Hence, fMRI consists of rapidly acquiring T2*-weighted images, where T2* relaxation time describes the observed relaxation in contrast to the true transverse relaxation in T2-weighted images, that are able to detect changes in blood oxygenation which correlates to the discovery of neuronal activity across different regions within the brain (Jenkinson and Chappell, 2018).

1.1.1.1 Preprocessing of fMRI

MRI data requires extensive preprocessing before starting any statistical analysis and includes correcting for artifacts, enhancing the signal within an image, and ensuring alignment within and between subjects' MRI scans. The following list provides a brief overview of commonly applied preprocessing steps for fMRI data where most of these steps also apply to structural MRI:

- **Distortion correction**: Dropout or geometric distortions occur due to inhomogeneities of the main magnetic field induced by the interfaces between air and tissue in the brain, such as sinus cavities or ear canals. Distortion correction alleviates this issue by applying a field map which characterises the main magnetic field.

- **Motion correction**: Head motion, caused by a subject moving in a MRI scanner, e.g. due to swallowing, causes a mismatch of all subsequent images recorded in the time series. The phenomenon is called “bulk motion” and can be corrected by realigning the images in the time series with a reference image.

- **Slice timing correction**: FMRI data is most commonly acquired with 2D MRI acquisition which means that 2D slices are collected consecutively to acquire a full 3D MRI scan. This can be problematic as the differences in acquisition times between the slices can cause bias in the downstream modelling of the expected BOLD signal which assumes that all slices have been collected at the same time. Slice timing correction reduces this effect by choosing a reference slice and aligning the time measurements of all other slices to the timing of the reference slice (Henson et al., 1999; Sladky et al., 2011). It should be noted that most recent neuroimaging pipelines for large-scale population health studies do not contain slice timing corrections as it can potentially exacerbate already existing artefacts throughout the entire time series and slice timing effects can be reduced with other measures, such as a combination of short repetition times and interleaved acquisitions, spatial smoothing, and the inclusion of temporal derivatives in statistical models.

- **Bias field correction**: Bias field correction is necessary for images acquired with MRI scanners of a field strength of greater or equal to 3T. While fMRI data is largely unaffected by the issue of broad variations in the intensity across the image due to inhomogeneities in the excitation of the head, the T1-weighted anatomical scan needed for downstream fMRI processing, such as registration or brain extraction, is. One example of bias field correction is the application of a high-pass filter which removes any low-frequency signals from the image (Friston et al., 2000).

- **Brain extraction**: The process of removing the skull and all other non-brain tissue, also called "skull-stripping", and is an important preprocessing step as many subsequent analyses depend on having a brain-only image.
- **Spatial smoothing**: The smoothing is accomplished with the convolution of a 3D image with a 3D Gaussian kernel in order to remove a certain level of high-frequency information. Smoothing reduces the variance of the noise and, as long as not too much high spatial frequency is lost, generally results in an increase in signal-to-noise ratio for voxel- or vertex-wise analyseswhere a voxel or a vertex are spatial locations in the volume or on the surface depending on the representation of a MRI scan. However, note that it has been argued by Mejia et al. (2020) that spatial smoothing can introduce signal contamination if performed in volumetric-based representations of MRI data as locations that are close in the volume are potentially far apart on the surface in the cortex, which is then a motivation to work with a cortical surface representation of the data (see Section 1.1.3). Nevertheless, spatial smoothing can still aid in decreasing the variability in spatial localisation across subjects that cannot be removed by registration and might be necessary to achieve a certain degree of smoothness required in statistical analyses, that for example involve Gaussian random field theory.
- **Co-registration**: The process of spatial normalisation that registers each scan collected in a fMRI time series to the subject's anatomical structural image.

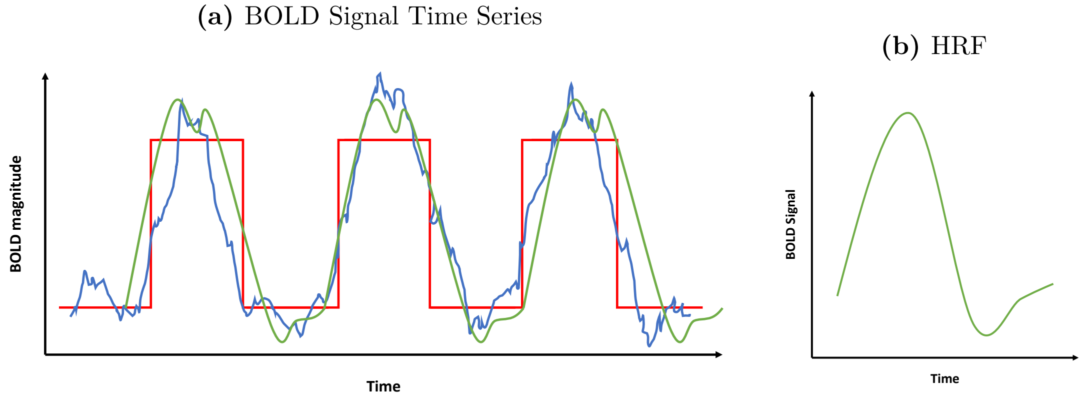
- **Registration**: Inter-subject registration or spatial normalisation of each subject's MRI scan to a common template, often to the Montreal Neurological Institute (MNI) template, in order to enable the statistical analysis between multiple subjects as it is essential to match up spatial locations between subjects before performing any group analyses.

For a more extensive overview on preprocessing pipelines, please refer to Glasser et al. (2013) for a description of the the minimal preprocessing pipelines of the Human Connectome Project (HCP), to Alfaro-Almagro et al., 2018 for the preprocessing pipeline of the UK Biobank (UK Biobank) and to Poldrack et al., 2011 for practical recommendations for preprocessing fMRI data.

1.1.1.2 Task-based fMRI Analysis

Functional MRI uses changes in BOLD signal across the brain and over time in order to evaluate brain function. There are two types of fMRI: resting-state and task-based fMRI. The former investigates the brain activity when a subject is at rest and the later studies brain activity when a subject is asked to perform a task in the scanner. Generally, in a fMRI experiment subjects are asked to respond to certain stimuli that a researcher has designed. These stimuli can be shown to the subject very briefly in event-based design or for a prolonged period in a block design. Moreover, stimuli are usually repeated within an imaging session in equally spaced time intervals in order to increase statistical power by acquiring more data from a subject. Figure 1.2 (a) provides an example of a block design where the stimulus onset timings recorded in red show that the stimuli are presented to the subject for a prolonged time period. In order to acquire a full fMRI time series, the BOLD signal for a subject is recorded across multiple time points as a separate image at each time point.

The resulting BOLD signal time series, shown in blue in Figure 1.2 (a), is then compared to the experimental design, shown in red in Figure 1.2 (a), to identify which areas in the brain are activated. For each region of the brain which is responsive to the task, the predicted BOLD activation is then acquired by convolving



(a) BOLD Signal Time Series

(b) HRF

Figure 1.2: (a) Illustration of BOLD fMRI time series in active voxel with the BOLD signal (blue), the stimulus time series (red) and the convolved approximation of the HRF with the stimulus onset timings which is the predicted BOLD activation (green). (b) Illustration of the HRF.

the hemodynamic response function (HRF) with the time series of the experimental design, shown in green in Figure 1.2 (a). The BOLD response measured with the HRF function across time then acts as a reasonable proxy to neuronal activity across the brain (Poldrack et al., 2011; Jenkinson and Chappell, 2018).

The statistical analysis of fMRI data is most commonly performed in two steps: a subject-level and a group-level analysis. Moreover, the analysis is performed one voxel or vertex at a time which is often referred to as a "mass-univariate" approach. In the first-level analysis, a separate linear regression model is constructed for each subject i and each spatial location j with the following equation:

$$\boldsymbol{Y}_{i,j} = \boldsymbol{X}_{i,j}\boldsymbol{\beta}_{i,j} + \boldsymbol{\epsilon}_{i,j} \qquad\qquad \boldsymbol{\epsilon}_{i,j} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2_{i,j}\boldsymbol{I}_{t_i}),$$

where t_i is defined as the number of time points collected in a fMRI study, $\boldsymbol{Y}_{i,j\in\mathbb{R}^{t_i}}$ contains the observed BOLD signal across time for each subject i and each spatial location j, $\boldsymbol{X}_{i,j\in\mathbb{R}^{t_i\times P}}$ is the design matrix containing the intercept, the predicted BOLD activation (see Figure 1.2) and other confounds for each subject i and each spatial location j, where P defines the number of covariates in the analysis, and $\boldsymbol{\epsilon}_i$ is defined as a random error term which is normally distributed with variance $\sigma^2_{i,j}$. The output of each regression in the first-level analysis is the subject-specific parameter estimates $\boldsymbol{\beta}_{i,j\in\mathbb{R}^P}$ at each voxel or vertex j and the within-subject variance for each contrast of interest, see Figure 1.3 (a) for an illustration of the subject-level analysis within the two-step procedure of fMRI analysis.

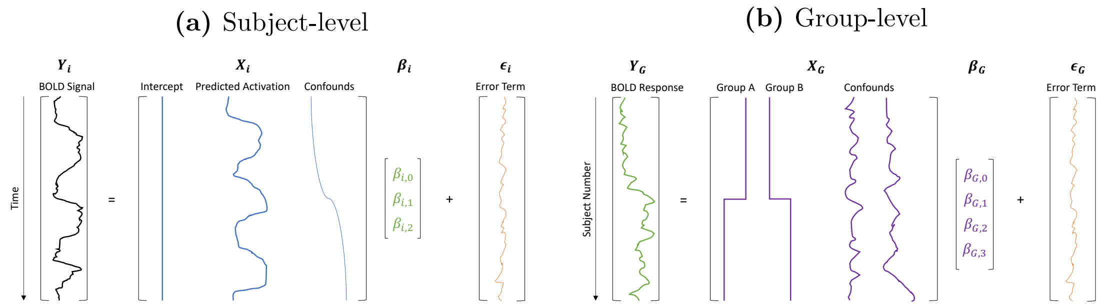

(a) Subject-level **(b)** Group-level

Figure 1.3: Illustration of the two-stage voxel- or vertex-wise modelling approach used for fMRI data. (a) In the first- or subject-level analysis, a regression associating the BOLD signal across all recorded time points with the predicted BOLD activation and other confounds is performed for each subject at each voxel or vertex separately. (b) In the second- or group-level analysis, a regression associating the coefficient of the predicted BOLD response, which was acquired during the first-level analysis for each subject and voxel or vertex, with group differences and other confounds is performed to make inference about group averages or between-group differences with $\hat{\boldsymbol{\beta}}_G$.

The second-level analysis takes the subject-specific parameter estimates for the BOLD signal, in Figure 1.3 (a) the estimates for $\beta_{i,j,1}$ for each subject $i = 1, \dots, N$ where N is the number of subjects in the study and spatial location $j = 1, \dots, M$ where M is the number of voxels or vertices, and inputs them into a group-level regression to identify group averages or between-group differences. The linear model for the group analysis can be expressed by

$$\boldsymbol{Y}_{j,G} = \boldsymbol{X}_{j,G}\boldsymbol{\beta}_{j,G} + \boldsymbol{\epsilon}_{j,G} \qquad\qquad \boldsymbol{\epsilon}_{j,G} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2_{j,G}\boldsymbol{I}_N),$$

where $\boldsymbol{Y}_{j,G}$ is defined by the subject-specific parameter estimates for the BOLD signal contrast $\hat{\boldsymbol{\beta}}_{j,1}$, $\boldsymbol{X}_{j,G}$ is the group-level design matrix containing group membership and other confounds, such as age or sex, and $\boldsymbol{\epsilon}_{j,G}$ is the random error term of the group analysis. The parameter vector of the group analysis $\boldsymbol{\beta}_{j,G}$ contains the coefficients of interest on which inference on group differences is performed on, see Figure 1.3 (b) for an illustration of the setup of a second stage group-level analysis. In summary, in a mass-univariate group-level fMRI analysis the collection of M regressions performed at each voxel or vertix location results in a parameter map which can be overlayed on a MNI template to display the results of the group-level fMRI analysis.

1.1.2 Brain Lesion Masks

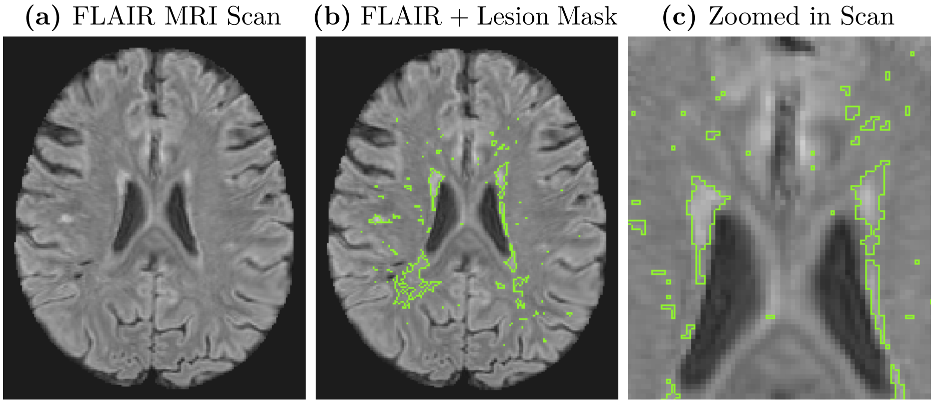


Figure 1.4: (a) T2-weighted FLAIR MRI scan in native space, (b) FLAIR MRI scan with overlaid contours of the respective lesion mask and (c) zoomed in version of FLAIR MRI scan with overlaid contours of a lesion mask from (b) to highlight the outlines of lesions.

It is also possible to derive other measures of interest from MRI scans, such as lesion masks, which quantify whether or not a lesion is present at each location within an image, see Figure 1.4 (b) for an example of contours of a lesion mask overlayed on a T2-weighted FLAIR MRI scan. Lesions can occur due to various reasons, such as ageing or disease development and progression, e.g. with multiple sclerosis. For example, lesions of presumed vascular origin occur more frequently in the brain with increased age and are associated with cognitive decline as the lesion burden increases (Wardlaw et al., 2013; Debette and Markus, 2010). In T2-weighted, FLAIR and proton density-weighted brain images those white matter lesions appear as hyperintense areas in the brain.

The MRI scans however also require preprocessing in order to be utilised for the quantitative analysis of lesions. Firstly, researchers create binary lesion masks for every subject which mark the presence or absence of a lesion at a location within the brain. Lesions segmentation can either be performed manually by radiologists or automatically via procedures, such as BIANCA (Griffanti et al., 2016). In this thesis we focus on large-scale lesion mapping studies, such as the UK Biobank (Miller et al., 2016), hence fully automated algorithms are preferable and allow for a fast and reproducible method of creating lesion maps for thousands of subjects. After performing manual or automatic lesion segmentation the masks are registered from native space to a common anatomical atlas for group analysis. Spatial alignment to a standard space therefore ensures that lesion localisations are comparable across subjects and analyses (Bordin et al., 2021). Note that binary image data are registered to an atlas where interpolation will produce grey values between 0 and 1. The traditional approach is to threshold the interpolated data at some value (e.g. 0.5) to keep the data in binary form (Kindalova et al., 2021).

1.1.3 Volumetric and Cortical Surface Representations

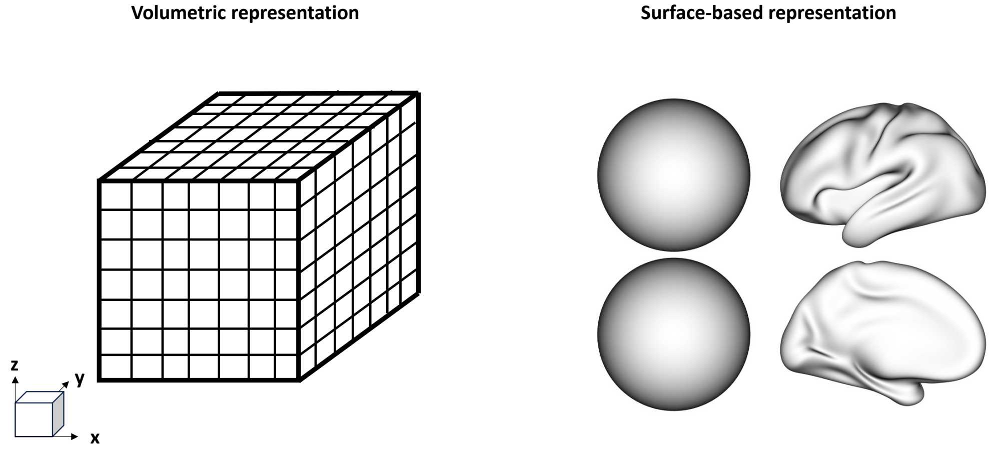


Figure 1.5: Illustration of volumetric and surface-based representation for MRI data.

In general, MRI data can be analysed and displayed in the form of volumetric as well as cortical surface images, see Figure 1.5 for an illustration of the two representations. Volumetric data consists of 3D brain volumes where each is composed of thousands or millions of equally sized volumetric elements that are called voxels. While the analysis of volumetric data is still the most common practice in the neuroimaging community, cortical surface data is becoming increasingly more popular with more preprocessing, analysis and display tools, such as the `Human Connectome Workbench` (Marcus et al., 2011) or `ciftiTools` for data visualisation in `R`, being available. Cortical surface data compared to volumetric data represents cortical grey matter as a 2D manifold surface (Fischl, 2012; Glasser et al., 2013).

Mejia et al. (2020) highlight the benefits of analysing MRI data with a cortical surface-based representation as it provides better whole brain visualisation, dimension reduction as only the surface is of interest in addition to being able to downsample the data without significant signal loss, removal of not needed tissue types and improved neurobiological significance of distances, see Figure 1.6. In order to improve the signal detected within fMRI studies, a common preprocessing step is to apply spatial smoothing of the data prior to the analysis. However, applying volumetric-based spatial smoothing introduces bias in the form of signal contamination and the potential for identifying more false positives due to the folding of the brain as areas that are nearby in the folded cortex can be far apart in the unfolded cortex. Hence, the benefit of smoothing on the cortical surface is that the signal-to-noise ratio is enhanced by strengthening the signal in the same brain region of interest (Brodoehl et al., 2020). To further highlight this issue, we display a zoomed in T1-weighted scan in Figure 1.6 where various locations in the image are marked. The figure illustrates that traditionally used Euclidean distances in volumetric representations can enlarge bias in Bayesian spatial models or when spatial smoothing is applied as a preprocessing step for fMRI data. For example, it would result in the mixing of signals from locations 1A, 1B, and 1C, as well as those from locations 2A and 2B, and in a Bayesian spatial model the signal coming from locations 1A, 1B, and 1C (and locations 2A and 2B) would be assumed to be highly correlated. However, when using a cortical surface-based representation the locations that do not contain any cortical grey matter would be removed from the image, such as location 1C and 2A, as they cannot exhibit any relevant task activation. Furthermore, the signal coming from locations 1A and 1B is not highly correlated but exhibits low dependence on the surface when measured with a geodesic distance.

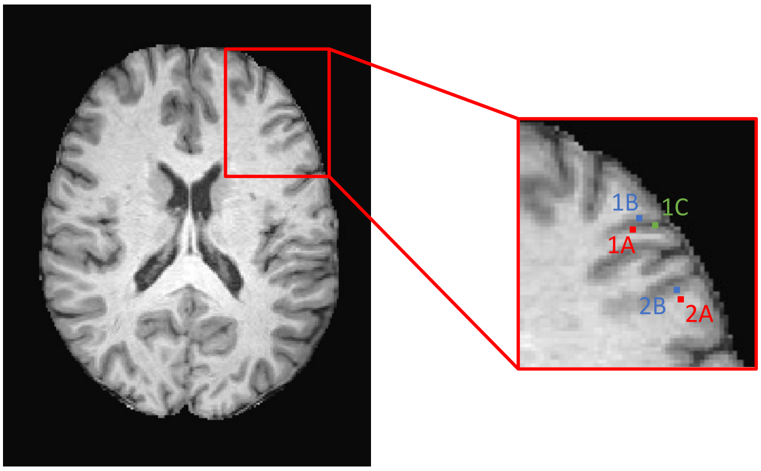


Figure 1.6: Illustration for distances in volumetric space. For one random subject from the UK Biobank, an axial slice of the T1-weighted image is displayed. Locations 1A, 1B, and 1C are close in terms of Euclidean distance in volumetric space, but are neurologically dissimilar, as location 1A lies in grey matter, 1B also falls in grey matter, however on the opposite sulcus bank, and 1C is located in cerebral spinal fluid between sulci banks. Therefore, locations 1A and 1B may exhibit task activation, while location 1B would not be expected to exhibit any task-related activation. Similarly, locations 2A and 2B are neighbouring in volumetric space, but location 2B lies in grey matter while location 2A lies in white matter and therefore would not be expected to exhibit task-related activation. (Figure replicated from Mejia et al. (2020).)

Lastly, cortical surface data has been shown to have improved cross-subject spatial alignment when registered to a common template by first mapping each hemisphere of the cortex onto the surface of a sphere with minimal distortion (Fischl et al., 1999a; Fischl et al., 1999b), see Figure 1.7 for an illustration of going from an inflated to a spherical representation of MRI data. The process of acquiring cortical-surface fMRI data from volumetric data is described in the following steps (Mejia et al., 2020):

1) Identification of cortical grey matter ribbon from a high-resolution structural MRI scan (Dale et al., 1999).

2) Application of a mesh to the white matter surface, the internal boundary of the cortical grey matter, to form a 2D manifold within each hemisphere, which is geometrically smoothed.

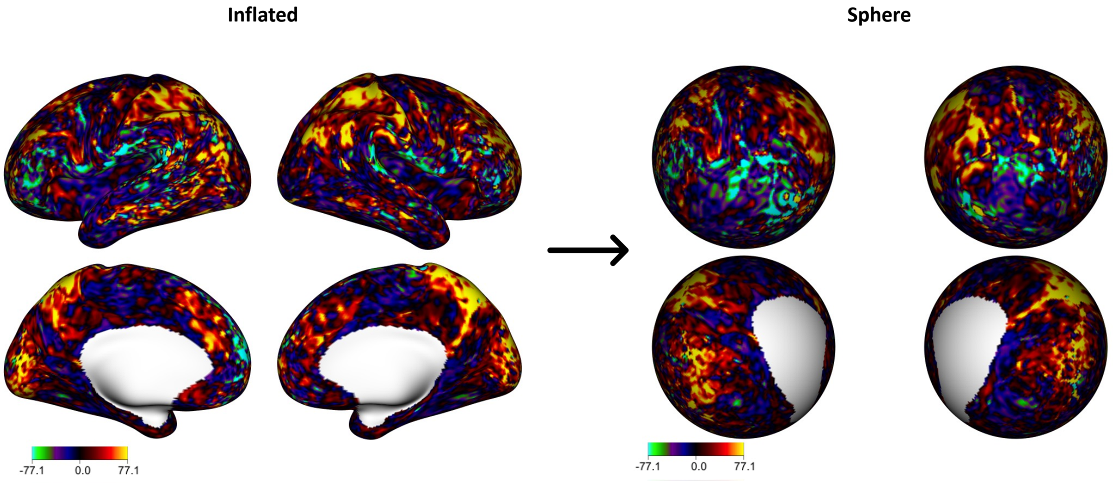


Figure 1.7: Example mapping of cortical surface task-based fMRI data, specifically contrast of parameter estimate (COPE) coefficients from the first level analysis of a task-based fMRI experiment of a random subject from the ABCD study, onto a sphere. The left side shows a four-way view of the COPE coefficients on an inflated surface showing the lateral or exterior (top) and medial or interior (bottom) views of both hemispheres. The right side shows the same COPE coefficients on a spherical surface which showcase the benefit of cortical surface data which can easily be mapped onto different surface representations and hence mathematically convenient measure of geodesic distance along the cortical surface can be leveraged.

3) Inflation of the surface to a sphere with minimal distortions (Fischl et al., 1999a; Fischl et al., 1999b).

4) Registration of subjects to a standard template space by aligning anatomical folding patterns (Fischl et al., 1999a; Fischl et al., 1999b).

5) Application of volume-to-surface transformation to each cortical grey matter ribbon for each fMRI volume to obtain cortical surface fMRI time series.

The output of the above process is a triangular mesh consisting of approximately 30,000 vertices within each hemisphere.

1.2 Large-scale Population Health Datasets

The last decade of brain imaging has brought immense insight into our understanding of the human brain. However, many findings suffer from small and unrepresentative samples. In addition, environmental and genetic factors that may explain individual differences are often ignored. These limitations are being addressed in large-scale epidemiological studies, such as the UK Biobank (UKBB) (Miller et al., 2016), the Human Connectome Project (HCP) (Van Essen et al., 2013) or the ABCD study (Casey et al., 2018), by collecting data on thousands (instead of tens) of subjects. While the main advantage of these data sources lies in their larger sample sizes, they are also more comprehensive due to their inclusion of multiple high-dimensional imaging modalities as well as numerous environmental factors, neuro-cognitive scores, and other clinical data. Many existing methods for brain mapping are either simplistic, ignoring complex spatial dependency, or are not scalable to large-scale studies.

1.2.1 UK Biobank

The UKBB is a prospective epidemiological study with a predominantly healthy cohort of approximately 500,000 participants, mostly with white British ancestry, aged between 40 to 69 years at recruitment. The study includes questionnaires data, physical and cognitive measures, biological samples, including genotyping (Sudlow et al., 2015), and additionally has been extended to provide imaging data for a subset of approximately 100,000 subjects from the original cohort (Miller et al., 2016). The brain imaging consists of six modalities: T1-weighted, T2-weighted FLAIR, susceptibility weighted, resting-state fMRI, task-based fMRI and diffusion-weighted MRI (Alfaro-Almagro et al., 2018). While the primary goal of brain imaging in a clinical setting is the diagnosis and monitoring of disease progression, the UKBB provides the option for researchers to identify predictive biomarkers which enable the possibility of early disease intervention or prevention. In the real data application of Chapter 2 and 3, we utilise the brain lesion masks derived from the T2-weighted FLAIR MRI scans to find associations between ageing and lesion incidence. Within our analysis we use the data, specifically the brain lesion masks, that has been preprocessed with the pipelines developed by Alfaro-Almagro et al. (2018).

1.2.2 ABCD Study

The Adolescent Brain Cognitive Development (ABCD) study is a population-based health study following the brain development and health of approximately 11,875 participants who are aged between 9 to 10 years old. The children are routinely examined across a time span of ten years into their adulthood across 21 acquisition sites. The primary goal of the study is to assess development and addiction behaviours of adolescents by measuring their brain structure and function (Casey et al., 2018; Karcher and Barch, 2021). However, the ABCD data contains various additional data sources and has therefore been used for other applications beyond its original study goal, such as identifying the effect of social media on childhood development via wearables data (Bagot et al., 2018), the overall impact of screen use in adolescents (Bagot et al., 2022), or the influences of puberty hormones or environmental factors towards the physical and mental development in children (Karcher and Barch, 2021). The study contains a breath of data sources including multiple imaging modalities, such as structural, resting state and task-based functional, and diffusion weighted MRI, as well as a collection of biospecimen data, behavioural assessments, wearables data and other environmental factors assessed via questionnaires which were answered by the children as well as their parents (Casey et al., 2018; Uban et al., 2018; Luciana et al., 2018). The real data application in Chapter 4 utilises the task-based fMRI data of the ABCD study, specifically the 2-back vs 0-back contrast of the emotional n-back task, alongside the total composite score for cognition and other confounds, such as age, sex, or family income. The data has been preprocessed with the minimally processed pipeline developed by Glasser et al. (2013).

1.3 Image-based Regression Problems

Neuroimaging data is most commonly analysed with mass-univariate approaches that fit a regression at each voxel or surface element independently. While this approach is fast, it ignores any spatial dependence that exists across the brain.

On the other hand, Bayesian spatial models address this limitation by building a single multivariate model that captures the spatial dependence across the brain. However, Bayesian spatial models are usually estimated via Markov chain Monte Carlo (MCMC) sampling which is computationally costly or even infeasible for larger sample sizes that can be found in large-scale population-based health studies, such as the ABCD study or the UKBB. In the following section, we will review various types of image-based regression problems - scalar-on-image, image-on-scalar, and image-on-image regression models - that tackle complex problems within the area of neuroimaging applications. Note that we generalise regression problems to generalised linear models in the following examples and hence, the output scalar or image can be binary in nature, for example.

1.3.1 Scalar-on-Image

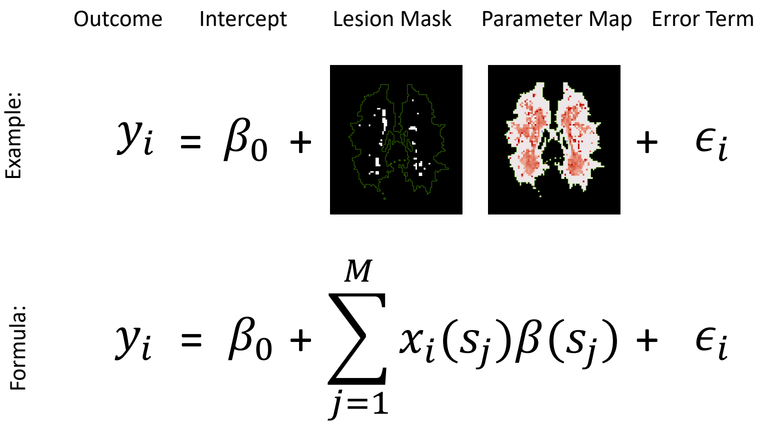


Figure 1.8: Scalar-on-image regression illustration where the output is a scalar and the input covariates are spatial locations of an image. The bottom row represents the general formula for scalar-on-image regression problems and the top row visualises an example of a scalar-on-image regression with a scalar output y_i, e.g. a disease severity score, intercept β_0, lesion mask (green outlines the white matter mask) / covariate map $\boldsymbol{X}_i$, parameter map $\boldsymbol{\beta}(\boldsymbol{s})$, and error term ϵ_i. Note that in the formula the output image $\boldsymbol{y}_i$ and the parameter map $\boldsymbol{\beta}(\boldsymbol{s})$ are unravelled vectors of size M but in the example we display them as 2D matrices for illustration purposes.

Scalar-on-image regressions model the associations between a scalar outcome of interest y_i for each subject $i = 1, \ldots, N$ and an image, where each image location $s_j = 1, \ldots, M$ for all M spatial locations is treated as a covariate $x_i(s_j)$, see

Figure 1.8 for an illustration. The model also includes an intercept β_0 and a random error term ϵ_i for each subject i. Statistical methods for scalar-on-image regressions can be split in projection-based approaches, sparsity-inducing models or a combination of both. Generally, strong model assumptions are needed to overcome the inherent non-identifiability problem due to the number of parameters in the model far exceeding the number of subjects (Happ et al., 2018). While various model approaches for scalar-on-image regression problems may provide a similar predictive performance, the estimated model coefficients may vary starkly between the approaches which renders the interpretability of the estimates difficult. Happ et al. (2018) provide an overview of possible model assumptions, such as using basis function approaches, dimension reductions methods, or smoothing and/or sparsity priors, to insure identifiability. In the following examples we highlight various state-of-the art improvements for scalar-on-image regression problems and how they achieve identifiability by enforcing certain structural assumptions.

Firstly, projection-based methods re-express a coefficient image by expanding it with a set of basis functions. Reiss and Ogden (2010) propose a functional principal components regression which models the image coefficients by approximating the coefficient function via B-splines. However, neuroimaging data often exhibits regions with sparse effects and/or sharp edges which the above approach is unable to estimate well. Another basis expansion approach was proposed by Reiss et al. (2015) which utilises a wavelet expansion for the regression coefficients. Unfortunately, the above method requires each predictor image to be of equal dimension and the image size needs to be a power of two which is prohibitive with many imaging modalities exhibiting various image sizes.

Alternatively, many approaches acquire model identifiability by proposing a Bayesian model and enforcing regularisation via smoothing and sparsity priors. Huang et al. (2013) aim to predict cognitive outcomes via diffusion tensor imaging by introducing an Ising prior into the model which performs variable selection and therefore aids in identifying locations in the image which are predictive of the outcome of interest. Goldsmith et al. (2014) extend the above model by modelling the regression coefficients as a product of two latent spatial processes through an Ising prior, which induces sparsity into the model, and a conditional autoregressive prior, which smooths the nonzero coefficients. Similarly, Li et al. (2015) place a Dirichlet process prior on the nonzero coefficients. However, these prior specifications are computationally costly and exhibit additional difficulties as the parameter estimation suffers from phase transitions where a small change in hyperparameters leads to a significant change in the fraction of coefficients estimated as zero or nonzero. Other methods focus on Bayesian nonparametric models, Kang et al. (2018) introduce a class of piecewise smooth, sparse and continuous spatially varying regression coefficient functions called soft-thresholded Gaussian process priors. This class of priors provides a gradual transition between zero and nonzero coefficients of nearby voxels.

1.3.2 Image-on-Scalar

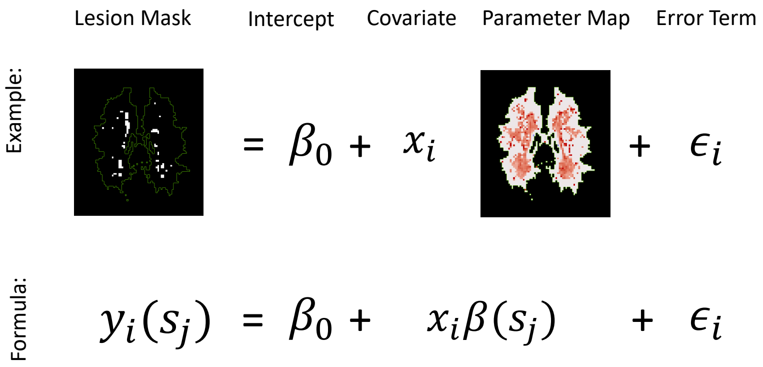


Figure 1.9: Image-on-scalar regression illustration where the output is an image and the input covariates are scalars. The bottom row represents the general formula for image-on-scalar regression problems and the top row visualises an example of a image-on-scalar regression with a lesion mask (green outlines the white matter mask) / output image $\boldsymbol{y}(\boldsymbol{s})$, intercept β_0, scalar covariate x_i, e.g. age, parameter map $\boldsymbol{\beta}(\boldsymbol{s})$, and error term ϵ_i. Note that in the formula the input image $\boldsymbol{X}_i$ and the parameter map $\boldsymbol{\beta}(\boldsymbol{s})$ are unravelled vectors of size M but in the example we display them as 2D matrices for illustration purposes.

In image-on-scalar regression problems the image $y_i(s_j)$ for each subject i is the response at every spatial location s_j for all $j = 1, \dots, M$ and the covariates x_i are scalar in nature, see Figure 1.9. Hence, providing insight into spatial patterns and heterogeneity across individuals by capturing the associations between clinical covariates, such as age, sex, disease duration, disease severity etc., and a type of imaging modality, such as structural, functional, or diffusion-weighted MRI. The complexity of image-on-scalar regression lies in the high-dimensional nature of the imaging data as the outcome, the spatial correlations and sparse signal within the data, and the low sample sizes compared to the number of parameters required for modelling the imaging outcome.

Initial approaches to image-on-scalar regression problems build a spatially varying coefficients model and place a smoothing prior, specifically a multivariate conditional autoregressive prior on the coefficients (Ge et al., 2014). However, this approach only utilises smoothing and induces no sparsity in the model. Zeng et al. (2022) build a Bayesian hierarchical model with a global-local spike-and-slab prior which performs image smoothing and variable selection simultaneously. The benefit of the global local feature is that sparsity is induced at two levels where the global prior achieves sparsity at a covariate level and the local prior achieves sparsity at a voxel level.

Other methods approximate the spatially varying coefficient functions in image-on-scalar regression problems with spline functions. For example, Li et al. (2020) use bivariate spline functions over triangulation. Another approach is formulated by Yu et al. (2021) which propose flexible multivariate splines over triangulations to handle the irregular domain of objects of interest on the images. Hence, being able to model spatial heterogeneity and being able to account better for spatial correlations.

Lastly, we provide an overview on some Bayesian nonparametric model approaches that utilise a different form of regularisation to handle spatially varying coefficient models. Zhang et al. (2023) estimate spatially varying coefficient functions via deep neural networks which aids in accounting for spatial smoothness, subject heterogeneity, and interpretability. A downside of Bayesian nonparametric models is usually that the computational cost is high due to MCMC sampling. Whiteman et al. (2023) propose a Bayesian spatial model for the analysis of group-level neuroimaging data with Gaussian process priors on the spatially varying coefficients and a non-stationary model for the error process. Computational tractability is ensured via the Vecchia approximation of Gaussian processes which is an ordered conditional approximation that leads to a sparse Cholesky factor of the precision matrix (Vecchia, 1988). Moreover, the above spatial model specification allows for better calibrated inference results compared to simply smoothing the data prior to analysis as a preprocessing step which is common practice in neuroimaging applications otherwise.

1.3.3 Image-on-Image

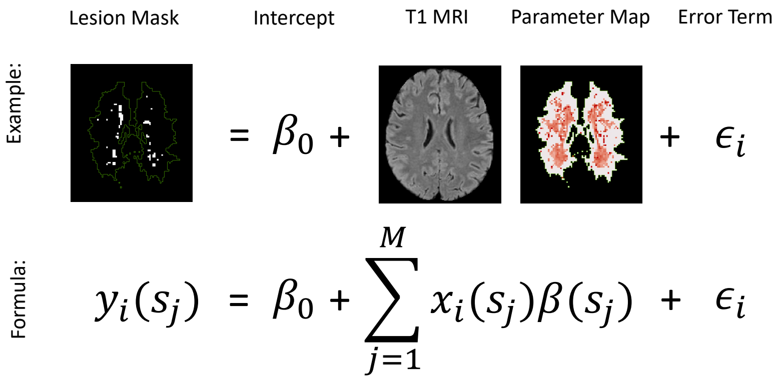


Figure 1.10: Image-on-image regression illustration where both the output and the input are an image with M spatial locations. The bottom row represents the general formula for image-on-image regression problems and the top row visualises an example of a image-on-image regression with a lesion mask (green outlines the white matter mask) / output image $\boldsymbol{y}(\boldsymbol{s})$, intercept β_0, T1-weighted MRI scan / covariate map $\boldsymbol{X}_i$, parameter map $\boldsymbol{\beta}(\boldsymbol{s})$, and error term ϵ_i. Note that in the formula the output image $\boldsymbol{y}_i$, the input image $\boldsymbol{X}_i$, and the parameter map $\boldsymbol{\beta}(\boldsymbol{s})$ are unravelled vectors of size M but in the example we display them as 2D matrices for illustration purposes.

Image-on-image regression approaches face the same issues as image-on-scalar regression, see Section 1.3.2, regarding the high dimensionality of the data and the complex spatial dependence structures between imaging outcome $y_i(s_j)$ and imaging predictors $x_i(s_j)$. Mass-univariate approaches are the most commonly used method to estimate coefficients in image-on-image regression problems. For example, Sweeney et al. (2013) propose voxelwise logistic regression models to associate a combination of imaging modalities, such as T1-weighted, T2-weighted, FLAIR, and proton attenuation MRI images, with lesion incidence. Similarly, Hazra et al.

(2019) uses a mass-univariate approach that includes prediction from neighbouring voxel locations within a certain Euclidean distance. Another key example for the need of image-on-image regression is proposed in the analysis by Tavor et al. (2016) who aim to associate task-based functional MRI with measurements from resting state functional MRI in order to show that individual differences in brain response are due to the structure of the brain and individual behaviours. Not only do these approaches neglect any spatial correlations in the data by assuming independence and an identical linear relationship across voxels but the latter also fits a model for each individual separately, averages the results across individuals, and then uses the average for out-of-sample predictions which contradicts the initial assumption of their work that individual variation is present in the imaging outcome, the tasked-based fMRI data.

On the other hand, Guo et al. (2022) develop a spatial Bayesian latent factor regression model to reduce the image dimensions with latent factors. Moreover, Gaussian process priors are placed on the spatially varying coefficient functions in order to capture the spatial dependence structures inherent in the data. Concretely, the outcome image is modelled with a linear combination of basis functions. In a next step, the basis functions are further decomposed into a few latent factors and a loading matrix. Lastly, a scalar-on-image regression is fitted by using the latent factors as an output and a summary of the imaging predictors as an input to the regression. The downside of this approach is that multiple imaging modalities are summarised as one entity in the scalar-on-image regression problem in order to lower the computational cost of the method by combining the cumulative effect across multiple imaging modalities. However, this is a disadvantage if there are imaging modalities with different image dimensions and one is interested in the individual effect of the separate imaging modalities (Guo et al., 2022). Another example of a Bayesian spatial model for image-on-image regression is given by Roy et al. (2021) who develop a product of independent Gaussian process priors with a smooth covariance kernel that model continuous signals and are sparse and piece-wise smooth. The latter work also provides an overview of all three types of regression - scalar-on-image, image-on-scalar, and image-on-image - with simulation studies for each regression type.

1.4 Bayesian Variable Selection

Variable selection has become particularly relevant for applications where the number of parameters far exceeds the number of subjects. Bayesian neuroimaging models usually suffer from exactly this problem and hence lend itself to utilising Bayesian variable selection. Generally, the process of variable selection allows for finding a subset of features that are most predictive of the outcome and hence decrease the identification of spurious associations. Other advantages of variable selection include the improved and simplified interpretation of coefficients in high-dimensional settings, the potential avoidance of overfitting and the better tackling of the issue of multicollinearity, and further insights into how the data was generated (Tadesse and Vannucci, 2022). There are three categories of approaches that employ variable selection: criteria-based methods, such as performing a subset selection in a regression based on, for example the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) (Miller, 2002), penalised likelihood methods, which shrink negligible coefficients to zero (Ormerod et al., 2017), and Bayesian approaches with shrinkage priors (George and McCulloch, 1993; Polson and Scott, 2010). In the following subsections we will mainly focus on Bayesian variable selection approaches, which can be divided into methods that use a mixture of two distributions, so called spike-and-slab priors, and unimodal continuous shrinkage priors, so called global-local shrinkage priors.

1.4.1 Spike-and-Slab Priors

One of the most commonly applied techniques for Bayesian variable selection is spike-and-slab regression which aims to identify a selection of predictors within a regression model. The original formulation of the spike-and-slab mixture prior places a mixture of a point mass at zero and a diffuse distribution on the coefficients (Mitchell and Beauchamp, 1988), see Figure 1.11 (a). This approach requires the

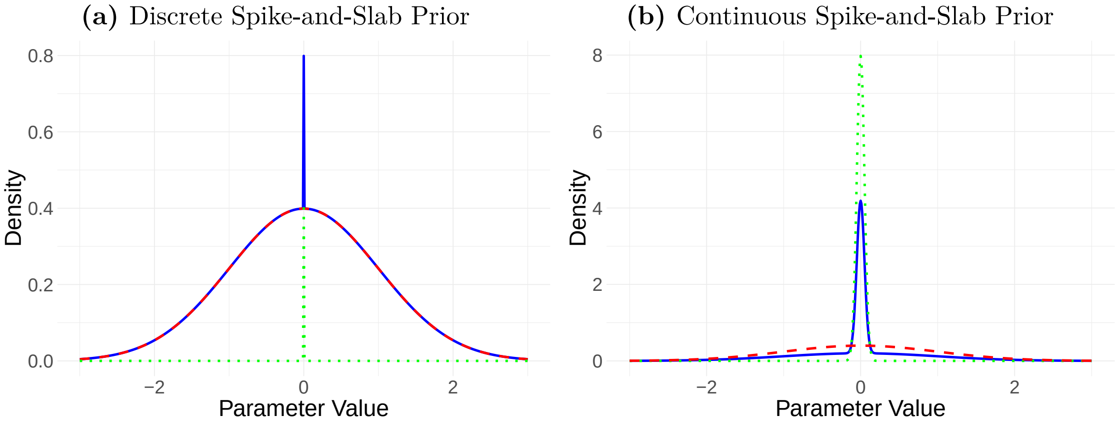


Figure 1.11: Illustration of (a) discrete spike-and-slab prior (solid blue) which is a mixture of a point mass prior at zero (spike; dotted green) and a diffuse Normal distribution (slab; dashed red) and (b) continuous spike-and-slab prior (solid blue) with a mixture of two Gaussian distributions with a very small variance (spike; dotted green) and a large variance (slab; dashed red). (Figure replicated from Tadesse and Vannucci (2022).)

calculation of posterior probabilities for all 2^P submodels, where P is the number of covariates. George and McCulloch (1993) and George and McCulloch (1997) have hence increased the computational feasibility of spike-and-slab regression problems by introducing the continuous mixture of Gaussians formulation where the spike distribution is not defined by a point mass prior but by a normal distribution with a small variance, see Figure 1.11 (b). While these are the most commonly applied versions of a spike-and-slab prior, there are many other versions of the spike-and-slab prior available. For example, other versions include non-local prior densities (Johnson and Rossell, 2010; Shin et al., 2018) or the spike-and-slab LASSO priors (Ročková and George, 2018; Deshpande et al., 2019).

The general setup of a simple normal spike-and-slab regression, modelling the association between a set of covariates $\boldsymbol{X} \in \mathbb{R}^{N \times P}$ and a dependent variable $\boldsymbol{y} \in \mathbb{R}^N$ via unknown parameter estimates $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)^T \in \mathbb{R}^P$, is hereby typically described by:

$$
\begin{aligned}
\boldsymbol{y} \mid \boldsymbol{\beta},\ \sigma^2 &\sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta},\ \sigma^2\boldsymbol{I}) \\
\text{(discrete)}\quad \beta_j \mid \gamma_j &\sim (1-\gamma_j)\,\delta_0(\beta_j) \ +\ \gamma_j\,\mathcal{N}(0,\ \nu_1) \\
\text{(continuous)}\quad \beta_j \mid \gamma_j &\sim (1-\gamma_j)\,\mathcal{N}(0,\ \nu_0) \ +\ \gamma_j\,\mathcal{N}(0,\ \nu_1) \\
\boldsymbol{\gamma} \mid \theta &\sim \text{Bernoulli}(\theta),
\end{aligned}
$$

where σ^2 is the residual variance, ν_0 is the spike variance, ν_1 is the slab variance, and $\delta_0(\cdot)$ is the Dirac function with a point mass at zero. The introduction of the binary latent variables $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_P)^T$ hereby enables the identification of high-probability subsets of predictors. Concretely, variable selection is performed by including the variable $\boldsymbol{x}_j$ in the model and estimating β_j with a non-zero value if $\gamma_j = 1$ and excluding variable $\boldsymbol{x}_j$ if $\gamma_j = 0$, in the case of the discrete formulation the coefficient value for β_j is exactly zero and in the case of the continuous formulation the coefficient value for β_j is shrunk to a value close to zero (Mitchell and Beauchamp, 1988; George and McCulloch, 1993). Continuous spike-and-slab regressions are therefore able to selectively shrink negligible coefficients to zero while leaving the larger parameters unaffected. The amount of regularisation is influenced by the spike and slab variance of the Gaussian mixture prior ν_0 and ν_1, respectively.

1.4.2 Global-local Shrinkage Priors

An alternative to mixture priors is given by unimodal priors which are also capable of inducing sparsity into models and therefore performing variable selection. Often the downside of models with spike-and-slab priors on the coefficients is their large computational cost due to the evaluation of 2^P models. However, it should be noted that this argument is no longer as impactful today as alternatives for model estimation, such as stochastic search variable selection (George and McCulloch, 1997) or variational inference (Carbonetto and Stephens, 2012; Titsias and Lázaro-Gredilla, 2011), have been proposed. Nevertheless, the computational cost of unimodal shrinkage priors is usually lower (Tadesse and Vannucci, 2022). Another advantage of global-local shrinkage priors, similarly to continuous mixture shrinkage priors, is that they only shrink coefficients to a value close to zero which is favourable in certain applications where many of the parameters are expected to be negligible but not exactly zero. Generally, the class of global-local scale mixture of normal priors is defined by

$$\begin{aligned} \beta_j \mid \tau^2, \ \lambda_j^2 &\sim \mathcal{N}(0, \ \tau^2\lambda_j^2) \\ \lambda_j^2 &\sim \pi(\lambda_j^2), \end{aligned}$$

where τ^2 defines the global variance component, λ_j^2 the local variance component for each coefficient $j = 1, \ldots, P$, and $\pi(\lambda_j^2)$ the prior distribution on λ_j^2 defining a particular type of global-local shrinkage prior. Polson and Scott (2010) provide an extensive overview of sparsity priors that can be achieved with the above formulation of global-local scale mixture priors by changing the prior formulation on λ_j^2. For example, the horseshoe prior (Carvalho et al., 2010) is defined by placing a Half-Cauchy prior on $\lambda_i^2 \sim C^+(0, 1)$. Hence, in the special case of the horseshoe prior the global variance parameter τ^2 is responsible for pulling all coefficients towards zero while letting some coefficients escape the shrinkage due to the thick tails of the Half-Cauchy distribution for the local variance parameters λ_j^2. Different levels of the global variance parameter τ^2 allow for varying degrees of sparsity in the model where a large value correlates with very little shrinkage and a small value induces a lot of regularisation (Piironen and Vehtari, 2017).

1.5 Approximate Posterior Inference

Posterior inference on the latent variables or model parameters $\boldsymbol{\psi}$ while accounting for the observed data $\boldsymbol{y}$ is most commonly performed via MCMC sampling through methods, such as Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970), Hamiltonian Monte Carlo (Girolami and Calderhead, 2011) or Gibbs algorithm (Geman and Geman, 1984), which acquire an approximate posterior density $p(\boldsymbol{\psi}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\psi})p(\boldsymbol{\psi})}{p(\boldsymbol{y})}$, where $p(\boldsymbol{y}|\boldsymbol{\psi})$ defines the likelihood, $p(\boldsymbol{\psi})$ the prior, and $p(\boldsymbol{y})$ the model evidence, by implementing the following steps:

1) Constructing an ergodic Markov chain on $\boldsymbol{\psi}$ whose stationary distribution is the posterior distribution $p(\boldsymbol{\psi}|\boldsymbol{y})$.

2) Sampling from the chain in order to acquire a collection of samples from the stationary distribution.

3) Approximating posterior distribution with summary statistics derived from the collection of samples.

While MCMC sampling is the gold standard for most applications due to guarantees of (asymptotically) exact samples from the target density, it can be prohibitive if it takes too long to reach the stationary distribution, a scenario which can occur if sample sizes are too large or models too complex (Blei et al., 2017). Neuroimaging applications in combination with large-scale population health datasets exhibit both of these issues and hence require faster approximate posterior inference techniques, such as variational inference, that can manage the complex model structures and can also be scaled to datasets of the size of the UKBB or the ABCD study. In the following subsections, we will describe how to acquire approximate posterior densities through the application of variational inference, a method which we utilise throughout this thesis for posterior inference.

1.5.1 Mean-field Variational Inference

Variational inference (Jordan et al., 1999; Wainwright and Jordan, 2008) approximates posterior densities through optimisation by applying the following steps:

1) Identify a family of approximate densities, called variational densities, $\mathcal{Q}$.

2) Find a member of that family $q^*(\boldsymbol{\psi}) \in \mathcal{Q}$ which matches the posterior $p(\boldsymbol{\psi}|\boldsymbol{y})$ as closely as possible by minimising the Kullback Leibler (KL) divergence between those two distributions

$$q^*(\boldsymbol{\psi}) = \underset{q(\boldsymbol{\psi}) \in \mathcal{Q}}{\arg\min}\, \mathrm{KL}(q(\boldsymbol{\psi})||p(\boldsymbol{\psi}|\boldsymbol{y})).$$

3) Acquire approximate posterior density through the optimised variational density $q^*(\boldsymbol{\psi})$.

The complexity of the optimisation problem is defined by the choice of variational family $\mathcal{Q}$ with the trade-off between choosing a flexible member of the family which captures the target posterior density well and selecting a simple variational family which facilitates efficient computation (Blei et al., 2017). Throughout this work, we focus on simpler variational families, such as mean-field variational inference (MFVI) (Bishop, 2006), instead of more flexible families, such as structured variational inference (Hoffman and Blei, 2015) or mixtures of variational densities (Bishop, 1998). In MFVI, the parameters of interest $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_M)^T$ are assumed to be mutually independent and hence can each be estimated by optimising an individual factor in the variational density $q(\boldsymbol{\psi})$, so that

$$q(\boldsymbol{\psi}) = \prod_{j=1}^{M} q_j(\psi_j),$$

for each density $q_j(\psi_j)$.

The KL divergence is defined by

$$\begin{aligned}\mathrm{KL}(q(\boldsymbol{\psi})||p(\boldsymbol{\psi}|\boldsymbol{y})) &= \mathbb{E}_q\left[\ln\left\{q(\boldsymbol{\psi})\right\}\right] - \mathbb{E}_q\left[\ln\left\{p(\boldsymbol{\psi}|\boldsymbol{y})\right\}\right] \\ &= \mathbb{E}_q\left[\ln\left\{q(\boldsymbol{\psi})\right\}\right] - \mathbb{E}_q\left[\ln\left\{p(\boldsymbol{y}|\boldsymbol{\psi})p(\boldsymbol{\psi})\right\}\right] + \ln\left\{p(\boldsymbol{y})\right\},\end{aligned}$$

where the expectations are taken with respect to the variational density $q(\boldsymbol{\psi})$. Substituting in Bayes rule for the posterior density $p(\boldsymbol{\psi}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\psi})p(\boldsymbol{\psi})}{p(\boldsymbol{y})}$ in the KL-divergence results in an optimisation objective which is usually not computable as it requires the computation of the log model evidence $\ln\{p(\boldsymbol{y})\}$. Hence, variational inference employs a different optimisation objective, the so called evidence lower bound (ELBO), where the maximisation of the ELBO $\mathcal{L}(q)$ is equivalent to the minimisation of the KL-divergence $\mathrm{KL}(q(\boldsymbol{\psi})||p(\boldsymbol{\psi}|\boldsymbol{y}))$ up to an added constant and is defined via Jensen's inequality (Jordan et al., 1999) by

$$\begin{aligned}\ln\{p(\boldsymbol{y})\} &\geq \mathcal{L}(q) \\ &= \mathbb{E}_q[\ln\{p(\boldsymbol{y}|\boldsymbol{\psi})\}] + \mathbb{E}_q[\ln\{p(\boldsymbol{\psi})\}] - \mathbb{E}_q[\ln\{q(\boldsymbol{\psi})\}] \\ &= \mathbb{E}_q[\ln\{p(\boldsymbol{y}|\boldsymbol{\psi})\}] - \mathrm{KL}(q(\boldsymbol{\psi})||p(\boldsymbol{\psi}))\end{aligned}$$

as the combination of expected likelihood and negative KL-divergence between the variational density $q(\boldsymbol{\psi})$ and the prior $p(\boldsymbol{\psi})$. Hence, the ELBO lower bounds the log of the model evidence as $\ln\{p(\boldsymbol{y})\} \geq \mathcal{L}(q)$ which holds for every $q(\boldsymbol{\psi})$ due to $\mathrm{KL}(\cdot||\cdot) \geq 0$. Hence, the above equations highlight that the log model evidence can be re-expressed through adding the KL-divergence $\mathrm{KL}(q(\boldsymbol{\psi})||p(\boldsymbol{\psi}|\boldsymbol{y}))$ and the ELBO $\mathcal{L}(q)$ together showcasing that maximising the ELBO is proportional to minimising the KL-divergence up to an added constant (Blei et al., 2017; Jordan et al., 1999).

1.5.2 Coordinate Ascent Variational Inference

The algorithm for solving the optimisation problem defined by variational inference is called coordinate ascent variational inference (CAVI) (Bishop, 2006); see Algorithm 1 for a step-by-step description. CAVI is an iterative algorithm which alternates the following steps until convergence, defined as the difference in ELBO values between two consecutive iterations dropping below a pre-defined convergence threshold ϵ:

1) For a factor j fix the other variational factors $\prod_{j \neq \ell} q_\ell(\psi_\ell)$.

2) Update the factor $q_j(\psi_j)$ via the exponentiated expected log of the complete conditional

$$q_j(\psi_j) \propto \exp\left\{\mathbb{E}_{-j}\left[\ln\{p(\psi_j|\boldsymbol{\psi}_{-j}, \boldsymbol{y})\}\right]\right\},$$

   where the complete conditional is given by the conditional density given all the other parameters and the expectation is taken with respect to the variational densities not equal to j.

3) Repeat steps 1) and 2) for all j and then iterate the whole process until convergence is reached.

Algorithm 1: Mean-field CAVI (adapted from Blei et al. (2017)).

```
Input: Joint $p(\boldsymbol{\psi}, \boldsymbol{y})$, data $\boldsymbol{y}$.
Output: Variational density $q(\boldsymbol{\psi}) = \prod_{j=1}^{M} q_j(\psi_j)$.
Initialisation: Variational factors $q_j(\psi_j)$, $\epsilon$ convergence threshold.
while $ELBO - ELBO_{old} < \epsilon$ do
    for $j \in \{1, \ldots, M\}$ do
        Set $q_j(\psi_j) \propto \exp\{\mathbb{E}_{-j}[\ln\{p(\psi_j|\boldsymbol{\psi}_{-j}, \boldsymbol{y})\}]\}$.
    Compute ELBO $\mathcal{L}(q) = \mathbb{E}_q[\ln\{p(\boldsymbol{\psi}, \boldsymbol{y})\}] - \mathbb{E}_q[\ln\{q(\boldsymbol{\psi})\}]$.
Return: Optimised variational density $q^*(\boldsymbol{\psi})$.
```

While variational inference can provide fast approximations to posterior inference problems, it is known to underestimate posterior variance (Yao et al., 2018). The under-penalisation of approximations with too-light tails leads to poor uncertainty quantification and hence should be used with care for this purpose. Moreover, variational inference is sensitive to parameter initialisation and convergence can be difficult to assess (Blei et al., 2017; Yao et al., 2018).

1.6 Contributions of Thesis

The main contribution of this thesis is the development of scalable Bayesian spatial models with shrinkage or sparsity priors for massive neuroimaging datasets. We achieve scalability to the large sample sizes of population health-based studies and large number of parameters in imaging applications by deriving variational inference and approximate posterior sampling algorithms. Moreover, we account for the spatial dependence between neighbouring locations in an MRI scan and for the sparse nature of effects in an image by proposing structured spike-and-slab priors or relaxed thresholded Gaussian process priors. There are three main chapters, the first two describe an image-on-scalar regression problem with an application to lesion mapping in the UKBB and the last one describes a scalar-on-image regression with an application to task-based fMRI in the ABCD study.

In Chapter 2, we develop a Bayesian spatial image-on-scalar regression model with a structured spike-and-slab prior, called BLESS, which we apply to a large-scale lesion mapping study, the UK Biobank. Within this work, we also utilise dynamic posterior exploration

In Chapter 3, we extend BLESS to provide more accurate uncertainty quantification of spatially varying coefficients by employing Bayesian bootstrap methods and a class of jittered spike-and-slab priors. Furthermore, we develop cluster size-based imaging statistics, such as credible intervals of cluster size and measures of reliability of cluster occurrence.

In Chapter 4, we develop a Bayesian scalar-on-image regression model with a relaxed-thresholded Gaussian process prior where we derive a variational approximation to the posterior. In our real data application, we apply the model to cortical surface data to identify associations between intelligence and task-based functional MRI, which evaluates the cognitive abilities of the participants performing the task, in the ABCD study.

2

Scalable Image-on-Scalar Regression with a Structured Spike-and-Slab Prior

Contents

In this chapter, we develop a hierarchical Bayesian spatial model with a structured spike-and-slab prior for the analysis of binary lesion masks. The spike-and-slab prior provides Bayesian variable selection, and we also propose a variational approximation to the posterior for parameter estimation and inference. The variational approximation allows us to scale our approach to large sample sizes and high-dimensional image-based regression problems. Lastly, we provide extensive simulation studies and a real data application to the UK Biobank where we aim to identify associations between lesion incidence and age while accounting for confounding variables, such as sex, head size scaling factor and age-by-sex interaction.

2.1 Introduction

2.1.1 Motivation for Analysis of Lesion Masks

Magnetic resonance imaging (MRI) is a non-invasive imaging technique to study human brain structure and function (see Section 1.1 for a review). Accumulated damages to the white matter, known as lesions, appear as localised hypo-/ hyperintensities in MRI scans (Wardlaw et al., 2013). The total burden of these lesions is often associated with cognitive disorders, ageing and cerebral small vessel disease (Wardlaw et al., 2013; Wardlaw et al., 2015). Lesion prevalence is higher for older adults (Griffanti et al., 2018) and for individuals with cerebrovascular risk factors, such as hypertension, alcohol consumption or smoking history (Rostrup et al., 2012). White matter lesions are also an overall indicator of poor brain health and have been found to triple the risk of stroke and double the risk of dementia and death, and are associated with cognitive impairment, functional decline, sensory changes or motor abnormalities (Debette and Markus, 2010). Not all white matter lesions however are attributed to ageing or an increased cerebrovascular risk burden. For example, white matter hyperintensities can also occur due to multiple sclerosis, Alzheimer's disease or as a result of a stroke (Debette and Markus, 2010; Prins and Scheltens, 2015). While white matter lesions due to vascular origin are a result of chronically reduced blood flow and incomplete infarction leading to altered cerebral autoregulation, the non-vascular demyelination as seen in multiple sclerosis is caused by an autoimmune response against myelin proteins (Sharma and Sekhon, 2021). Regardless of etiology, an important clinical feature is the spatial location of lesions; while noting that lesions exhibit a high level of variability, together with the size and number of lesions, for both between and within subjects, as seen in the binary lesion masks in Figure 2.1. Elderly patients tend to present scattered lesions which later form to confluent lesions whereas white matter lesions of non-vascular origin have a particularly heterogeneous presentation where the disease course can result in rapid progression or alternation between relapses and remissions (Sharma and Sekhon, 2021). Identifying spatial locations in the brain where lesion incidence is associated with different covariates (e.g. age, hypertension, cardiovascular disease) is known as lesion mapping and is an essential tool to locate the brain regions that are particularly vulnerable to damage from various risk factors and inform development of interventions to reduce incidence or severity of disease (Veldsman et al., 2020).

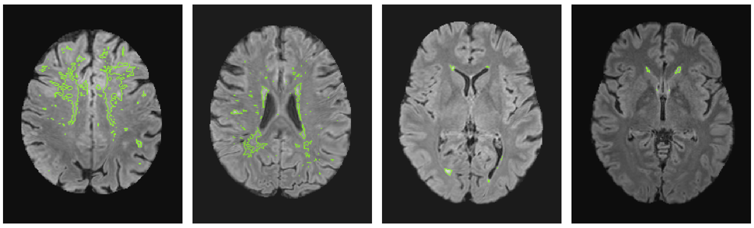


Figure 2.1: Contours of binary lesion masks from four healthy subjects from the UK Biobank with varying lesion numbers and sizes, where green outlines indicate lesions which show the heterogeneity of lesion incidence at various 2D axial slices from the 3D lesion mask, see Section 1.1.2 for further detail.

2.1.2 Mass-univariate Methods and Other Spatial Models

Lesion mapping falls under the category of image-on-scalar regression problems, where the output is an image and the input is a scalar quantity, for which we provide an overview in Section 1.3.2 in the Introduction. In this section, we focus on approaches which are commonly, but not exclusively, used for lesion mapping. The standard practice for lesion mapping is mass-univariate (Rostrup et al., 2012). In this approach a logistic regression model is fitted at each voxel or spatial location independently, any form of spatial dependence among neighbouring locations is ignored. Moreover, most methods fail to address the problem of complete separation which often occurs in logistic regression models when the output variable separates a subject-specific predictor variable or a combination of input features perfectly and hence leads to infinite and biased maximum-likelihood estimates (Firth, 1993). This problem can be addressed with a logistic regression approach known as Firth

Regression, which utilises a penalised likelihood approach and produces mean-bias reduced parameter estimates (Firth, 1993; Kosmidis et al., 2020).

Bayesian spatial models on the other hand are capable of accounting for the spatial dependence structure among neighbouring voxels in a single joint model. For example, Ge et al. (2014) have developed a Bayesian spatial generalised linear mixed model (BSGLMM) with a probit link function where the probability of lesion presence is modelled via a linear combination of fixed and random effects and subject-specific covariates. BSGLMM places a spatial smoothing prior directly on the parameters, specifically a conditional autoregressive model prior (Besag, 1974), which may induce bias due to oversmoothing of regression coefficients. Moreover, BSGLMM relies on MCMC methods for posterior computation which do not scale well to the large sample sizes found in the UK Biobank.

In order to address the limitations of previous methods, we propose a multivariate Bayesian model for lesion mapping in large-scale epidemiological studies that (1) uses variable selection and shrinkage priors, (2) takes into account the spatial dependency through a parameter that controls the level of sparsity rather than directly smoothing regression coefficients, and (3) relies on variational inference for an approximation to the posterior. Hence, this allows us to fit the model to thousands of subjects and appropriately account for the spatial dependency in lesion mapping studies containing over 50,000 voxel locations. We also want to acknowledge that other model choices in the literature may better capture the association between lesions and covariates (Li et al., 2020; Zeng et al., 2022; Whiteman, 2022); however, we favour a model that enables us to scale parameter estimation and inference to large-scale epidemiological studies, see Section 1.3 for a detailed literature review.

2.1.3 Bayesian Variable Selection

In this section, we provide a short overview on the literature of Bayesian variable selection. For a more thorough review, please see Section 1.4. We utilise Bayesian variable selection to improve brain lesion mapping by shrinking small coefficients towards zero, thus helping with prediction, interpretation and reduction of spurious associations in high-dimensional settings. A commonly applied technique for Bayesian variable selection is spike-and-slab regression which aims to identify a subset of predictors within a regression model. The original spike-and-slab mixture prior places a mixture of a point mass at zero and a diffuse distribution on the coefficients (Mitchell and Beauchamp, 1988). George and McCulloch (1993) and George and McCulloch (1997) have increased the computational feasibility of spike-and-slab regressions by introducing a continuous mixture of Gaussians formulation where the spike distribution is defined by a normal distribution with a small variance rather than a point mass prior. The binary latent variable, sampled from a Bernoulli distribution with an inclusion probability, determines which mixture component a variable belongs to and enables variable selection. Overall, the options of continuous shrinkage priors in the literature are large, see Piironen and Vehtari (2017) for a comparison of different methods.

The spike-and-slab regression is also able to incorporate spatial information, replacing the exchangeable Bernoulli prior on the inclusion indicator variables, with a structured spatial prior using a vector of inclusion probabilities. Previous examples of introducing structure within a spike-and-slab regression include the placement of a logistic regression product prior (Stingo et al., 2010) on the latent variables in order to group biological information for a genetics application, an Ising prior which incorporates structural information for a high-dimensional genomics application (Li and Zhang, 2010) or a structured spike-and-slab prior with a spatial Gaussian process prior (Andersen et al., 2014).

2.1.4 Approximate Posterior Inference

The gold standard of parameter estimation and inference for spike-and-slab regression with a continuous mixture of Gaussians prior is Gibbs sampling (George and McCulloch, 1993). However, in high-dimensional regression settings as well as large sample size scenarios other more scalable approximate methods are required due to the intense computational burden.

Expectation propagation (EP) (Minka, 2013) or variational inference (Jordan et al., 1999) algorithms redefine the problem of approximating densities through optimisation (Blei et al., 2017). Both of these methods have been extensively studied for spike-and-slab regression problems (Hernández-Lobato et al., 2013; Carbonetto and Stephens, 2012). The EP algorithm however poses several challenges as it is computationally intensive for even moderate sample sizes, there is no guarantee of convergence, and its poor performance for multimodal posteriors due to the problematic need to incorporate all modes in its approximation (Bishop, 2006). Poor variational approximations can arise due to slow convergence, a simplistic choice of variational families, or due to underestimation of the posterior variance as the KL-divergence tends to under-penalise thin tails (Yao et al., 2018).

2.1.4.1 Local Variational Approximations

For a review on variational inference, we refer the reader to Section 1.5. In this work we make use of local variational approximations. Generally, local approximations identify bounds on functions over individual variables within a model compared to the previously described global approach to variational inference which identifies a bound on a function over all random variables within the model (Bishop, 2006). In our work, we specifically require a bound on the logistic sigmoid function which is defined by

$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$

and is neither convex nor concave. Jaakkola and Jordan (2000) develop a variational approximation to the posterior for logistic regression problems, where the logistic sigmoid function appears in the likelihood. The lower bound on the sigmoid function can be represented through the functional form of a Gaussian where firstly the sigmoid function is transformed by taking the log and then by decomposing the function, so that

$$\begin{aligned}\ln\{\sigma(x)\} &= -\ln\{1+\exp(-x)\}\\ &= -\ln\{\exp(-x/2)[\exp(x/2)+\exp(-x/2)]\}\\ &= x/2 - \ln\{\exp(x/2)+\exp(-x/2)\},\end{aligned}$$

where the latter term $f(x) = -\ln\{\exp(x/2)+\exp(-x/2)\}$ is a convex function of the variable x^2. By utilising the framework of complex duality (Rockafellar, 1972; Jordan et al., 1999) one can approximate a convex function $f(x)$ by a simpler linear function $\lambda x - f(x)$ where the tightest bound is obtained by optimising the introduced variational parameter λ which identifies the slope of the linear function. The optimal linear function is given by the tangent line to the function $f(x)$ by identifying the intercept $g(\lambda) = \max_x\{\lambda x - f(x)\}$. Hence, in the case of the logistic sigmoid function we identify the lower bound for the now convex function $f(x)$ by using the conjugate function which is defined as

$$g(\lambda) = \max_{x^2}\left\{\lambda x^2 - f(\sqrt{x^2})\right\}$$

and the stationarity condition which yields

$$0 = \lambda - \frac{dx}{dx^2}\frac{d}{dx}f(x) = \lambda + \frac{1}{4x}\tanh\left(\frac{x}{2}\right).$$

By rearranging above equation, one can define ξ as a new variational parameter in lieu of λ where the value x corresponds to the contact point on the tangent line for a particular value λ. This definition leads to the following equation:

$$\lambda(\xi) = -\frac{1}{4\xi}\tanh\left(\frac{\xi}{2}\right) = -\frac{1}{2\xi}\left[\sigma(\xi) - \frac{1}{2}\right].$$

Ultimately, this expression also allows for (1) the simplification of the conjugate function $g(\lambda)$, (2) the redefinition of the bound $f(x)$, and (3) the final expression of the bound on the sigmoid function via the variational parameter ξ, so that

$$\begin{aligned}&(1)\ g(\lambda) = \lambda(\xi)\xi^2 - f(\xi) = \lambda(\xi)\xi^2 + \ln\{\exp(\xi/2)+\exp(-\xi/2)\}\\ &(2)\ f(x) \geq \lambda x^2 - g(\lambda) = \lambda x^2 - \lambda\xi^2 - \ln\{\exp(\xi/2)+\exp(-\xi/2)\}\\ &(3)\ \sigma(x) \geq \sigma(\xi)\exp\left\{(x-\xi)/2 - \lambda(\xi)(x^2-\xi^2)\right\}.\end{aligned}$$

The bound now has the form of an exponential of a quadratic function of x which lends itself to acquiring a conjugate Gaussian posterior defined through a logistic function and when using a Gaussian prior on the parameters. For more details on this approach, we refer the reader to a description by Bishop (2006) and the original work by Jaakkola and Jordan, 2000.

2.2 Methods

Our model for **B**ayesian **L**esion **E**stimation with a **S**tructured **S**pike-and-Slab (**BLESS**) prior is formulated as a Bayesian spatial hierarchical generalised linear model. While we specifically focus on neuroimaging applications within this thesis, the model can be applied to any form of spatial binary data on a lattice and can equally be extended to various neuroimaging modalities other than lesion masks.

Throughout this thesis we use boldface to indicate a vector or matrix. We model the binary data $y_i(s_j)$ for every subject $i = 1, \dots, N$ at voxel location $s_j \in \mathcal{B} \in \mathbb{R}^3$, $j = 1, \dots, M$, with a Bernoulli random variable with lesion probability $p_i(s_j)$. For computational reasons, we choose to model the binary data via a probit link function which defines the relationship between the conditional expectation $\eta_i(s_j)$ and the linear combination of input features $\boldsymbol{x}_i$ containing P subject-specific covariates, spatially varying parameters $\boldsymbol{\beta}(s_j)$ and a spatially varying intercept $\beta_0(s_j)$. Despite the availability of the bound for the logit, described in Section 2.2.2, it is only an approximation that breaks the conjugacy in the later stages of the hierarchical model and hence we use the probit link in combination with a data augmentation approach for defining the likelihood. While the data comprise an image for each subject, we store the data as unraveled M-vectors $\boldsymbol{y}_i$ for each subject or N-vectors $\boldsymbol{y}(s_j)$ for each voxel.

The Bayesian spatial generalised linear model for subject i at location s_j is specified as

$$[y_i(s_j)|p_i(s_j)] \sim \text{Bernoulli}[p_i(s_j)] \tag{2.1}$$

$$\Phi^{-1}\left(p_i(s_j)\right) = \eta_i(s_j) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(s_j) + \beta_0(s_j), \tag{2.2}$$

where the equations reflect the random and systematic component, respectively and the link function is given by the cumulative Gaussian density $\boldsymbol{\Phi}(\cdot)$.

Furthermore, we reparameterise the Bayesian probit regression model defined in Equation (2.1) and (2.2) exactly via the data augmentation approach by Albert and Chib (1993) by introducing latent normal variables in Equation (2.3) and (2.4) into the model in order to ease the computational complexity. This approach assumes that the probit regression has an underlying normal regression structure on latent continuous data. These independent continuous latent variables $z_i(s_j)$ for every subject $i = 1, \dots, N$ and spatial location $j = 1, \dots, M$ are drawn from the following normal distribution

$$z_i(s_j)|\eta_i(s_j) \sim \mathcal{N}(\eta_i(s_j), 1) \tag{2.3}$$

where the conditional probability of $y_i(s_j) = 1$ is given by

$$\Pr[y_i(s_j) = 1|z_i(s_j)] = \begin{cases} 1, & z_i(s_j) > 0, \\ 0, & z_i(s_j) \leq 0. \end{cases} \tag{2.4}$$

2.2.1 Prior Specifications

We build a Bayesian hierarchical regression model by placing a continuous version of a spike-and-slab prior on the spatially varying P-coefficient vector $\boldsymbol{\beta}(s_j)$. The continuous mixture of Gaussians with two different variances, consisting of the spike and the slab distribution, is given by

$$\beta_p(s_j) \mid \gamma_p(s_j) \ \sim \ \mathcal{N}(0, \ \nu_0 \left[1 - \gamma_p(s_j)\right] \ + \ \nu_1 \gamma_p(s_j)), \tag{2.5}$$

where $\gamma_p(s_j)$ is a latent binary indicator variable for covariate $p = 1, \dots, P$ and locations $j = 1, \dots, M$, ν_0 is the spike variance and ν_1 is the slab variance which determine the amount of regularisation. Variable selection is implemented via the latent variables $\gamma_p(s_j)$ localising the spatial effect of each variable. Due to the continuous spike-and-slab specification, the variance within the spike distribution is always $\nu_0 > 0$ which ensures the continuity of the spike distribution and therefore the derivation of closed form solutions of the variational parameter updates. The slab variance on the other hand is set to a fixed value to include the range of all possible values of the spatially varying coefficients. The combination of a small spike variance and a large slab variance with latent indicator variables for every covariate and location introduces a selective spatial shrinkage property that shrinks smaller coefficients close to zero and leaves the large parameters unaffected.

In order to account for the spatial dependence across the brain, we place an independent logistic regression prior with non-exchangeable inclusion probabilities on the latent binary indicator variables $\boldsymbol{\gamma}(s_j)$ sampled from a Bernoulli distribution, similar to Stingo et al. (2010). The prior is non-exchangeable because we incorporate structural information via the sparsity parameter $\boldsymbol{\theta}(s_j) \in \mathbb{R}^P$ which ensures that certain voxel locations are more likely to be included in the model than others. In the context of brain imaging data this means that voxels that are nearby each other are expected to have a similar inclusion probability. Specifically, we model the latent variables $\gamma_p(s_j)$ via

$$\gamma_p(s_j) \mid \theta_p(s_j) \;\sim\; \text{Bernoulli}(\sigma\left[\theta_p(s_j)\right]), \tag{2.6}$$

where $\sigma(\cdot)$ is a sigmoid function.

The hierarchical spatial regression model is completed by placing a spatial prior on the sparsity parameter $\boldsymbol{\theta}^{\mathrm{T}} = \left[\boldsymbol{\theta}^{\mathrm{T}}(s_1), \ldots, \boldsymbol{\theta}^{\mathrm{T}}(s_M)\right]$: a length PM column vector. We choose a multivariate conditional autoregressive (MCAR) prior as a spatial prior for computational reasons (Gelfand and Vounatsou, 2003). Alternative priors could be considered in lieu of this type of a simple smoothing prior; however, this would significantly increase the computational complexity of the model at hand.

The full conditional distribution for $\boldsymbol{\theta}(s_j)$ is given by the following multivariate normal distribution and utilises the notation defined by Mardia (1988):

$$\left[\boldsymbol{\theta}(s_j) \mid \boldsymbol{\theta}(-s_j), \boldsymbol{\Sigma}^{-1}\right] \;\sim\; \text{MVN}\left(\frac{\sum_{s_r \in \partial s_j} \boldsymbol{\theta}(s_r)}{n(s_j)},\; \frac{\boldsymbol{\Sigma}}{n(s_j)}\right), \tag{2.7}$$

where $\boldsymbol{\Sigma}$ is a symmetric positive definite smoothing matrix. The sum $\sum_{s_r \in \partial s_j}$ defines the sum over the neighbourhood voxels at location s_j, ∂s_j defines the set of neighbours at location s_j and $n(s_j)$ is the cardinality of the neighbourhood set. For our MRI scans we consider only neighbours sharing a face, so therefore most of the interior of the brain has $n(s_j) = 6$ neighbours whereas locations near the brain mask have $n(s_j) < 6$.

We then describe the joint distribution over the sparsity parameters, up to a proportionality constant, by utilising Brooks's lemma (Brook, 1964) which is given by:

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{\Sigma}) \propto \exp\left\{ -\frac{1}{2} \sum_{s_j \sim s_{j'}} [\boldsymbol{\theta}(s_j) - \boldsymbol{\theta}(s_{j'})]^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} [\boldsymbol{\theta}(s_j) - \boldsymbol{\theta}(s_{j'})] \right\}, \tag{2.8}$$

where the sum $\sum_{s_j \sim s_{j'}}$ describes the sum over neighbourhood voxels, and $s_j \sim s_{j'}$ indicates that s_j and $s_{j'}$ are neighbours. This joint prior distribution is improper and not identifiable according to Besag (1986). However, the posterior of $\boldsymbol{\theta}$ is proper, if there is information in the data with respect to the sparsity parameters. Lastly, we finish specifying the Bayesian hierarchical regression model by placing an uninformative, conjugate Wishart prior over the precision matrix $\boldsymbol{\Sigma}^{-1}$ to fully specify the model with

$$\boldsymbol{\Sigma}^{-1} \sim \mathrm{Wishart}(\nu,\ \boldsymbol{I}), \tag{2.9}$$

where the degrees of freedom are given by $\nu = P$ and the scale matrix is defined by the identity matrix $\boldsymbol{I}$ (Ge et al., 2014).

2.2.2 Variational Posterior Approximation

A variational approximation to the posterior allows for using optimisation instead of MCMC sampling for a more scalable approach to posterior inference. We opt for variational inference due to the non-conjugacy in the hierarchical model induced by specifying a logistic function around the sparsity parameters $\boldsymbol{\theta}$ in the inclusion probabilities of the spike-and-slab priors. Local variational approximations solve this problem by finding a bound on an individual set of variables via a first-order Taylor approximation (Jaakkola and Jordan, 2000). For general variational inference, we then require the full joint distribution of the Bayesian spatial regression model, consisting of the likelihood $p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\beta_0})$ and the joint prior $p(\boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\beta_0}, \gamma, \boldsymbol{\theta}, \boldsymbol{\Sigma}^{-1})$, which is given by

$$p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\beta}_0, \gamma, \boldsymbol{\theta}, \boldsymbol{\Sigma}^{-1}) = p(\boldsymbol{Y}|\boldsymbol{Z})p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\beta}_0)p(\boldsymbol{\beta}_0)p(\boldsymbol{\beta}|\gamma)p(\gamma|\boldsymbol{\theta}) \\ p(\boldsymbol{\theta}|\boldsymbol{\Sigma}^{-1})p(\boldsymbol{\Sigma}^{-1}). \tag{2.10}$$

We write the entire set of model parameters as $\boldsymbol{\Psi} = \{\boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\beta_0}, \gamma, \boldsymbol{\theta}, \boldsymbol{\Sigma}^{-1}\}$ where the conditional distribution of each model parameter $\boldsymbol{\psi}$, where the elements in the model parameters are indexed by j, is obtained as $p(\boldsymbol{\psi}|\boldsymbol{y}) = \frac{p(\boldsymbol{\psi},\boldsymbol{y})}{p(\boldsymbol{y})}$. We acquire an approximation to the exact posterior by firstly specifying a family of densities $\mathcal{Q}$ over each element of the model parameter ψ_j and secondly identifying the parameters of the candidate distribution $q(\psi_j) \in \mathcal{Q}$ that minimises the Kullback-Leibler (KL) divergence, given by

$$q^*(\psi_j) = \underset{q(\psi_j)\in\mathcal{Q}}{\arg\min}\, \mathrm{KL}\ \{q(\psi_j)\ ||\ p(\psi_j|\boldsymbol{y})\}. \tag{2.11}$$

We aim to minimise the difference between the exact posterior $p(\psi_j|\boldsymbol{y})$ and the variational distribution $q(\psi_j)$ to find the best approximate distribution $q^*(\psi_j)$. Hence as part of the variational approach, rather than computing the KL-divergence which contains the log-marginal of the data, a quantity that is often not computable, the evidence lower bound (ELBO) $\mathcal{L}(q)$ (Blei et al., 2017)

$$\ln\{p(\boldsymbol{y})\} \geq \mathcal{L}(q) = \mathbb{E}_{q(\boldsymbol{\Psi})}\left[\ln\{p(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Psi})\}\right] - \mathbb{E}_{q(\boldsymbol{\Psi})}\left[\ln\{q(\boldsymbol{\Psi})\}\right] \tag{2.12}$$

is optimised instead. The derivation of the variational distributions and the ELBO can be found in Appendix Section A.1. The variational density $q_j(\psi_j)$ is derived by taking the exponentiated expected log of the complete conditional given all the other parameters and the data which is defined by $q_j(\psi_j) \propto \exp\{\mathbb{E}_{-j}[\log\{p(\psi_j|\boldsymbol{\psi}_{-j}, \boldsymbol{X})\}]\}$ where the expectation is over the fixed variational density of other variables $\boldsymbol{\psi}_{-j}$, given by $\prod_{\ell\neq j} q_\ell(\psi_\ell)$. By determining the variational distributions q, we successively update each parameter $\boldsymbol{\psi}$, while holding the others fixed, via mean-field coordinate ascent variational inference (Bishop, 2006). Further details on initialisation and convergence of variational inference are also found in the Appendix Section A.1.

2.2.3 Dynamic Posterior Exploration

Dynamic posterior exploration (DPE) (Ročková and George, 2014) is an annealing-like strategy which fixes the slab variance to a large, fixed value. The procedure works by starting in a smooth posterior landscape and aims to discover a sparse, multimodal posterior by gradually decreasing the value of the spike parameter until it approximates the spike-and-slab point mass prior. When the starting spike variance is large, we should be able to easily identify a small set of local optima by maximising the ELBO. Thereafter, the technique uses the result as a warm start for the next optimisation with a reduced spike variance which leads to a more peaked posterior until the last value within a range of spike variances is evaluated and a stable solution to the optimisation problem is found.

Algorithm 2: DPE for BLESS-VI

```
Result: Estimated parameter estimates from variational posterior
        distribution by performing DPE.
Set: $V = \{\nu_0^{(1)}, \dots, \nu_0^{(K)}\}$: sequence of spike variances ; $\nu_1$: large, fixed slab
  variance; $\boldsymbol{\psi}^{(K)}$: initialised parameters ($\boldsymbol{\psi} = \{\boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\Sigma}^{-1}\}$); $\epsilon$:
  convergence criterion; $K$: number of spike variances in DPE sequence
for $k = K, \dots, 1$ do
    Set current spike variance $\nu_0 = \nu_0^{(k)}$ in DPE sequence.
    while $ELBO - ELBO_{old} < \epsilon$ do
        for $j \in \{1, \dots, M\}$ do
            Set $q_j(\psi_j) \propto \exp\{\mathbb{E}_{-j}[\ln\{p(\psi_j|\boldsymbol{\psi}_{-j}, \boldsymbol{y})\}]\}$.
        Compute ELBO $\mathcal{L}(q) = \mathbb{E}_q[\ln\{p(\boldsymbol{\psi}, \boldsymbol{y})\}] - \mathbb{E}_q[\ln\{q(\boldsymbol{\psi})\}]$.
    Set initial values for next DPE iteration to variational posterior mean of
      optimised variational densities $\hat{\boldsymbol{\psi}}^{(k)}$ and use as warm start.
    Return: Optimised variational densities $q^*(\cdot)$ at spike variance $\nu_0^{(1)}$ in
      the DPE sequence.
```

The process of dynamic posterior exploration can be split into three parts, see the description below and Algorithm 2. Firstly, we perform parameter estimation via variational inference over a sequence of K increasing spike variances $\nu_0 \in V = \{\nu_0^{(1)}, \dots, \nu_0^{(K)}\}$. After the initial evaluation of the backwards DPE procedure with $\nu_0^{(K)} \leq \nu_1$, every subsequent optimisation is run with a successively smaller ν_0 and initialised with the previously estimated variational parameters as a “warm start” solution. Secondly, the output of every optimisation run within the sequence of spike parameter values V is thresholded via the posterior inclusion probabilities. The thresholding rule for BLESS is based on the following inclusion probabilities

$$\hat{\gamma}_p(s_j) = \begin{cases} 1, & \text{if } P(\hat{\gamma}_p(s_j) = 1 | \boldsymbol{Y}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{0}}, \hat{\boldsymbol{\theta}}) > 0.5, \\ 0, & \text{if } P(\hat{\gamma}_p(s_j) = 1 | \boldsymbol{Y}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{0}}, \hat{\boldsymbol{\theta}}) \leq 0.5, \end{cases} \tag{2.13}$$

which is equivalent to the local version of the median probability model defined by Barbieri and Berger (2004) and Barbieri et al. (2021). Furthermore, the determination of active versus inactive voxels based on the inclusion probability $P(\hat{\gamma}_p(s_j) = 1 | \boldsymbol{y})$ is equivalent to thresholding the parameter estimates $\hat{\boldsymbol{\beta}}$ themselves where the threshold is given by the intersection of the weighted mixture of the spike-and-slab priors (George and McCulloch, 1993; Ročková and George, 2014). For BLESS, we choose the former thresholding rule based on the posterior inclusion probabilities considering that thresholding the parameters $\hat{\boldsymbol{\beta}}$ would require the calculation of a different set of intersection points for every coefficient due to the non-exchangeable nature of the spatial prior within the inclusion probability. Thirdly, the estimated posterior with the smallest spike variance ν_0 within the range of parameters V is used. We do not assert that this ν_0 is optimal per se, but that our annealing-like strategy obviates the need for a precise determination of ν_0 as the estimates for the larger effects tend to stabilise at a particular solution of variational posterior parameters.

This behaviour can be validated by two types of plots. Regularisation plots enable the examination of the estimated coefficients over a sequence of spike variances. For each ν_0, the colour of the parameter values indicates whether or not a variable is included in (red) or excluded from (blue) the model based on the thresholded posterior probability of inclusion. Figure 2.2(a) illustrates how the negligible coefficients are drawn to zero as the values of ν_0 decrease, while the larger parameters of the active voxels stabilise and are unaffected by regularisation. Hence, for the plot in Figure 2.2(a) this occurs at a log-spike variance $\log(\nu_0) \leq -6$ where a local optimum has been identified and any further decrease in spike variance only leads to further shrinkage of the negligible coefficients.

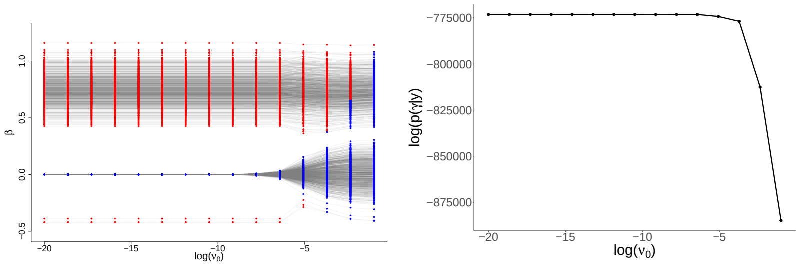

(a) Regularisation Plot **(b)** Marginal Plot

Figure 2.2: (a) Regularisation plot (active voxel: red, inactive voxel: blue) and (b) plot of marginal posterior of $\hat{\boldsymbol{\gamma}}$ under $\nu_0 = 0$ over a sequence of equidistant $\nu_0 \in V$ within log-space for the simulation study described in Section 2.3.1 for sample size $N = 1{,}000$ and base rate intensity $\lambda = 3$, where base rate intensity defines the base rate number of lesion presence in each quadrant of a simulated image. Both plots indicate that parameter estimates have stabilised past spike variances of $\log(\nu_0) \leq -6$ within the DPE procedure. The regularisation plot also shows how negligible (blue) coefficients are progressively shrunk towards 0 while the larger (red) coefficients remain almost unregularised.

A complementary plot, especially useful when overplotting makes the regularisation plot difficult to interpret, is the log-marginal posterior plot $\ln\{\pi_{\nu_0=0}(\boldsymbol{\gamma}|\boldsymbol{Y}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}^{-1})\}$ for the latent variables. The maximum value of this quantity yields the posterior closest to approximating the point mass prior which is the goal of backwards DPE. Since our model contains intractable integrals, we use a variational approximation to the marginal posterior of $\boldsymbol{\gamma}$ under the prior of $\nu_0 = 0$. We utilise Jensen's inequality to bound the marginal probability integrating out the parameters $\boldsymbol{\beta}$, $\boldsymbol{\beta_0}$ and the latent variables $\boldsymbol{Z}$ via their respective variational approximation. The other model parameters $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Sigma}}^{-1}$ are regarded as nuisance parameters. Specifically, the log-marginal posterior under $\nu_0 = 0$ and its approximation

$$\ln\{\pi_{\nu_0=0}(\boldsymbol{\gamma}|\boldsymbol{Y}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}^{-1})\} \tag{2.14}$$

$$= \ln\left\{\int\int\int q(\boldsymbol{Z}, \boldsymbol{\beta}_\gamma, \boldsymbol{\beta_0}) \frac{p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\beta}_\gamma, \boldsymbol{\beta_0}, \gamma, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}^{-1}|\boldsymbol{X}_\gamma)}{q(\boldsymbol{Z}, \boldsymbol{\beta}_\gamma, \boldsymbol{\beta_0})} d\boldsymbol{Z} d\boldsymbol{\beta}_\gamma d\boldsymbol{\beta_0}\right\} \tag{2.15}$$

$$\geq \mathbb{E}_{q(\boldsymbol{Z}, \boldsymbol{\beta}_\gamma, \boldsymbol{\beta_0})}\left[\ln\left\{p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\beta}_\gamma, \boldsymbol{\beta_0}, \gamma, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}^{-1}|\boldsymbol{X}_\gamma)\right\}\right] - \tag{2.16}$$

$$\mathbb{E}_{q(\boldsymbol{Z}, \boldsymbol{\beta}_\gamma, \boldsymbol{\beta_0})}\left[\ln\left\{q(\boldsymbol{Z}, \boldsymbol{\beta}_\gamma, \boldsymbol{\beta_0})\right\}\right],$$

where $\boldsymbol{\beta}_\gamma = \boldsymbol{\beta}\hat{\boldsymbol{\gamma}}$, can be used to determine whether or not the parameters identified as active have stabilised by checking a single quantity rather than the solution path of all parameters of the model. Figure 2.2(b) illustrates the marginal

$\ln\{\pi_{\nu_0=0}(\boldsymbol{\gamma}|\boldsymbol{Y}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}^{-1})\}$, showing a plateau for any log-spike variance $\log(\nu_0) \leq -6$ indicating a good approximation of the point mass prior. The marginal plot can be used in an equivalent manner to the regularisation plot as a sanity check for visualising the stabilisation of large effects and continued shrinkage for the negligible coefficients at the end of the annealing-like process.

2.3 Results

2.3.1 Simulation Study

In this section, we firstly explain the process of simulating lesion data where the ground truth is known but the data generating mechanism is intentionally different from the model at hand. We perform various simulation studies to assess the performance of BLESS by comparing the parameter estimates and predictive performance of our method to the mass-univariate approach, Firth regression (Firth, 1993), and the Bayesian spatial model, BSGLMM (Ge et al., 2014). For comparison, the latter is adopted to fit a Bayesian hierarchical modelling framework, similar to BLESS, where we add a spatially varying intercept $\boldsymbol{\beta}_0(s_j)$ to match the setup of BLESS. For Firth regression, which fits an independent probit regression model with a mean bias reduction for every voxel location, we use the `R` package `brglm2` (Kosmidis, 2021).

The main aim of many neuroimaging studies lies in the provision of accurate inference results. We therefore tailor the assessment of simulation studies to the evaluation of inference results rather than on coverage probabilities. We compare inference results by assessing true positive (TP), false positive (FP), true negative (TN), and false negative (FN) discoveries in the following measures: (1) sensitivity/true positive rate ($\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$), (2) true discovery rate ($\text{TDR} = \frac{\text{TP}}{\text{TP+FP}}$), (3) specificity/1 - false positive rate ($\text{FPR} = \frac{\text{FP}}{\text{FP+TN}}$), and (4) false discovery rate ($\text{FDR} = \frac{\text{FP}}{\text{FP+TP}}$). Note that all above inference evaluation criteria are averaged results across a 100 simulated datasets for increased robustness of the results, so for example $\text{FPR} = n_{\text{sim}}^{-1}\text{FPR}_k$ where $k = 1, \ldots, n_{\text{sim}}$ and n_{sim} is the number of simulated datasets. Lastly, we provide extensive simulation studies on the performance of BLESS compared to a frequentist, mass-univariate approach as well as a Bayesian spatial model with a simulation study addressing varying sample sizes N, base rate intensities λ, and sizes of effect within an image. Base rate intensities hereby provide an indicator for the magnitude of various regression coefficient effect sizes where a smaller λ value yields smaller regression coefficients.

2.3.1.1 Data Generating Process

For simulating the data, we adopt a data generating process that is different from our model in order to guarantee a fair comparison between the method we propose, BLESS, to the other methods, BSGLMM and Firth regression. We therefore use a data generating mechanism which simulates homogeneous regions of lesions proposed by Ge et al. (2014), with intensities that vary over subjects, which allows us to provide a fair comparison among the three methods evaluated. For our study, we consider $P = 2$ effects in addition to an intercept, which we label sex and group (e.g. patient and control). We simulate 2-D binary lesion masks of size 50×50, thus $M = 2,500$, with homogeneous effects in each 25×25 quadrant. The effect of sex leads to 4 times more lesions on the right side of an image for female subjects compared to the baseline. The second effect of group membership introduces an effect of 4 times more lesions within the lower left quadrant of an image for subjects within group 2. A Poisson random variable with base rate λ determines the number of lesions. The true parameter values for the spatially varying coefficients $\boldsymbol{\beta}(\boldsymbol{s})$ are not set but rather inferred by drawing a large sample through the data generating mechanism described below:

1) Draw the number of lesions n_ℓ for quadrant, $\ell \in \{I, II, III, IV\}$ starting with quadrant I in the top left corner and going clockwise where a quadrant contains $M/4$ voxels, within an image from a Poisson random variable $n_\ell \sim Poisson(\lambda c)$ where λ dictates the base rate number of lesion presence and c is the effect setting.

$$c = \begin{cases} (1,1,1,1), & \text{if } x_{gender} = 0 \text{ (male) and } x_{group} = 1, \\ (1,1,1,4), & \text{if } x_{gender} = 0 \text{ (male) and } x_{group} = 2, \\ (1,4,4,1), & \text{if } x_{gender} = 1 \text{ (female) and } x_{group} = 1, \\ (1,4,4,4), & \text{if } x_{gender} = 1 \text{ (female) and } x_{group} = 2. \end{cases}$$

2) Draw n_ℓ lesion locations uniformly within each quadrant.

3) Draw the lesion sizes from a discrete random variable, taking on lesion size $\{1, 9, 25\}$ with equal probabilities.

4) Create binary lesion masks $\boldsymbol{y}_i$ by combining the lesion locations from the 4 quadrants sampled above for every subject $i = 1, \dots, N$.

Note that lesions are allowed to intersect with each other and merge into bigger lesion formations. In subsequent evaluations we exclude the outer two edge voxels for every quadrant in order to reduce edge effects. Figure 2.3 shows an example of the generated binary lesion masks for randomly selected subjects and the empirical lesion probabilities for a subset of $N = 250$ subjects for every configuration of sex and group membership at $\lambda = 3$. The overall maximum lesion intensity lies at approximately 0.25.

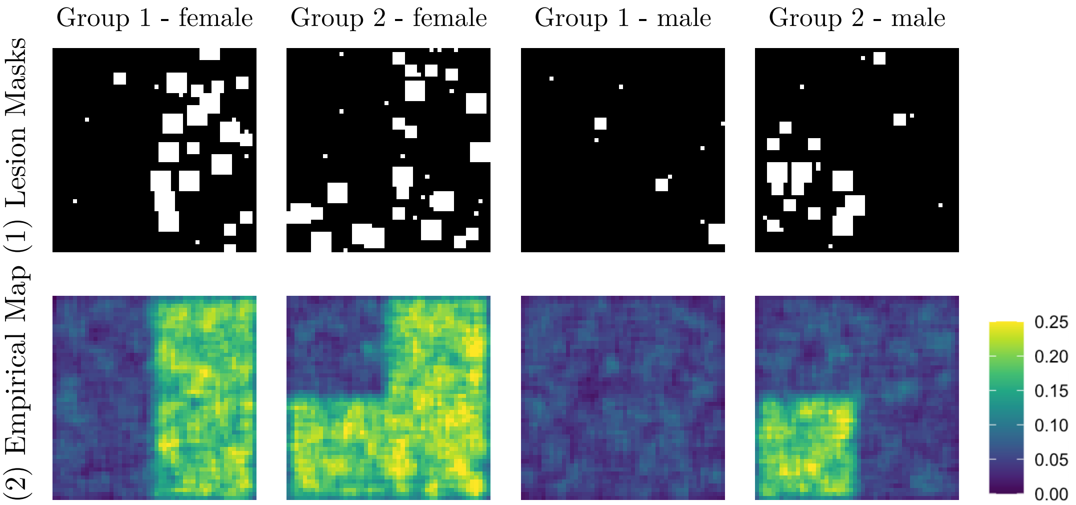


Figure 2.3: (1) Binary lesion masks indicating lesion presence $y_i(s_j) = 1$ and lesion absence $y_i(s_j) = 0$ for every subject $i = 1, \dots, N$ and voxel location $j = 1, \dots, M$ for the 4 different configurations of covariates of sex and group membership. (2) Aggregated empirical lesion maps for every combinations of male vs. female and group 1 vs. group 2.

We compare the point estimates of the parameter estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ to the truth where we determine truth by averaging the (homogeneous) rate within each quadrant for each of the four group-sex subject types, for a very large sample size of 100,000 observations. Under the assumption of homogeneity, we can acquire the true coefficient values by solving the following equation for each of the four group-sex combinations

$$\Phi^{-1}\left(p_i(s_j)\right) = \beta_0(s_j) + \beta_1(s_j)x_{i,1} + \beta_2(s_j)x_{i,2}, \tag{2.17}$$

which leads to the ground truth parameter estimates $\beta_0 = -1.7933$, $\beta_1 = 0.6559$ and $\beta_2 = 0.6237$ for a base rate intensity $\lambda = 3$, for example. Note that under the homogeneity assumption we can acquire a single true value by averaging across the quadrants of effect for the covariate sex (quadrant II and III or the covariate group (quadrant IV) or the intercept (all quadrants) for reporting purposes. However, for the evaluation of results we use the acquired true values for each spatial location $j = 1, \dots, M$ (see the definition of evaluation criteria in Section 2.3.1.2). The other averaged true parameter values for the intercept as well as the effects sex and group are reported in Table 2.1.

	β_0	β_1	β_2
$\lambda = 1$	-2.0868	0.5882	0.5586
$\lambda = 2$	-1.7933	0.6559	0.6237
$\lambda = 3$	-1.6089	0.7051	0.6685

Table 2.1: True parameter estimates averaged across quadrants for different base rate intensities $\lambda = 1, 2, 3$, which represent different magnitudes of lesion numbers and therefore small, medium, and large regression coefficients respectively. β_0 is averaged across all four quadrants, β_1 is averaged across its area of effect which is across quadrant II and III, and lastly β_2 is averged across its area of effect which is across quadrant IV.

2.3.1.2 Results Interpretation

Our simulation study evaluates the performance of BLESS for a broad set of scenarios with varying sample sizes $N = \{500;\ 1{,}000;\ 5{,}000\}$, base rate intensities $\lambda = \{1, 2, 3\}$ and sizes of spatial effect, where 25% (group effect) or 50% (sex effect) of the image are active, compared to BSGLMM and Firth regression. The true and estimated parameter estimates are available in the appendix alongside more sensitivity analyses from simulation studies with different spatial priors, varying magnitudes of the slab variance, and various neighbourhood structures in Section A.4, A.5 and A.6 respectively. We focus on the effect map for the covariate sex and generate 100 datasets for each sample size and base rate scenario to provide robustness by averaging over the results of each dataset. Parameter estimate results are then evaluated through absolute bias $n_{\text{sim}}^{-1} M^{-1} \sum_{k=1}^{n_{\text{sim}}} \sum_{j=1}^{M} |\hat{\beta}_k(s_j) - \beta(s_j)|$, variance of each model's parameter estimate $n_{\text{sim}}^{-1} M^{-1} \sum_{k=1}^{n_{\text{sim}}} \sum_{j=1}^{M} \text{Var}(\hat{\beta}_k(s_j))$, and mean squared error (MSE) $n_{\text{sim}}^{-1} M^{-1} \sum_{k=1}^{n_{\text{sim}}} \sum_{j=1}^{M} (\hat{\beta}_k(s_j) - \beta(s_j))^2$ for all $j = 1, \dots, M$ and $k = 1, \dots, n_{\text{sim}}$. The predictive results are evaluated through comparing the true lesion rate to the predicted lesion rate with the evaluation metrics absolute bias, variance and MSE in each quadrant $\ell \in \{I, II, III, IV\}$ and across each simulated dataset $k = 1, \dots, n_{\text{sim}}$. The true lesion rate is acquired by drawing 100,000 subjects for each sex and group membership combination and averaging the lesion masks across those draws. The predicted lesion rate is acquired by calculating the predicted probability of a lesion occurring at each location $j = 1, \dots, M$, for each subject $i = 1, \dots, N$, and across each simulated dataset $k = 1, \dots, n_{\text{sim}}$, so that the predicted lesion probability for a single subject, voxel, and simulated dataset is acquired by $\Phi(\hat{\beta}_0(s_j) + \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}(s_j))$.

The quality of parameter estimates and prediction for BLESS, BSGLMM and Firth regression are evaluated via absolute bias, variance and mean squared error (MSE) in Table 2.2. BLESS exhibits low bias for the evaluation of the parameter estimates for the sex effect and moreover outperforms the mass-univariate approach when comparing the quality of the coefficients via MSE. For example, the MSE of the parameter estimates for a small sample size $N = 500$ and low base rate intensity $\lambda = 1$ is approximately 5 times larger for Firth regression with a value of 0.0563 compared to our method BLESS with a value of 0.0106. This showcases how BLESS adequately regularises negligible coefficients to zero while the larger effects are unaffected by shrinkage. The quality of the predictive performance is determined by comparing the true empirical lesion rates to the estimated lesion

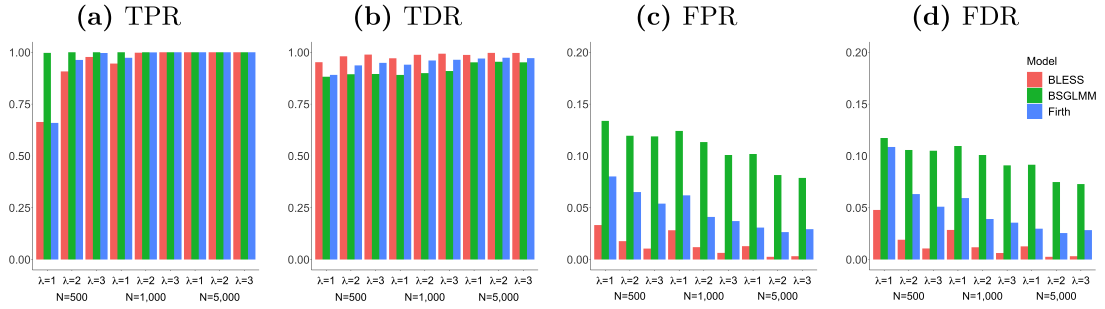
Parameter Estimate:	**Bias**			**Variance**			**MSE**		
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	**0.1022**	**0.0942**	**0.0849**	**0.0014**	**0.0019**	**0.0020**	**0.0106**	**0.0024**	**0.0020**
BSGLMM	0.4412	0.4388	0.4290	0.0117	0.0080	0.0067	0.0125	0.0082	0.0068
Firth	0.1670	0.1402	0.1344	0.0562	0.0348	0.0272	0.0563	0.0348	0.0272
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	**0.0662**	**0.0601**	**0.0581**	**0.0010**	**0.0010**	**0.0010**	**0.0010**	**0.0011**	**0.0010**
BSGLMM	0.3594	0.3404	0.3275	0.0063	0.0045	0.0039	0.0064	0.0046	0.0039
Firth	0.1290	0.0993	0.0912	0.0271	0.0171	0.0135	0.0271	0.0171	0.0135
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	**0.0401**	**0.0389**	**0.0333**	**0.0002**	**0.0002**	**0.0002**	**0.0002**	**0.0002**	**0.0002**
BSGLMM	0.2534	0.2493	0.2481	0.0018	0.0014	0.0012	0.0018	0.0014	0.0012
Firth	0.0645	0.0498	0.0409	0.0053	0.0034	0.0027	0.0053	0.0034	0.0027

Predictive Performance:	**Bias**			**Variance**			**MSE**		
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	0.0078	0.0052	0.0032	0.0011	0.0017	0.0018	0.0022	0.0031	0.0034
BSGLMM	**0.0027**	**0.0025**	**0.0020**	**0.0002**	**0.0004**	**0.0007**	**0.0002**	**0.0004**	**0.0007**
Firth	0.0170	0.0140	0.0122	0.0009	0.0016	0.0022	0.0018	0.0032	0.0043
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	0.0015	**0.0007**	**0.0013**	0.0004	0.0005	0.0007	0.0008	0.0010	0.0012
BSGLMM	**0.0010**	0.0018	0.0020	**0.0001**	**0.0002**	**0.0003**	**0.0001**	**0.0002**	**0.0003**
Firth	0.0082	0.0082	0.0056	0.0005	0.0008	0.0011	0.0009	0.0016	0.0021
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	0.0006	**0.0000**	0.0008	0.0001	**0.0001**	**0.0001**	0.0001	0.0002	0.0003
BSGLMM	**0.0004**	0.0008	**0.0001**	**0.0000**	**0.0001**	**0.0001**	**0.0000**	**0.0001**	**0.0001**
Firth	0.0010	0.0015	0.0020	0.0001	0.0002	0.0002	0.0002	0.0003	0.0004

Table 2.2: Evaluation of parameter estimates from the methods, BLESS, BSGLMM and Firth Regression via bias, variance and MSE of the spatially varying coefficients $\hat{\boldsymbol{\beta}}_1$, and the predictive performance. Improved bias and MSE for prediction for BLESS compared to Firth regression due to selective shrinkage property of BLESS.

probabilities. BLESS yields slightly better predictive results with respect to MSE, by exhibiting less biased estimates, compared to Firth regression for all scenarios except for the instance with low sample size and base rate intensity ($N = 500$, $\lambda = 1$) where Firth regression exhibits a slightly lower MSE. This result motivates the usage of BLESS for studies with larger sample sizes where BLESS outperforms the mass-univariate approach.

Our simulation study enforces 50% of the voxels as active on the right side of an image for the covariate sex. Hence, by knowing the true location of the effect, we can evaluate the quality of the inference results of BLESS compared to BSGLMM and Firth regression. Effect detection for BLESS is determined by utilising the latent variables $\hat{\boldsymbol{\gamma}}$, marking voxels s_j significant if $P(\hat{\gamma}_p(s_j) = 1|\boldsymbol{y}) > 0.5$. For BSGLMM and Firth regression we acquire test statistics $t = \hat{\beta}/\sigma_{\hat{\beta}}$ and threshold them at a significance level of 5%. We perform a multiple testing adjustment via a Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). All methods have



(a) TPR **(b)** TDR **(c)** FPR **(d)** FDR

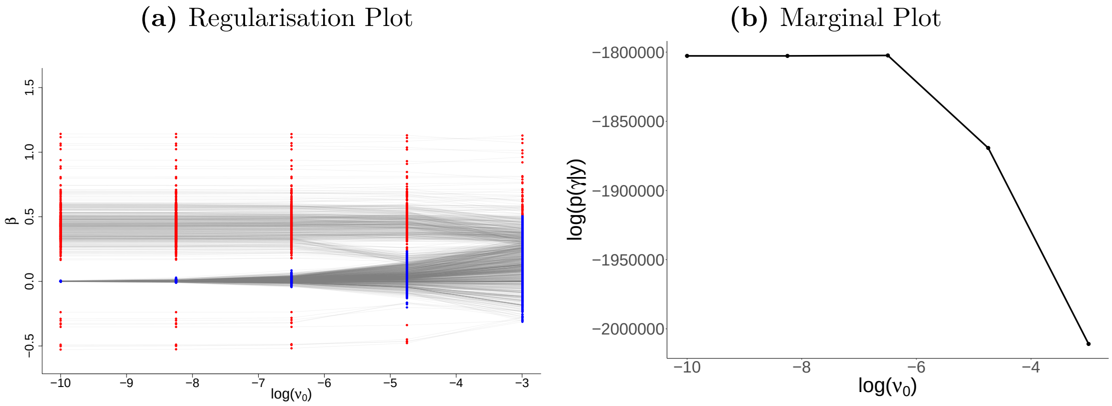
Figure 2.4: Evaluation of inference results from the methods, BLESS, BSGLMM and Firth Regression (FDR correction at 5%) via True Positive Rate (TPR), True Discovery Rate (TDR), False Positive Rate (FPR) and False Discovery Rate (FDR) for parameter estimate $\hat{\beta}_1$. BLESS outperforms Firth regression and BSGLMM with consistently high TPRs and low FPRs for various sample sizes and base rate intensities.

comparable results with respect to their performance in parameter estimation and prediction. However, the evaluation of the inference results in Figure 2.4 showcases that the Bayesian spatial model BSGLMM has a particularly high number of false positives and hence a very low level of specificity compared to the other methods. BLESS's key advantage is therefore shown by comparable levels of sensitivity and high values of specificity for all configurations of sample size and base rate intensity.

2.3.2 UK Biobank Application

2.3.2.1 Dataset Description and Model Estimation

Our motivating data set is from the UK Biobank, a large-scale biomedical database containing imaging data from predominately healthy individuals. With a target of 100,000 subjects, there is currently imaging data available for 40,000 participants (Miller et al., 2016). We refer the reader to Miller et al. (2016) for a detailed description of the scanning and processing protocols and Section 1.2.1 for a general overview of the UK Biobank. Our goal is to map the influence of risk factors on the incidence of white matter hyperintensities to understand their potential clinical significance and how they may contribute to neurological and cognitive deficits. Our data set consists of $N = 38{,}331$ subjects for which white matter hyperintensity binary lesion masks have been generated via the automatic lesion segmentation algorithm BIANCA (Griffanti et al., 2016). The binary lesion maps in subject space are then registered to a common 2mm MNI template across subjects. Each 3D binary image with voxel size $2 \times 2 \times 2$ mm^3 and dimensions $91 \times 109 \times 91$ contains a total of 902,629 voxel locations. Our region of interest lies in the white matter tracts of the brain and hence the total number of voxels is restricted to $M = 54{,}728$ by masking the 3D-lesion masks. We are interested in modelling the influence of age on lesion incidence while accounting for the confounding variables sex, head size scaling factor and the interaction of age and sex. In order to ensure interpretability across studies we have chosen confounds based on research by Alfaro-Almagro et al. (2021) where a head size scaling factor is commonly included to normalise brain tissue volumes for head size compared to the MNI template. The mean age of the participants in our study is 63.6 years (± 7.5 years) and 53.04% of individuals are female (20,332 women).



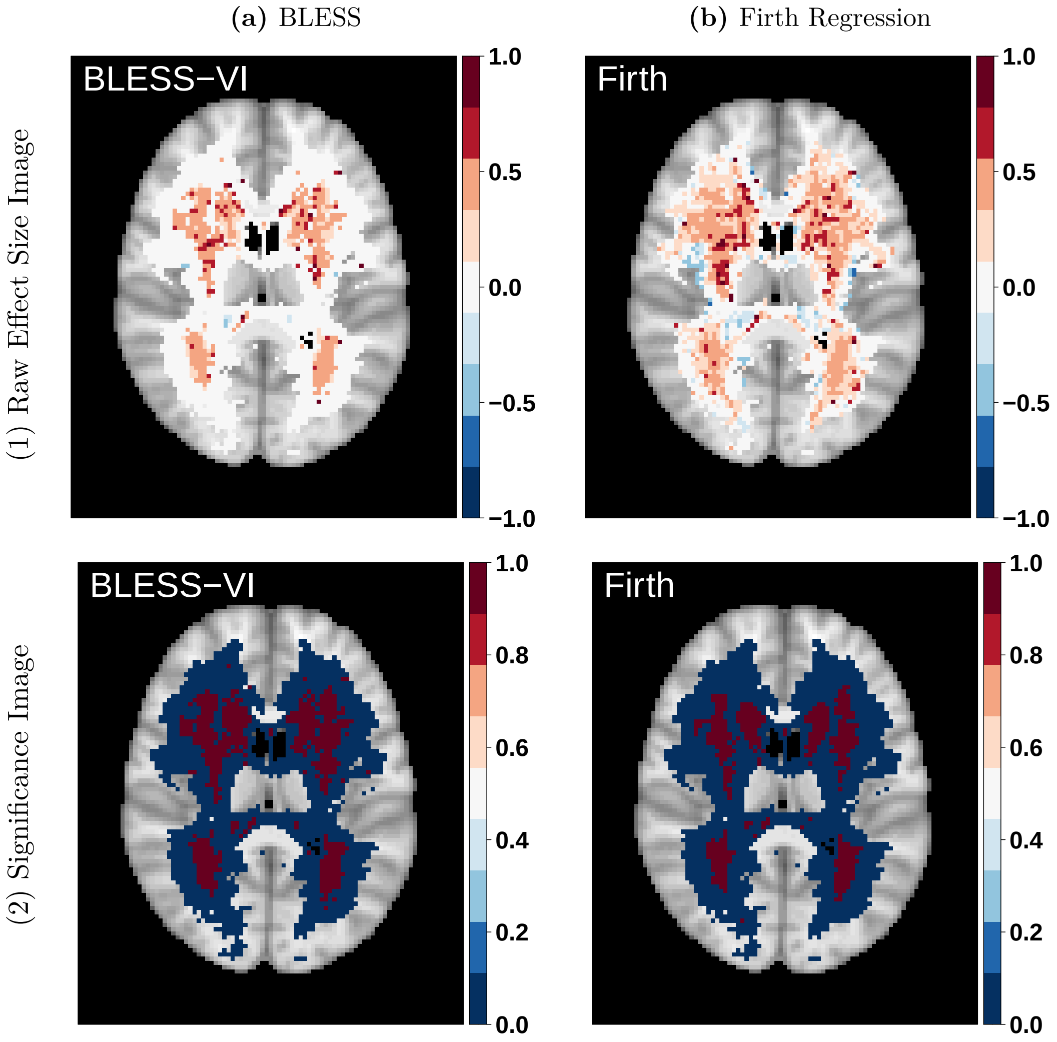
(a) Regularisation Plot

(b) Marginal Plot

Figure 2.5: (a) Regularisation plot for the age coefficients of an axial slice ($z = 45$, third dimension of the 3D image, plotting all $54,728$ coefficients would lead to severe overplotting) (active voxel: red, inactive voxel: blue) and (b) plot of marginal posterior of $\hat{\gamma}$ under $\nu_0 = 0$ over a sequence of equidistant $\nu_0 \in V$ within log-space ($\nu_0 = \{\exp(-10), \ldots, \exp(-3)\}$). Parameters stabilise across warm-start initialisations. Small effects are shrunk to 0 (blue) and large effects are almost unregeluarised (red).

For model estimation, we firstly perform backwards dynamic posterior exploration over $\nu_0 = \{\exp(-10), \ldots, \exp(-3)\}$ to help with the optimisation of the variational parameters; otherwise, we fit the model identically to the simulation study as described in the previous section. The regularisation and marginal plot for this application to the UK Biobank are displayed in Figure 2.5 and exhibit the same behaviour as in the simulation studies where likely negligible coefficients are shrunk to zero and relevant coefficients are left unaffected by regularisation, see the regularisation plot in Figure 2.5 (a).

2.3.2.2 Results Interpretation



(a) BLESS

(b) Firth Regression

(1) Raw Effect Size Image

(2) Significance Image

Figure 2.6: Comparison of results between (a) BLESS and (b) Firth Regression for a single axial slice ($z = 45$, third dimension of 3D image). (1) spatially varying age coefficient maps. (2) Thresholded age significance maps where the threshold for BLESS is determined via the probability of inclusion/exclusion $P(\gamma_p(s_j)|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) \geq 0.5$ and Firth regression via the test statistic $t = |\hat{\beta}/\hat{\sigma}_{\hat{\beta}}| \geq 1.96$ (significant voxels: red, not signficant voxels: blue, FDR-correction applied at 5%). The parameter map estimated via BLESS in (1a) exhibits a larger spatial area with values close to 0 compared to Firth in (1b).

Figure 2.6 compares the raw age effect size images of our method (a) BLESS, estimated via variational inference, to (b) the mass-univariate approach Firth

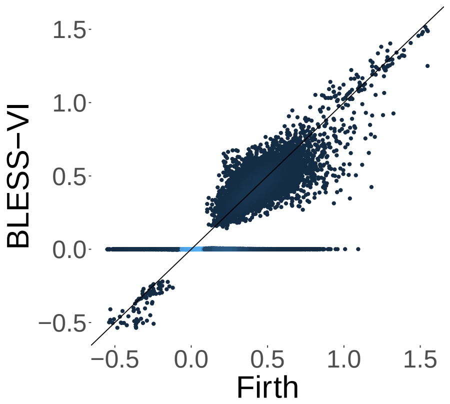


Figure 2.7: Scatterplot comparing the age coefficients for all voxel locations within the 3D image (lighter values indicate higher density of values). The scatterplot shows that BLESS regularises small effects almost completely to 0 compared to Firth.

regression. It should be noted that we omit the comparison to the other baseline method BSGLMM as the computation of the Bayesian spatial model becomes infeasible due to the large sample size of this study. We highlight how BLESS sufficiently regularises the negligible age coefficients to zero while leaving the larger effects unaffected. This is a direct consequence of the structured spike-and-slab prior placed on the spatially varying coefficients. Furthermore, the spatial MCAR prior allows the sparsity dictating parameters within the spike-and-slab prior to borrow strength from their respective neighbouring voxels. We further illustrate this behaviour by plotting the coefficients of the feature age of the entire 3D effect map of the brain in the scatterplot in Figure 2.7 where the plot shows the induced shrinkage of small effects to zero via BLESS while the Firth regression parameter estimates vary for the negligible effects and exhibit non-zero values.

For inference, we threshold the posterior probability of inclusion at 0.5 in order to acquire its respective binary significance map. Hence, we exploit variable selection as a means to conduct inference. On the other hand, the mass-univariate Firth regression ignores any form of spatial dependence and hence requires the application of a multiple testing correction where we adjust the p-values with a FDR correction (Benjamini and Hochberg, 1995) at a significance level of 5%. The results in Figure 2.6 indicate a slightly larger extent of spatial activation for BLESS compared to Firth regression for a sample size of $N = 2,000$. For the covariate age, in the regression model estimated via Firth regression 10,171 voxels are deemed active based on uncorrected p-values. On the other hand, only 6,278 voxels pass the FDR adjusted threshold whereas in BLESS 8,257 effects are detected via simply thresholding the posterior inclusion probabilities suggesting that similarly to our simulation study in Section 2.3.1 BLESS is more sensitive than Firth regression in the case of a small sample size and low lesion rate scenario.

2.4 Discussion and Future Work

We have proposed a novel Bayesian spatial generalised linear model with a structured spike-and-slab prior for the analysis of binary lesion data. Our main contribution is the development of a scalable version of a Bayesian spatial model that is able to diminish spurious associations by shrinking negligible coefficients to zero, to increase model interpretation via Bayesian variable selection and to provide a model that is also easily extendable to other neuroimaging modalities, such as functional MRI with a continuous response variable.

The computational tractability of our method is also facilitated by using a data augmentation approach for the probit model and an analytical approximation to estimate the parameters in the logistic function in the Bernoulli prior on the latent variables within the spike-and-slab distribution (Albert and Chib, 1993; Jaakkola and Jordan, 2000). For future work, switching the probit link to a logit link enables the interpretation of the spatially varying coefficients via log-odds ratios. Moreover, advances in Bayesian inference for efficient posterior estimation of logistic regressions using Pólya-Gamma latent variables developed by Polson et al. (2013) and Durante and Rigon (2019) connect the approach by Jaakkola and Jordan (2000) to the exact Pólya-Gamma data augmentation by defining the adaptive bound by Jaakkola and Jordan (2000) through a probabilistic interpretation. Therefore, enabling potential gains in accuracy and computational efficiency while providing an explanation for the excellent empirical performance of the framework by Jaakkola and Jordan (2000).

Lastly, we would like to address how variational inference underestimates the posterior variance of the spatially varying coefficients (Yao et al., 2018). While we show empirically that the point estimates of our parameters are accurate in simulation studies, neuroimaging applications often also require an accurate assessment of the uncertainty in the parameters in order to perform reliable inference. Hence, we extend our work in the next chapter to the development of an approximate posterior sampling algorithm for BLESS rather than using a variational approximation alone for optimisation. We hereby aim to capture the posterior variance more appropriately while remaining scalable to the large sample size of the UK Biobank.

3

Scalable Uncertainty Quantification and Cluster Size-Based Imaging Statistics for Large-scale Lesion Mapping Applications

Contents

The previous chapter introduced a method, called BLESS, for structured Bayesian variable selection by imposing a spike-and-slab prior on the spatially varying coefficients with a smoothing prior on the sparsity parameter in the inclusion probabilities for an image-on-scalar regression problem setting where the image outcome is binary at each spatial location. Moreover, the method ensured scalability to larger sample sizes by utilising variational inference for acquiring posterior estimates. For more details on the model setup and inference algorithm, we refer the reader to Section 2. While the method provided an increase in the scalability of parameter estimation and inference to sample sizes of the scale of the UK Biobank, it led to a severe under-estimation of posterior variance of the spatially varying coefficients which is common for variational approximations (Yao et al., 2018). In this chapter, we therefore aim to extend our previous work to account for the under-estimation of posterior variance by applying approximate posterior sampling, through Bayesian bootstrap methods and a class of jittered spike-and-slab priors, for parameter estimation and inference. Furthermore, we derive cluster size-based imaging statistics, a tool used in neuroimaging analysis to enhance inference, as a by-product of our scalable and more accurate uncertainty quantification of the spatially varying coefficients.

3.1 Introduction

3.1.1 Approximate Posterior Sampling

Approximate posterior sampling is able to capture marginal posterior densities more accurately than variational densities while remaining highly scalable due to embarrassingly parallel implementations (Fong et al., 2019). The cornerstone of these methods lies in the Bayesian bootstrap (BB) (Rubin, 1981) and the Weighted Likelihood Bootstrap (WLB) (Newton and Raftery, 1994). The WLB randomly re-weights the likelihood with Dirichlet weights for the observations and maximises this likelihood with respect to the parameter of interest. Modern extensions not only re-weight the likelihood but also introduce perturbations on the prior, such as in the Weighted Bayesian Bootstrap (WBB) formulated by Newton et al. (2021) or the Posterior Bootstrap proposed by Fong et al. (2019). More generally, Lyddon et al. (2018) and Fong et al. (2019) developed Bayesian nonparametric learning (BNL) routines using the WLB which utilise parametric models to achieve posterior sampling through the optimisation of randomised objective functions.

Our focus lies on the recently introduced method by Nie and Ročková (2022) which combines Bayesian bootstrap methods with a new class of jittered spike-and-slab LASSO (BB-SSL) priors and obtains samples via optimisation of many independently perturbed datasets by re-weighting the likelihood and by jittering the prior with a random mean shift. This procedure is equivalent to adding pseudo-samples from a prior sampling distribution as in the case of BNL (Fong et al., 2019). We argue that for high-dimensional datasets with large sample sizes, where memory allocation is already a computational concern, the approach by Nie and Ročková (2022) is favourable as it merely requires storing a set of mean shift parameters compared to an arbitrarily large number of pseudo-samples (Fong et al., 2019). In the following sections, we will introduce the concepts discussed above in more detail and showcase their connections with each other.

3.1.1.1 Weighted Likelihood Bootstrap

The Weighted Likelihood Bootstrap generalises the Bayesian bootstrap, developed by Rubin (1981) for nonparameteric models, to parameteric and semiparametric models and can be used to approximate posterior distributions (Newton and Raftery, 1994). Generally, the WLB approximates a posterior by acquiring a random sample of parameter estimates from a distribution of the parameter space through maximising a weighted likelihood function. The weight vector, by which each likelihood term for each data point is re-weighted, is determined by a draw from some probability distribution where in the most generic case a weight vector is drawn from a uniform Dirichlet distribution where the concentration parameter $\alpha = 1$.

For illustrative purposes, we will use a simple Bayesian linear regression setup to highlight the differences between the WLB, WBB, posterior bootstrap, and BB-SSL where the response vector $\boldsymbol{y} = (y_1, \dots, y_N)^T$ is linked with a Gaussian distribution to the predictors $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)^T \in \mathbb{R}^{P \times N}$ so that $y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i$ with $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ for each observation $i = 1, \dots, N$ where the residual noise variance is $\sigma^2 > 0$ and the parameter vector $\boldsymbol{\beta} \in \mathbb{R}^P$ is possibly sparse. We assume σ^2 is known throughout this introduction and we also center the output $\boldsymbol{y}$ and input $\boldsymbol{X}$ in order to omit an intercept term. The likelihood function L is therefore defined by $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^N \phi(y_i; \boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2)$, where $\phi(\cdot)$ is a Gaussian density, for all data points $i = 1, \dots, N$ and some prior distribution $\pi(\boldsymbol{\beta})$.

Algorithm 3: Weighted Likelihood Bootstrap

```
Result: Draw approximate samples $\tilde{\boldsymbol{\beta}}$ by independently maximising
        randomly re-weighted likelihood functions $\tilde{L}$.
Input: Samples $(\boldsymbol{X}, \boldsymbol{y})$.
Set: $\alpha$: concentration parameter; $B$: number of bootstrap samples
for $b = 1, \ldots, B$ do
    1) Sample weights $\boldsymbol{w}^{(b)} = (w_1^{(b)}, \ldots, w_N^{(b)}) \sim \text{Dirichlet}(\alpha, \ldots, \alpha)$.
    2) Maximise the re-weighted likelihood, so that
      $\tilde{\boldsymbol{\beta}}^{(b)} = \arg\max_{\boldsymbol{\beta}} \tilde{L}^{\boldsymbol{w}^{(b)}}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{X}; \boldsymbol{y})$ where
      $\tilde{L}^{\boldsymbol{w}^{(b)}}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{X}; \boldsymbol{y}) = \prod_{i=1}^{N} \phi(y_i; \boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2)^{w_i^{(b)}}$.
```

As shown in Algorithm 3, the WLB produces a random sample of the parameters $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}^{(1)}, \ldots, \tilde{\boldsymbol{\beta}}^{(B)})$ by maximising the weighted likelihood function $\tilde{L}$ for each bootstrap iteration $b = 1, \ldots, B$ where for each iteration a new weight vector $\boldsymbol{w}^{(b)}$ is drawn from a Dirichlet distribution with a concentration parameter α to re-weight the likelihood function L with. Note that other distributions can be considered for re-weighting the likelihood function; however, for the Dirichlet distribution with $\alpha = 1$ Newton and Raftery (1994) show that the WLB is first order correct and hence possesses consistency in the estimation of mean and variance of the posterior distribution. Overall, the WLB is very easy to implement, as it only requires a maximum likelihood estimator, and it is also computationally fast due to its embarrassingly parallelisable nature. However, the WLB is not able to incorporate any prior information and therefore is unlikely to possess good higher order approximation properties. Nevertheless, using sampling importance re-sampling or density estimation after applying the WLB can alleviate some of these inaccuracies (Newton and Raftery, 1994).

3.1.1.2 Weighted Bayesian Bootstrap

The Weighted Bayesian Bootstrap accounts for prior information when drawing approximate posterior samples by not only perturbing the likelihood with random weights but also the prior terms (Newton et al., 2021). It should be noted that this idea has also been introduced by Lyddon (2018) in the form of the so-called weighted h-likelihood bootstrap for mixed effects models. Equivalently to the WLB, the WBB is computationally fast and scalable to large datasets as it replaces sequential posterior sampling with repeated optimisation which can be performed in an embarrassingly parallel manner. Algorithm 4 describes the process of acquiring approximate samples $\tilde{\boldsymbol{\beta}}^{(b)}$ from the posterior by optimising pseudo-posteriors created by re-weighting the likelihood function and also the prior with random weights $\boldsymbol{w}^{(b)}$ and $\tilde{\boldsymbol{w}}^{(b)}$ respectively.

```
Algorithm 4: Weighted Bayesian Bootstrap
Result: Draw approximate samples $\tilde{\boldsymbol{\beta}}$ from posterior distribution by
        independently maximising randomly re-weighted likelihood
        functions $\tilde{L}$ and prior terms $\tilde{\pi}(\boldsymbol{\beta})$.
Input: Samples ($\boldsymbol{X}$, $\boldsymbol{y}$).
Set: $B$: number of bootstrap samples
for $b$ = 1, ..., $B$ do
    1) Sample weights $w_i^{(b)} \overset{iid}{\sim} Exp(1)$ for each data point $i = 1, \dots, N$ where
       $(w_1^{(b)}, \dots, w_N^{(b)}) \overset{iid}{\sim} Exp(1)$ is equivalent to
       $\left(\frac{w_1}{\sum_i w_i}, \dots, \frac{w_N}{\sum_i w_i}\right) \sim \text{Dirichlet}(1, \dots, 1)$ for the likelihood terms and
       weights $\tilde{\boldsymbol{w}}^{(b)} \overset{iid}{\sim} Exp(1)$ (if weights are not fixed at 1) for the prior term.
    2) Maximise the re-weighted likelihood and prior term, so that
       $\tilde{\boldsymbol{\beta}}^{(b)} = \arg\max_{\boldsymbol{\beta}} \tilde{L}^{\boldsymbol{w}^{(b)}}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{X}; \boldsymbol{y})\pi(\boldsymbol{\beta})^{\tilde{\boldsymbol{w}}^{(b)}}$ where
       $\tilde{L}^{\boldsymbol{w}^{(b)}}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{X}; \boldsymbol{y}) = \prod_{i=1}^{N} \phi(y_i; \boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma^2)^{w_i^{(b)}}$.
```

The WBB achieves first-order correctness (consistency and asymptotic normality) for the low-dimensional Bayesian LASSO regression for a wide class of random weight distributions for the likelihood and prior terms alike (Newton et al., 2021). However, we are particularly interested in the behaviour of the WBB when it comes to thresholding procedures. Nie and Ročková (2022) provide a discussion on the drawbacks of the WBB when it comes to high-dimensional regression problems where sparsity is induced with a spike-and-slab LASSO prior on the coefficients $\boldsymbol{\beta}$. Specifically, they show that accurate uncertainty quantification for negligible coefficients is not guaranteed as many of the posterior samples may be exactly 0 when the pseudo-posterior is optimised in the WBB which hence leads to a severe under-estimation of posterior variance on the negligible coefficients.

3.1.1.3 Posterior Bootstrap

The posterior bootstrap was developed by Lyddon et al. (2018) and Fong et al. (2019) as a scalable posterior sampling technique in lieu of computationally intensive MCMC methods through their advancements on Bayesian nonparametric learning. BNL uses statistical models without assuming that the underlying model is true and generates exact posterior samples by formulating a Bayesian nonparametric model by placing a single (Fong et al., 2019) or a mixture of (Lyddon et al., 2018) Dirichlet process prior centred on a parametric model on the unknown data distribution and returning a nonparametric posterior over the parameter of interest. We will focus on the simpler prior choice with a single Dirichlet process prior which retains many of the theoretical properties of the original mixture formulation; however, increases the computational performance significantly as sampling from a Bayesian posterior is no longer required. Performing posterior inference through BNL is advantageous as it accounts for model misspecification and is embarrassingly parallel (Fong et al., 2019).

More generally, the observed data is assumed to be $\boldsymbol{y}_{1:N} \overset{iid}{\sim} F_0$, where $\boldsymbol{y}_{1:N}$ is a sequence of i.i.d. observables and F_0 is the unknown sampling distribution, and some parameter is expressed by $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^P$, which is indexed by a family of probability densities $\mathcal{F}_{\boldsymbol{\Theta}} = \{f_{\boldsymbol{\theta}}(\boldsymbol{y}); \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. In traditional Bayesian updating the unknown sampling distribution is assumed to be included in the model $F_{\boldsymbol{\Theta}}$ which is an assumption that is not necessary for BNL. Moreover, the parameter of interest is defined as $\boldsymbol{\theta}_0(F_0) = \arg\min_{\boldsymbol{\theta}} \int l(\boldsymbol{y}, \boldsymbol{\theta}) dF_0(\boldsymbol{y})$, where $l(\boldsymbol{y}, \boldsymbol{\theta})$ is a loss function that is utilised to acquire a statistic of interest. While we are mostly interested in specifying the loss function as $l(\boldsymbol{y}, \boldsymbol{\theta}) = -\log f_{\boldsymbol{\theta}}(\boldsymbol{y})$, where $f_{\boldsymbol{\theta}}$ is a density of a parametric model, other loss functions return other statistics of interest, such as $l(\boldsymbol{y}, \boldsymbol{\theta}) = |\boldsymbol{y} - \boldsymbol{\theta}|$ results the median or $l(\boldsymbol{y}, \boldsymbol{\theta}) = (\boldsymbol{y} - \boldsymbol{\theta})^2$ returns the mean (Fong et al., 2019).

After choosing the loss function, a Dirichlet process prior is placed on the unknown sampling distribution F_0, so that $[F|\alpha, F_\pi] \sim \mathrm{DP}(\alpha, F_\pi)$, where α is a concentration parameter measuring the strength of our prior belief and F_π is a prior base measure. The base measure F_π is responsible for capturing the prior information about the unknown sampling distribution. Some common examples of prior choice for the base measure include $f_\pi = \int f_{\boldsymbol{\theta}}(\boldsymbol{y}) d\pi(\boldsymbol{\theta})$ for the density of F_π, where $\pi(\boldsymbol{\theta})$ is the prior distribution on $\boldsymbol{\theta}$, and the empirical distribution of historical data $\hat{\boldsymbol{y}}_{1:\hat{N}}$ defined by $F_\pi(\boldsymbol{y}) = \frac{1}{\hat{N}} \sum_{i=1}^{\hat{N}} \delta_{\hat{y}_i}(\boldsymbol{y})$ where $\delta(\cdot)$ is the Dirac function. The concentration parameter α on the other hand defines the strength of our prior belief in the base measure F_π and can be seen as the effective sample size from the base measure F_π. Fong et al. (2019) propose two options for choosing the concentration parameter in a principled manner through 1) simulation of the prior distribution of $\boldsymbol{\theta}$ and tuning its variance or 2) through the a priori variance of the mean functional. BNL includes various special cases introduced in the prior sections, such as the Bayesian Bootstrap if $\alpha = 0$ and the loss function is not defined by a parametric model, the WLB if $\alpha = 0$ and $l(y, \theta) = -\log f_{\boldsymbol{\theta}}(\boldsymbol{y})$, and the WBB if $\alpha = 0$ and $l(\boldsymbol{y}, \boldsymbol{\theta}) = -\log f_{\boldsymbol{\theta}}(\boldsymbol{y}) + \lambda\pi(\boldsymbol{\theta})$.

The BNL posterior on F is conjugate and hence defined by a Dirichlet process $[F|\boldsymbol{y}_{1:N}] \sim \mathrm{DP}(\alpha + N, G_N)$, where $G_N = \frac{\alpha}{\alpha+N} F_\pi + \frac{1}{\alpha+N} \sum_{i=1}^{N} \delta_{y_i}$ is the weighted sum of the prior base measure F_π and the empirical distribution $F_n = \frac{1}{N} \sum_{i=1}^{N} \delta_{y_i}$. The posterior on $\boldsymbol{\theta}$ is determined by $\pi(\boldsymbol{\theta}|\boldsymbol{y}_{1:N}) = \int \pi(\boldsymbol{\theta}|F) d\pi(F|\boldsymbol{y}_{1:N})$. Draws of the posterior distribution on $[F|\boldsymbol{y}_{1:N}]$ are almost surely discrete and therefore $\boldsymbol{\theta}_0(F_0) = \arg\min_{\boldsymbol{\theta}} \int l(\boldsymbol{y}, \boldsymbol{\theta}) dF_0(\boldsymbol{y})$ simplifies to $\theta(F) = \arg\min_{\boldsymbol{\theta}} \sum_{t=1}^{\infty} w_t l(\tilde{\boldsymbol{y}}, \boldsymbol{\theta})$ where the stick-breaking construction defines $\boldsymbol{w}_{1:\infty} \sim \mathrm{GEM}(\alpha + N)$ and $\tilde{\boldsymbol{y}}_{1:\infty} \overset{iid}{\sim} G_N$. It is apparent that computing the posterior $\pi(\boldsymbol{\theta}|\boldsymbol{y}_{1:N})$ is intractable as sampling from the Dirichlet process posterior $[F|\boldsymbol{y}_{1:N}]$ requires infinite computational resources. Nevertheless, the so-called posterior bootstrap algorithm, described in Algorithm 5, proposes the usage of approximate weights $\tilde{\boldsymbol{w}}_{1:T} \sim \mathrm{Dirichlet}(\frac{\alpha}{T}, \ldots, \frac{\alpha}{T})$ with a finite truncation limit T, to acquire posterior estimates for our parameter of interest $\boldsymbol{\theta}$.

Algorithm 5: Posterior Bootstrap

```
Result: Draw approximate samples $\tilde{\boldsymbol{\beta}}$ from posterior distribution through
        BNL by optimising randomised objective functions.
Input: Samples $(\boldsymbol{X}, \boldsymbol{y})$.
Set: $\alpha$: concentration parameter; $F_\pi$: prior base measure; $T$: truncation
 limit ; $B$: number of bootstrap samples
for $b = 1, \ldots, B$ do
    1) Draw prior pseudo-samples $(\tilde{\boldsymbol{x}}^{(b)}_{1:T}, \tilde{\boldsymbol{y}}^{(b)}_{1:T}) \sim F_\pi$.
    2) Draw random weights $(\boldsymbol{w}^{(b)}_{1:N}, \tilde{\boldsymbol{w}}^{(b)}_{1:T}) \sim \text{Dirichlet}(1, \ldots, 1, \frac{\alpha}{T}, \ldots, \frac{\alpha}{T})$.
    3) Optimise randomised objective function, so that
       $\tilde{\boldsymbol{\beta}}^{(b)} = \arg\max_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} w_i^{(b)} l(\boldsymbol{x}_i, y_i, \boldsymbol{\beta}, \sigma^2) + \sum_{t=1}^{T} \tilde{w}_t^{(b)} l(\tilde{\boldsymbol{x}}_t^{(b)}, \tilde{y}_t^{(b)}, \boldsymbol{\beta}, \sigma^2) \right\}$,
       where the loss function in this example is characterised by the
       log-likelihood function $l(\boldsymbol{x}_i, y_i, \boldsymbol{\beta}, \sigma^2) = \ln\{\phi(y_i; \boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2)\}$.
```

In summary, the overview of BNL, specifically the posterior bootstrap, showcase how one can draw posterior samples through placing a prior on the sampling distribution F and then optimising a randomised objective function consisting of a combination of randomly weighted loss functions evaluated at the observed samples $\boldsymbol{y}_{1:N}$ and the pseudo-samples $\tilde{\boldsymbol{y}}_{1:T}$ which are drawn from the prior base measure F_π (Fong et al., 2019; Nie and Ročková, 2022).

3.1.1.4 Bayesian Bootstrap Spike-and-Slab LASSO

In this section, we will extend our canonical example by placing a spike-and-slab-type prior distribution on the regression coefficients $\boldsymbol{\beta}$, so that $\pi(\boldsymbol{\beta}|\boldsymbol{\gamma}) = \prod_{j=1}^{P} [\gamma_j \psi_0(\beta_j) + (1 - \gamma_j)\psi_1(\beta_j)]$, where $\boldsymbol{\gamma} \in \mathbb{R}^P$ is a vector of indicator variables, where the probability of prior inclusion defined by $\mathbb{P}(\gamma_j = 1|\theta) = \theta$, $\psi_0(\beta)$ is the spike distribution, and $\psi_1(\beta)$ is the slab distribution. Moreover, the work by Nie and Ročková (2022) focuses on providing accurate but also scalable uncertainty quantification of spike-and-slab LASSO (SS-LASSO) priors which replace the traditional mixture of normal distributions with a mixture of Laplace distributions, where the spike distribution is represented by $\psi_0(\beta) = \frac{\lambda_0}{2} \exp\{-|\beta|\lambda_0\}$ and the slab distribution by $\psi_1(\beta) = \frac{\lambda_1}{2} \exp\{-|\beta|\lambda_2\}$ with λ_0 and λ_1 as hyperparameters (Ročková and George, 2018). Models with SS-LASSO priors can be estimated via fast nonconvex penalised likelihood methods and in our simple Gaussian regression example the posterior mode can be found by the following expression: $\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \left\{ \prod_{i=1}^{N} \phi(y_i; \boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2) \times \int \prod_{j=1}^{P} \pi(\beta_j|\theta) d\pi(\theta) \right\}$, where $\pi(\beta_j|\theta) = \theta\psi_1(\beta_j) + (1-\theta)\psi_0(\beta_j)$ is obtained by integrating out the indicator variables γ_j.

The downside of this approach is the lack of proper uncertainty quantification which inspired the development of the Bayesian Bootstrap Spike-and-slab LASSO (BB-SSL) method by Nie and Ročková (2022). They propose another extension of the WLB by not only re-weighting the likelihood with Dirichlet weights but also by re-centering the prior mean of the SS-LASSO priors with a random shift. Similar to the methods described above, repeated optimisation, which can be performed in parallel, is able to provide an approximate sample of the posterior, see Algorithm 6. Moreover, Nie and Ročková (2022) show that for some induced perturbations the approximate posterior contracts around the truth at the same rate as the actual posterior which is a desirable property which neither the WLB or the WBB formally proved.

```
Algorithm 6: BB-SSL
  Result: Draw approximate samples $\tilde{\boldsymbol{\beta}}$ from posterior distribution by
          maximising pseudo-posterior obtained by re-weighting the
          likelihood and perturbing the prior.
  Input: Samples ($\boldsymbol{X}$, $\boldsymbol{y}$).
  Set: $\lambda_0 \gg \lambda_1$: spike and slab variances; $\alpha$: concentration parameter; $B$:
    number of bootstrap samples
  for $b = 1, \ldots, B$ do
      1) Draw random weights $\boldsymbol{w}_{1:N}^{(b)} \sim \text{Dirichlet}(\alpha, \ldots, \alpha)$.
      2) Draw random jitter / prior shift of the mean $\boldsymbol{\mu}^{(b)}$ with $\mu_j^{(b)} \overset{iid}{\sim} \psi_0(\mu)$
        for all $j = 1, \ldots, P$ from the spike distribution.
      3) Optimise pseudo-posterior through a pseudo-MAP estimator, so that
        $\tilde{\boldsymbol{\beta}}^{(b)} = \arg\max_{\boldsymbol{\beta}} \left\{ \tilde{L}^{\boldsymbol{w}^{(b)}}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{X}; \boldsymbol{y}) \times \int_\theta \pi(\boldsymbol{\beta}|\boldsymbol{\mu}^{(b)}, \theta) d\pi(\theta) \right\}$.
```

BB-SSL is therefore able to incorporate a random perturbation in the likelihood as well as the prior; however, in a more principled manner than the WBB which inflates the prior contribution with a fixed or random prior weight drawn from a Dirichlet distribution, see Section 3.1.1.2. It is also interesting to highlight the similarities of the BB-SSL procedure to the other methods, such as the WLB which can be achieved by not jittering the SS-LASSO distribution with a mean shift vector $\boldsymbol{\mu}^{(b)}$ at every bootstrap iteration but rather keeping the original zero centred formulation. More interestingly, Nie and Ročková (2022) show that the BB-SSL procedure, which adds a prior perturbation to the SS-LASSO prior, is equivalent to the posterior bootstrap, which adds prior pseudo-samples $(\tilde{\boldsymbol{X}}_{1:T}, \tilde{\boldsymbol{y}}_{1:T})$ from a prior base measure F_π to the optimisation of the pseudo-posterior, so that $\tilde{\boldsymbol{x}}_t \sim F_N(\boldsymbol{x}) = \frac{1}{N}\delta(\boldsymbol{x}_i)$ and $\tilde{y}_t|\tilde{\boldsymbol{x}}_t = \hat{y}_t + \tilde{\boldsymbol{x}}_t^T \boldsymbol{\mu}$ where μ is drawn from the spike distribution and $\hat{y}_t = y_i$ where i satisfies $\tilde{\boldsymbol{x}}_t = \boldsymbol{x}_i$. Hence, approximate posterior samples from the posterior bootstrap can be drawn by optimising for

$$\tilde{\boldsymbol{\beta}}^{(b)} \overset{d}{\approx} \arg\max_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} w_i^{*(b)} \log[\phi(y_i; \boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2)] + \log \left[\int \prod_{j=1}^{P} \pi(\beta_j - \frac{\alpha}{\alpha+N} \mu_j^{*(b)} | \theta) d\pi(\theta) \right] \right\} \\ - \frac{\alpha}{\alpha+N} \boldsymbol{\mu}^{*(b)},$$

where $(w_1^*, \dots, w_N^*)^T \sim N \times \text{Dirichlet}(1 + \frac{\alpha}{N}, \dots, 1 + \frac{\alpha}{N})$ are weights and $\boldsymbol{\mu}^*$ is drawn from the spike distribution, to resemble the formulation of the BB-SSL. A draw from the BB-SSL however is acquired by optimising a pseudo-posterior as an objective, where the likelihood has been re-weighted with Dirichlet weights $(w_1, \dots, w_N)^T \sim \text{Dirichlet}(\alpha, \dots, \alpha)$ and the prior perturbed with $(\mu_1, \dots, \mu_P)^T \sim \psi_0(\mu)$, so that

$$\tilde{\boldsymbol{\beta}}^{(b)} = \arg\max_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} w_i^{(b)} \log \left[\phi(y_i; \boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2) \right] + \log \left[\int \prod_{j=1}^{P} \pi(\beta_j - \mu_j^{(b)} | \theta) d\pi(\theta) \right] \right\}.$$

The above equations provide similar draws $\tilde{\boldsymbol{\beta}}^{(b)}$ from the posterior even though the posterior bootstrap and BB-SSL have different objective functions when it comes to optimisation. Hence, this is showing that enlarging the sample through pseudo-samples in the posterior bootstrap and perturbing the prior mean of the SS-LASSO priors in the BB-SSL leads to a similar outcome. The only differences between the techniques are the definition of the concentration parameter α as well as adding the random shift $\frac{\alpha}{\alpha+N}\mu_j^*$ back to the objective function in the posterior bootstrap formulation which leads to less variance in the posterior estimation (Nie and Ročková, 2022).

3.1.2 Cluster Size-Based Imaging Statistics

Another advantage of using bootstrapping techniques is that we can acquire posteriors on complex imaging statistics, such as cluster size. In neuroimaging, a cluster is a contiguous set of voxels with voxelwise statistics that exceed some cluster-defining threshold (Poline and Mazoyer, 1993). Cluster-wise inference which declares clusters significant when they exceed a given size, has been found to be more sensitive to detect effects than voxelwise inference (Friston et al., 1996), albeit having less spatial specificity as the null hypothesis is specified at the level of clusters. The existing methods can only produce p-values, while we are able to compute credible intervals for cluster size or any other spatial feature of interest by incorporating bootstrapping into our model.

3.2 Methods

In this section, we extend the methodology of BLESS, developed in Section 2 for large-scale lesion mapping studies, to account for uncertainty quantification via Bayesian bootstrap ideas and a class of jittered spike-and-slab priors. Within this chapter, we now refer to the original formulation of BLESS where posterior inference is exclusively performed via variational inference (VI), specifically following the dynamic posterior exploration approach described in Section 2.2.3, as BLESS-VI and call our extension BB-BLESS which stands for Bayesian Bootstrap - BLESS.

3.2.1 Model

$$
\text{(1) Probit model} \begin{cases} [y_i(s_j) | p_i(s_j)] \sim \text{Bernoulli}(p_i(s_j)) \\ \Phi^{-1}\left\{\mathbb{E}[y_i(s_j) | p_i(s_j)]\right\} = \eta_i(s_j) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}(s_j) + \beta_0(s_j) \\ \Phi^{-1}\left\{\Pr[y_i(s_j) = 1 | \eta_i(s_j)]\right\} = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}(s_j) + \beta_0(s_j) \end{cases}
$$

$$
\text{(2) Latent model} \begin{cases} \Pr[y_i(s_j) | z_i(s_j)] = \begin{cases} 1, & z_i(s_j) > 0, \\ 0, & z_i(s_j) \leq 0, \end{cases} \\ z_i(s_j) \sim \mathcal{N}(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}(s_j) + \beta_0(s_j),\ 1) \end{cases}
$$

$$
\text{(3) Spike-and-slab prior} \begin{cases} \beta_p(s_j) \sim \mathcal{N}(0,\ \nu_0(1 - \gamma_p(s_j)) + \nu_1 \gamma_p(s_j)) \\ \gamma_p(s_j) \sim \text{Bernoulli}(\sigma(\theta_p(s_j))) \end{cases}
$$

$$
\text{(4) MCAR prior} \begin{cases} [\boldsymbol{\theta}(s_j) \mid \boldsymbol{\theta}(-s_j), \boldsymbol{\Sigma}] \sim \mathcal{N}\left(\frac{\sum_{s_r \in \partial s_j} \boldsymbol{\theta}(s_r)}{n(s_j)}, \frac{\boldsymbol{\Sigma}}{n(s_j)}\right) \\ \boldsymbol{\Sigma}^{-1} \sim \text{Wishart}(\nu,\ \boldsymbol{I}) \end{cases}
$$

The BLESS model hierarchy can be split into four parts: (1) Probit model, (2) latent model, (3) spike-and-slab prior, and (4) MCAR prior. Firstly, the probit model describes a generic Bayesian Generalised Linear Model (GLM) with a probit link function which relates the lesion probability $p_i(s_j)$ for every subject i at every voxel s_j to the linear predictor $\eta_i(s_j) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(s_j)$, where $\boldsymbol{x}_i$ contains the subject specific data and $\boldsymbol{\beta}(s_j)$ is a spatially varying parameter. Secondly, the latent model describes the commonly used data augmentation approach (Albert and Chib, 1993) for Bayesian probit regression which assumes that the binary outcomes $y_i(s_j)$ have an underlying normal regression structure on latent continuous variables $\boldsymbol{z}_i(s_j)$. Thirdly, a spatially varying, continuous version of the spike-and-slab prior (George and McCulloch, 1993) on the parameters $\boldsymbol{\beta}(s_j)$ in the form of a mixture of normal distributions where ν_0 is the spike variance, ν_1 defines the slab variance and $0 < \nu_0 < \nu_1$. Lastly, the parameter $\boldsymbol{\theta}(s_j)$ introduces the spatial structure within the probability of inclusion/exclusion $\sigma(\boldsymbol{\theta}(s_j))$, where $\sigma(\cdot)$ is the logistic function, with a multivariate conditional autoregressive (MCAR) prior. For further detail on the model configuration, we refer the reader to Section 2.2.

3.2.2 Bayesian-Bootstrap BLESS Algorithm

Algorithm 7: BB-BLESS

```
Result: Sample of parameter estimates $\tilde{\boldsymbol{\beta}}$ from approximate posterior
        distribution by re-weighting the likelihood function and re-centring
        the prior mean of the spike-and-slab prior.
Set: $\nu_0 = \nu_0^{DPE}$ spike variance (determined via DPE); $\nu_1$: large, fixed slab
  variance; $\alpha$: concentration parameter; $\epsilon$: convergence criterion; $B$: number
  of bootstraps
for $b$ = 1, ..., $B$ do
    1) Sample weights $\boldsymbol{w}^{(b)} \sim N \times \text{Dirichlet}(\alpha, \dots, \alpha)$.
    2) Sample mean shifts $\boldsymbol{\mu}^{(b)}(s_j)$ for all $j = 1, \dots, M$ from
      $\mu_p(s_j) \sim \mathcal{N}(0, \nu_0)$.
    3) Calculate $\tilde{\boldsymbol{\beta}}^{(b)}$ by acquiring variational posterior mean via
      approximating pseudo-posterior, defined in Equation (3.1), through
      variational inference.
```

We use Bayesian bootstrapping techniques (Nie and Ročková, 2022) to obtain approximate posterior samples. The approach for BB-BLESS is threefold and is described by Algorithm 7: Firstly, weights $w_i^{(b)}$ are sampled for every observation $i = 1, \ldots, N$ and every bootstrap iteration $b = 1, \ldots, B$ from a Dirichlet distribution, which has been scaled by the sample size N, in order to re-weight the likelihood following the weighted likelihood bootstrap by Newton and Raftery (1994). Every weight $w_i^{(b)}$ hereby perturbs the contribution of each observation to the likelihood. Secondly, a prior mean shift $\mu_p^{(b)}(s_j)$ is sampled for every covariate $p = 1, \ldots, P$, voxel $j = 1, \ldots, M$ and bootstrap iteration $b = 1, \ldots, B$ from the spike distribution. For initialisation, the variational parameters estimated when performing DPE for BLESS-VI (BLESS estimated via variational inference alone) at the target spike variance value ν_0 are used as initial values to BB-BLESS. This prior mean shift $\mu_p^{(b)}(s_j)$ is then used to center the prior for $\pi(\boldsymbol{\beta}(s_j)|\boldsymbol{\gamma}(s_j))$ on $\boldsymbol{\mu}^{(b)}(s_j)$ instead of $\mathbf{0}$ (Nie and Ročková, 2022). This combination of Bayesian bootstrap methods and the jittering of the spike-and-slab prior allows for approximate posterior sampling by repeatedly optimising the updated ELBO with respect to its variational parameters to approximate a posterior density. The variational posteriors for all other nuisance parameters are also re-fitted for every bootstrap sample. Thirdly, we acquire a sample $\tilde{\boldsymbol{\beta}}^{(b)}(s_j)$ by optimising the ELBO with respect to the spatially varying coefficient $\tilde{\boldsymbol{\beta}}^{(b)}(s_j)$. The following is a variational approximation to the pseudo-posterior, defined by a re-weighted likelihood and perturbed prior:

$$\begin{aligned} q(\tilde{\boldsymbol{\beta}}^{(b)}(s_j)) \propto \exp\Bigg\{ \mathbb{E}_{q(\boldsymbol{Z},\boldsymbol{\beta_0},\boldsymbol{\gamma},\boldsymbol{\theta},\boldsymbol{\Sigma}^{-1})} \Bigg[\ln \Bigg\{ \prod_{i=1}^{N} & [p(y_i(s_j)|z_i(s_j)) \\ & p(z_i(s_j)|\boldsymbol{\beta}(s_j), \boldsymbol{\beta_0}, \boldsymbol{x}_i)]^{w_i^{(b)}} \times \\ & p(\boldsymbol{\beta}(s_j)|\boldsymbol{\mu}^{(b)}(s_j), \boldsymbol{\gamma}(s_j))\ p(\boldsymbol{\beta_0})\ p(\boldsymbol{\gamma}|\boldsymbol{\theta})\ p(\boldsymbol{\theta}|\boldsymbol{\Sigma}^{-1})\ p(\boldsymbol{\Sigma}^{-1}) \Bigg\} \Bigg] \Bigg\}, \end{aligned} \tag{3.1}$$

where the Dirichlet weights are $(w_1^{(b)}, \ldots, w_N^{(b)}) \sim \text{Dir}(\alpha, \ldots, \alpha)$ and the jitter is drawn via the spike distribution $\mu_p^{(b)}(s_j) \sim \mathcal{N}(0, \nu_0)$. Each bootstrap sample $\tilde{\boldsymbol{\beta}}^{(b)}(s_j)$ is acquired by taking the marginal variational posterior mean of the pseudo-posterior defined in Equation (3.1) where the nuisance parameters are approximately marginalised out. Note that we prefer the variational posterior mean as opposed to the maximum-a-posteriori (MAP) estimate for each bootstrap draw due to the computational tractability of the former. Using the MAP estimate would result in having to use numerical optimisation at each iteration as some updates do not have a closed-form solution. However, taking the variational posterior mean as a quantity of interest is a natural choice because we then acquire the $\tilde{\boldsymbol{\beta}}(s_j)$ which maximises the variational posterior in the case of a normal posterior. We also acknowledge that while we do not provide theoretical guarantees for some aspects of this work, i.e. regarding likelihoods transformed in data augmentation strategies, we do validate all our work with numerical simulations. Specifically, regarding the treatment of the WLB in the application of the data augmentation approach by Albert and Chib (1993) for the likelihood of probit models in Equation (3.1). The full derivations of this method can be found in Appendix Section B.1.

3.3 Results

Firstly, we provide extensive simulation studies where we evaluate BB-BLESS (estimated via approximate posterior sampling defined in Algorithm 7) by comparing its performance to BLESS-VI (BLESS estimated via variational inference) and BLESS-Gibbs (BLESS estimated via Gibbs sampling) where we define the latter as the gold standard MCMC technique. Secondly, we showcase the scalability of our approximate posterior sampling algorithm by applying our work to the UK Biobank studying the association between lesion incidence and age while taking into account confounding variables, such as sex, head size scaling factor and age-by-sex interactions (Alfaro-Almagro et al., 2018). Thirdly, we demonstrate the benefit of cluster size-based imaging statistics on the UK Biobank application which can be acquired as a byproduct of the approximate posterior sampling algorithm. We also highlight that we are able to derive prevalence maps of cluster size which indicate the strength of our belief in cluster occurrence at a particular spatial location.

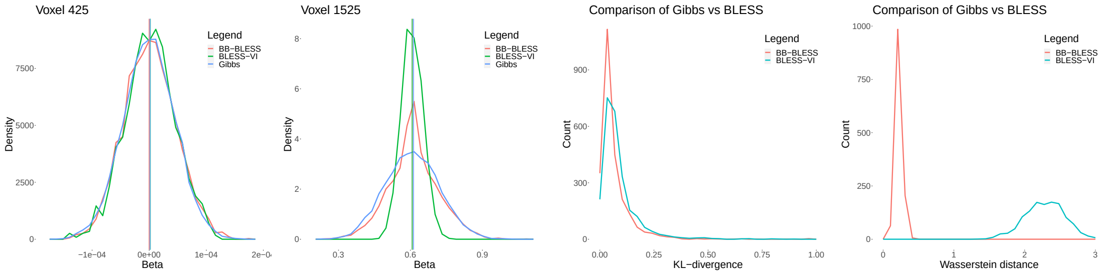
3.3.1 Simulation Study

The simulation study setup is identical to Section 2.3.1 from the previous chapter where the data generating process is different from the model for a fair comparison between methods. In this simulation study, we focus on a low base rate and sample size scenario ($N = 500$, $\lambda = 1$) as well as a high base rate and sample size scenario ($N = 1{,}000$, $\lambda = 3$) where we want to assess the performance of BLESS-VI, BB-BLESS and BLESS-Gibbs on two scenarios with small and large regression coefficients based on their base rate intensity $\lambda = 1$ and $\lambda = 3$. The posterior quantities of BLESS-VI are acquired by running a separate backwards dynamic posterior exploration procedure for every dataset with an equispaced spike sequence of $\nu_0 = \exp\{-20, \ldots, -1\}$ of length 15 and a slab variance of $\nu_1 = 10$. The method is initialised with the coefficients of Firth regression where we use the parameter estimates and respective inference results from the final run in the backwards DPE procedure ($\nu_0 = \exp(-20)$). We estimate BB-BLESS by drawing $B = 1,000$ bootstrap replicates and Dirichlet weights with a concentration parameter $\alpha = 1$. We run the Gibbs sampler for 15,000 iterations and discard 5,000 iterations as burn-in. The performance of BB-BLESS and BLESS-Gibbs is then greatly improved by utilising the output of the backwards DPE procedure as parameter initialisation for the respective parameter estimation techniques. For completeness, we report the results of the previously determined baseline methods, Firth regression and BSGLMM, alongside our work on BLESS-VI, BB-BLESS, and BLESS-Gibbs.

We also use a simulation framework developed by Kindalova et al. (2021) to generate realistic looking lesion masks utilising summary statistics from Firth regression based on UK Biobank data as truth. The results are similar and can be found in the Appendix Section B.4.

3.3.1.1 Results Interpretation

Firstly, we examine the marginal posterior densities of a random active and inactive voxel. As expected, the posterior variance from BLESS estimated via variational inference is underestimated as the posterior distribution is very peaked around the posterior mean (Figure 3.1a, b). On the other hand, the posterior estimated via BB-BLESS aligns well with the distribution acquired via the gold standard method of Gibbs sampling. This is further illustrated by comparing the marginal posterior densities of all voxels within an effect image via KL-divergence and Wasserstein distance in Figure 3.1c, d. Both methods show the higher quality of posterior approximation via BB-BLESS compared to BLESS-VI when calculating the discrepancy of the distributions acquired via approximate methods and Gibbs sampling.


(a) Inactive Voxel **(b)** Active Voxel **(c)** KL-divergence **(d)** Wasserstein metric

(e) Posterior Mean: Gibbs vs BB-BLESS **(f)** Posterior Mean: Gibbs vs BLESS-VI **(g)** Posterior Sd: Gibbs vs BB-BLESS **(h)** Posterior Sd: Gibbs vs BLESS-VI

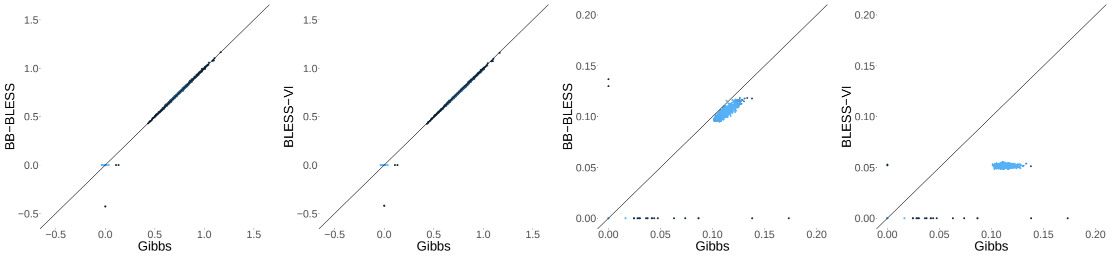


Figure 3.1: Comparison of marginal posterior distributions for an (a) inactive and (b) active voxel between BB-BLESS, BLESS-Gibbs, and BLESS-VI where the posterior mean is indicated via a vertical line. Overall evaluation of marginal posterior distributions for all voxels between Gibbs and BB-BLESS and BLESS-VI via (c) KL-divergence and (d) Wasserstein distance. Comparison of posterior quantities, such as posterior mean (e)-(f) and standard deviation (g)-(h), of the parameter estimates for all voxels for $N = 1,000$ and $\lambda = 3$ (lighter values indicate higher density of values). Parameters acquired via BLESS-VI exhibit similar point estimates to BB-BLESS and Gibbs but their posterior distributions are too peaked and variances are underestimated.

Figure 3.1 illustrates that both BB-BLESS and BLESS-VI are able to better capture the posterior mean of all voxel locations within an image when compared to BLESS estimated via Gibbs sampling. However, BLESS-VI severely underestimates the posterior standard deviation for both active and inactive voxels. Lastly, we compare the inference results of our method BLESS-VI, where we use the marginal posterior probability of inclusion as a proxy for inference, to the approximate posterior sampling technique BB-BLESS and the gold standard of BLESS-Gibbs, for which we determine activation via test statistics $t = \hat{\beta}/\sigma_{\hat{\beta}}$, for two simulation study setups. BLESS estimated via Gibbs sampling yields high sensitivity and a very low false positive rate for both settings in Table 3.1. More importantly, the inference results for BB-BLESS and BLESS-VI are very similar, i.e. the false positive rate for both BB-BLESS and BLESS-VI lies at 2.57% for a sample size of $N = 500$ and base rate intensity of $\lambda = 1$. Hence, we showcase empirically that, when it comes to inference, thresholding posterior inclusion probabilities in BLESS-VI yields similar results to the approximate posterior sampling approach BB-BLESS which determines effect detection via test statistics. Hence, if a researcher is uninterested in the additional features of BB-BLESS, such as acquiring uncertainty estimates of coefficients or more complex imaging statistics, then the application of BLESS-VI alone can be considered for parameter estimation and inference, as we get empirically similar voxelwise inference results in our simulation studies at a lower overall computational cost for BLESS-VI compared to BB-BLESS.

	$\boldsymbol{N = 500}$ **and** $\boldsymbol{\lambda = 1}$				$\boldsymbol{N = 1,000}$ **and** $\boldsymbol{\lambda = 3}$			
$\boldsymbol{\hat{\beta}_1}$	**Bias**	**Variance**	**MSE**		**Bias**	**Variance**	**MSE**	
BLESS-Gibbs	0.0501	0.0406	0.0414		0.0267	0.0065	0.0065	
BB-BLESS	0.0999	0.0171	0.0245		0.0423	0.0063	0.0064	
BLESS-VI	0.1022	**0.0013**	**0.0121**		0.0581	**0.0010**	**0.0010**	
BSGLMM	0.4412	0.0120	0.0147		0.3275	0.0039	0.0040	
Firth	0.1670	0.0539	0.0542		0.0912	0.0118	0.0118	
$\boldsymbol{t_{\hat{\beta}_1}}$	**TPR**	**TDR**	**FPR**	**FDR**	**TPR**	**TDR**	**FPR**	**FDR**
BLESS-Gibbs	0.7319	**0.9998**	**0.0001**	**0.0002**	0.9999	**0.9999**	**0.0001**	**0.0001**
BB-BLESS	0.6263	0.9606	0.0257	0.0394	0.9999	0.9970	0.0031	0.0030
BLESS-VI	0.6263	0.9606	0.0257	0.0394	**1.0000**	0.9970	0.0031	0.0031
BSGLMM	**0.9991**	0.9004	0.1128	0.0996	**1.0000**	0.9027	0.1088	0.0973
Firth	0.6566	0.8953	0.0768	0.1047	**1.0000**	0.9637	0.0379	0.0363

Table 3.1: Evaluation of parameter estimates and inference results for BLESS (Gibbs, BB, VI), BSGLMM and Firth for two simulation study scenarios ($N = 500$, $\lambda = 1$ and $N = 1,000$, $\lambda = 3$) for the effect of sex. MSE of parameter estimates for BB-BLESS and BLESS-Gibbs are comparable. Inference results from thresholding posterior inclusion probabilities for BLESS-VI and test statistics for BB-BLESS also achieve similar rates.

3.3.1.2 Run Time of Posterior Inference Algorithms

The main contribution of our work is the reliable but also scalable posterior uncertainty quantification of the spatially varying coefficients. In Table 3.2, we show the differences in run time between BLESS, estimated via Gibbs sampling, approximate posterior sampling, and variational inference, and the baseline mass-univariate approach Firth regression and spatial Bayesian model BSGLMM for various sample sizes $N = \{500;\ 1,000;\ 5,000\}$ and base rate intensities $\lambda = \{1, 2, 3\}$.

	N=500			**N=1,000**			**N=5,000**		
Model	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS-Gibbs	13:57	13:58	14:02	14:43	12:50	14:58	18:04	21:49	21:42
BB-BLESS	00:03	00:03	00:03	00:04	00:04	00:03	00:09	00:09	00:10
BLESS-VI	00:09	00:10	00:10	00:12	00:12	00:12	00:32	00:32	00:32
BSGLMM	06:44	07:44	08:23	08:26	08:35	08:14	15:14	15:49	15:04
Firth	00:01	00:01	00:01	00:02	00:01	00:01	00:04	00:04	00:03

Table 3.2: Average computational times (in hours) for all methods, BB-BLESS, BLESS-VI, BLESS-Gibbs, BSGLMM, and Firth Regression, across 100 simulated datasets. The average time for BB-BLESS is given by the average over all datasets as well as the average over all bootstrap iterations considering that the method is parallelisable for each bootstrap iteration. BLESS-Gibbs was run for 15,000 iterations with 5,000 iterations as burn-in which an effective sample size of 891 was achieved, BB-BLESS drew 1,000 assumed independent bootstrap samples, and BLESS-VI ran for cumulatively 3,541 iterations across the DPE procedure.

The methods relying on MCMC sampling, such as BLESS-Gibbs and BSGLMM, require the most run time to converge towards a posterior solution where BSGLMM requires between 6 to 15 hours and BLESS-Gibbs between 14 and 22 hours depending on the sample size and base rate intensity scenario. On the other hand, the mass-univariate approach only requires a 1 to 3 minutes to acquire parameter estimates which is unsurprising considering that the method is parallelisable across voxels due to the independent model configuration at each spatial location. BLESS-VI achieves run times between 10 to 30 minutes in this simulation setup and hence a computational improvement of approximately 83.7 to 43.63 times compared to BLESS-Gibbs which utilises MCMC sampling.

BLESS-VI also has a small sequential component in its posterior inference algorithm due to the included DPE procedure which is an annealing-like strategy that starts at an initial spike variance value (while keeping the slab variance at a fixed, large value) and gradually decreasing the spike variance value while utilising the previous model estimation via variational inference as a warm-start initialisation. Naturally, models estimated with a large spike variance do not provide a good solution for a posterior which we know to be highly multi-modal and sparse. Hence, the run time for BLESS-VI is larger than for BB-BLESS which takes the DPE solution of BLESS-VI to initialise its algorithm whereas BLESS-VI requires sequential model estimation for various spike variances within a range of $0 < \nu_0 < \nu_1$. Nevertheless, even when adding the run time of BLESS-VI to BB-BLESS, our algorithm BB-BLESS is 70 to 32 times faster than BLESS-Gibbs while possessing a similar level of accurate posterior uncertainty quantification for the spatially varying coefficients.

For a computational complexity analysis of BLESS-VI (and therefore a single bootstrap iteration of BB-BLESS), we refer the reader to Section B.2.

3.3.2 UK Biobank Application

In the real data application, we use the same UK Biobank cohort as in Section 2.3.2.1 with the aim of analysing the associations between lesion incidence and age while accounting for nuisance variables, such as sex, head size scaling factor and age-by-sex interaction (Alfaro-Almagro et al., 2021). The lesion masks of $N = 38,331$ subjects are generated via the automatic lesion segmentation algorithm BIANCA (Griffanti et al., 2016), the total number of voxels considered for analysis is determined by the white matter mask with $M = 54,728$ voxels, and the number of covariates in the analysis is $P = 4$. For further details, we refer the reader to Section 2.3.2.1.

For the model estimation of BB-BLESS, we acquire $B = 1,500$ bootstrap replicates in which we re-weight the likelihood by drawing Dirichlet weights for every subject with a concentration parameter $\alpha = 1$ and perturb the prior mean of the structured spike-and-slab prior by drawing a jitter from $\mathcal{N}(0, \nu_0)$. We initialise the parameters via the results from the DPE procedure and validate the behaviour of the annealing-like strategy by examining the regularisation and marginal plot, see Section 2.3.2.1. This approximate posterior sampling method remains highly scalable as each optimisation can be performed in parallel.

3.3.2.1 Results Interpretation

Figure 3.2 compares the raw effect size images of (a) BB-BLESS and (b) BLESS-VI where the shrinkage effect of the structured spike-and-slab priors for the negligible coefficients towards zero is shown in the white areas of the image maps. Both parameter maps are almost identical which is also reflected in the scatterplot in Figure 3.3 where the comparison between BLESS-VI and BB-BLESS coefficients showcases the alignment of posterior mean estimates between the two parameter estimation procedures. This similarity of methods is expected as BB-BLESS only provides more accurate uncertainty quantification for the spatially varying coefficients; however, the point estimates should remain the same for BLESS-VI and BB-BLESS.

For inference, we threshold the test statistics of BB-BLESS at a significance level of 5%. In contrast, for BLESS-VI we threshold the posterior probability of inclusion at 0.5 in order to acquire its respective binary significance map (Barbieri and Berger, 2004). For the covariate age, in the regression model estimated via BB-BLESS 8,385 effect locations are detected by utilising the full posterior to derive test statistics and similarly in BLESS-VI 8,257 effects are detected via simply thresholding the posterior inclusion probabilities. Acquiring similar binary significance results is beneficial as practitioners who have no need for proper uncertainty quantification can remain using BLESS-VI for posterior inference. BLESS-VI requires fewer computational resources than BB-BLESS which needs a pre-defined B bootstrap iteration for parameter estimation and inference.

3.3.2.2 Cluster Size-Based Imaging Statistics

Cluster-extent based thresholding is the most commonly used inference technique for statistical maps in neuroimaging studies. By proposing BB-BLESS and hence by sampling from an approximate posterior, we are able to provide novel cluster size-based imaging statistics, such as cluster size credible intervals, in addition to reliable uncertainty quantification of the spatially varying coefficients.

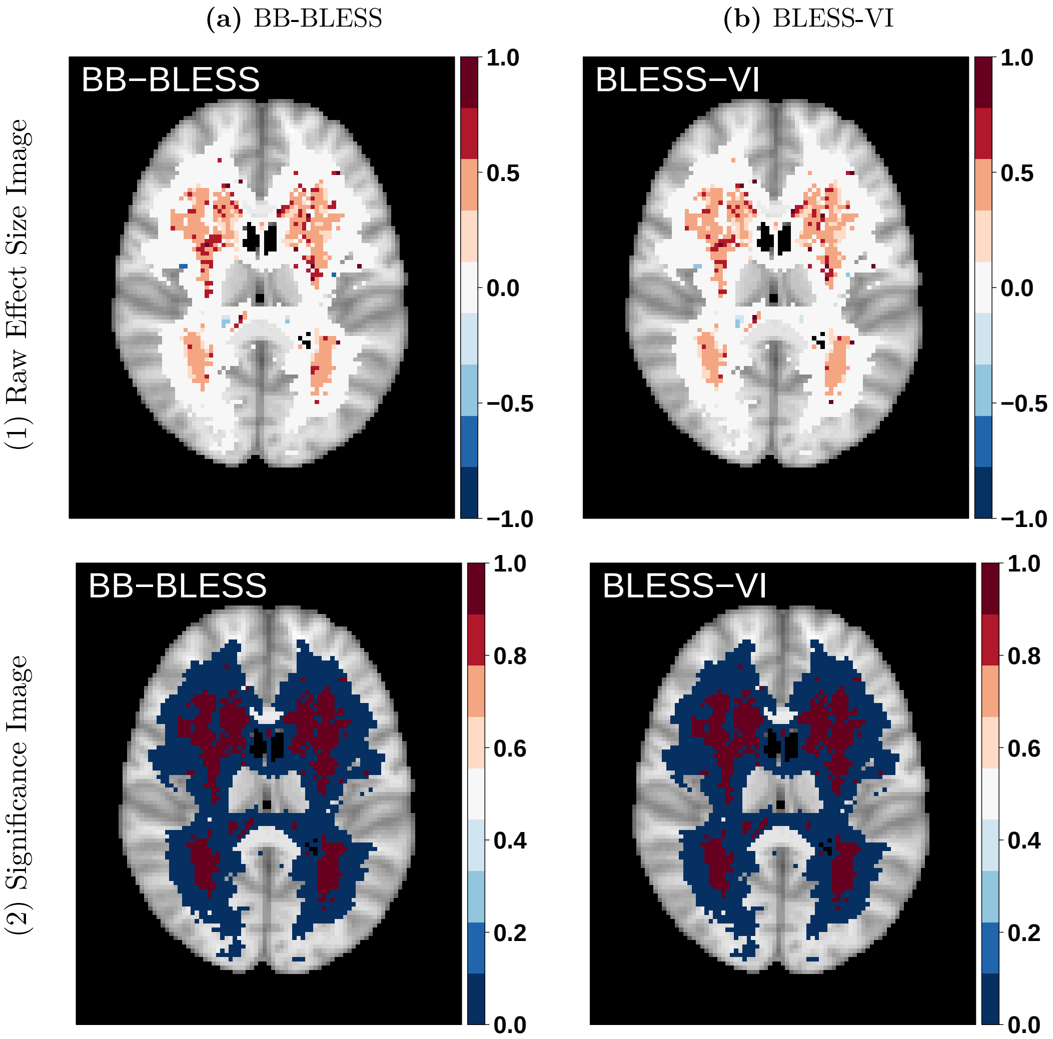


Figure 3.2: Comparison of results between (a) BB-BLESS and (b) BLESS-VI Regression for a single axial slice ($z = 45$, third dimension of 3D image). (1) spatially varying age coefficient maps. (2) Thresholded age significance maps where the threshold for BLESS-VI is determined via the probability of inclusion/exclusion $P(\gamma_p(s_j)|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) \geq 0.5$ and BB-BLESS via the test statistic $t = |\hat{\beta}/\hat{\sigma}_{\hat{\beta}}| \geq 1.96$ (significant voxels: red, not significant voxels: blue, FDR-correction applied at 5%). The parameter maps estimated via BB-BLESS in (1a) and BLESS-VI (1b) are almost identical which also applies for the significance maps showcasing that binary significance is nearly equivalent for both thresholding mechanism whereas the quality of uncertainty quantification is better for BB-BLESS compared to BLESS-VI.

We have shown that the raw age effect size image of the posterior mean for BB-BLESS in Figure 3.2 obtained as an average over the bootstrap replicates $\bar{\beta}(s_j) = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}^{(b)}(s_j)$ is almost identical to BLESS-VI. From our simulation studies, we expect the point estimates from both approximate methods to be

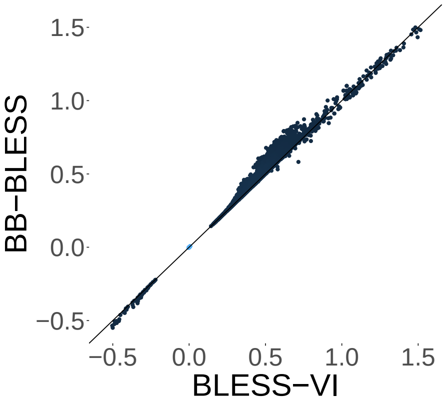


Figure 3.3: Scatterplot comparing the age coefficients for all voxel locations within the 3D image (lighter values indicate higher density of values). The scatterplot shows that BLESS-VI and BB-BLESS yield almost identical parameter estimates.

identical. Moreover, BB-BLESS is able to capture the uncertainty of coefficients more reliably than BLESS-VI. This is important not only for providing uncertainty quantification at a population level but also for evaluating the predictive performance via posterior predictive checks and ensuring model robustness via calibration plots (see Section B.5). With the latter we show that our model is well calibrated in predictive uncertainty and hence the model performance of BLESS does not change to a large extent when using new out-of-sample data. The approximate posterior samples can also be utilised to calculate cluster size-based imaging statistics which require test statistics in their estimation. The statistical map in the top middle part of Figure 3.4, acquired by the standardisation of the raw effects $t(s_j) = \bar{\beta}(s_j)/\sigma_\beta(s_j)$ with the posterior standard deviation $\sigma_\beta(s_j)$, shows that voxelwise inference based on thresholding the posterior probabilities of inclusion from BLESS-VI at 0.5 is similar to thresholding test statistics at a significance level of $\alpha = 5\%$.

We now highlight two novel cluster size approaches, based on cluster size inference and cluster size mapping, that can be calculated via BB-BLESS. In the first approach we acquire credible intervals of cluster size by utilising the more accurate posterior standard deviation estimates of BB-BLESS to standardise the bootstrap samples. The resampled statistical maps are then thresholded by a cluster-defining threshold of 2.3 (equivalent to thresholding p-values at a significance level of 0.01) which generates cluster size maps for every bootstrap replicate. We then build a distribution of cluster size by identifying the intersection of each bootstrap cluster with the observed one and recording its respective cluster size. The observed cluster size map is hereby defined as the cluster size map acquired by thresholding the test statistic map of the posterior mean divided by the posterior standard deviation of BB-BLESS with the same cluster-defining threshold. The observed cluster map is necessary to map overlapping clusters when wanting to plot a cluster size distribution for a specific cluster. A distribution over cluster sizes for any cluster within the brain allows us to calculate an array of statistical quantities, such as credible intervals of cluster size. In the top right part of Figure 3.4, we display the cluster size distribution for the largest cluster identified across the brain alongside its 95% credible interval which ranges between a cluster size of 4,063 and 4,265 voxels and contains the observed cluster size value of 4,179 voxels.

In a second cluster size mapping approach, we compute the voxelwise posterior probability of the standardised effect exceeding 2.3. This allows us to create a map of not just large effect but reliably large effect voxels. Comparing the cluster map with the occurrence map provides a measure for the reliability of cluster occurrence at a particular location within the brain. Due to the large spatial extent of the effect of age across the brain and viewing the central axial slice of the 3D maps, almost all voxel locations have a cluster prevalence close to 1. For these locations we then report posterior mean and standard deviation of reliable cluster size at locations where the prevalence of a cluster exceeds 50%.

To summarise, Figure 3.2 highlights how BLESS is able to reduce the identification of spurious associations for high-dimensional problems by shrinking the model's negligible coefficients to zero and leaving larger effects unaffected. In the UK Biobank, where we study how age is associated with occurrence of lesions, age has a very large effect on lesion incidence. However, many studies require methods to identify much subtler risk factors for lesion incidence and BLESS is able to identify these smaller effects alike with a higher level of specificity and sensitivity compared to the mass-univariate approach. Our methods aid a more accurate spatial localisation

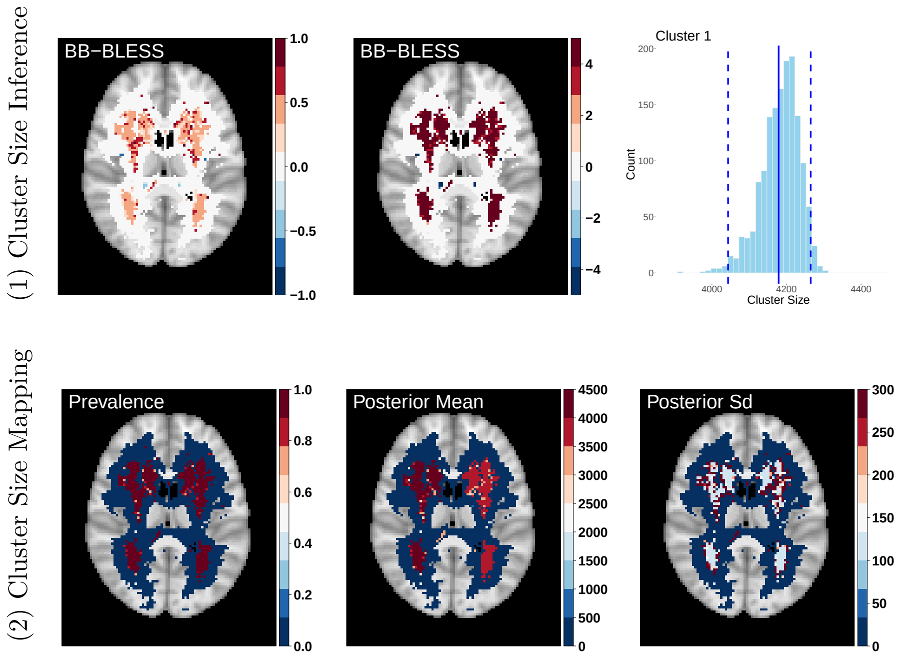


Figure 3.4: (1) Cluster Size Inference: Top left: Raw age effect size image. Top middle: Test statistic map for age effect. Top right: Cluster size distribution for the largest cluster detected by a cluster defining threshold of 2.3 (The solid line indicates the observed cluster size from BLESS-VI and the dashed lines signify the 95% credible interval of cluster size.). **(2) Cluster Size Mapping:** Lower left, middle, right: Prevalence, posterior mean and posterior standard deviation map of cluster size, where the latter two statistics are determined for instances where the prevalence map exceeds a probability of 50%. The prevalence map here indicates that both clusters have reliably large effects with values close to 1.

of effects where we show that the effect of age on lesion incidence predominantly covers periventricular and deep white matter regions. Hence, our model provides us with a tool to identify the brain regions impacted by lesion occurrence within a large population and to determine the impact on cognitive, sensory, or autonomic loss potentially induced by a higher lesion burden due to increased age. In Figure B.12 we also highlight the change of lesion incidence across the brain in 1000 out-of-sample subjects under 50 years and over 75 years old. While the lesion location is still focused around the ventricles, the predicted lesion incidence increases greatly with age which validates previous research findings (Kindalova et al., 2021).

3.4 Discussion and Future Work

With this work, we are able to provide uncertainty estimates of parameter maps which help in the assessment of spatial associations at a population level, attaching risk assessments for identified biomarkers for diseases, and enabling the acquisition of cluster size-based imaging statistics. The UK Biobank application for analysing the effect of age on lesion occurrence only identifies two big clusters, which validates our expectation based around the magnitude of the effect for age. More importantly, BB-BLESS has the unique advantage to provide us with prevalence statements of cluster size quantities. A spatial map that can aid decisions for follow-up studies, when resources are scarce and a researcher needs to know the reliability of large effect voxels and cluster occurrence across the brain.

We would therefore like to stress that we limit our application of BB-BLESS to a clinical application aiming at identifying the association between age and lesion incidence in a large-scale population health study, the UK Biobank. However, it is a well established finding in the analysis of white matter hyperintensities that age is one of the strongest predictors of lesion incidence (Wardlaw et al., 2013). Therefore, the study of more subtle risk factors for disease poses an interesting future research direction, as for example the further exploration of the cognitive impact of cerebrovascular risk-related white matter lesions (Veldsman et al., 2020).

For future work, it is also of interest to study the theoretical properties of BB-BLESS or more generally the effect of using weighted likelihood bootstrap in combination with a prior perturbation of mixture of normal piors with variational inference used for optimisation in high-dimensional regression problems, specifically with respect to the choice of optimal concentration parameter of the Dirichlet weights and distribution to draw prior mean perturbations from.

4

Scalable Scalar-on-Image Cortical Surface Regression with a Relaxed-Thresholded Gaussian Process Prior

Contents

In this chapter, we switch our focus from volumetric- to cortical surface-based approaches and from image-on-scalar to scalar-on-image regression problems. Specifically, we develop a hierarchical Bayesian spatial scalar-on-image regression with a relaxed thresholded Gaussian process prior in order to respect the potentially smooth and spatially-continuous nature of cortical surface fMRI data. With this work, we aim at increasing the scalability of Gaussian process priors for neuroimaging applications by using variational inference and the Karhunen-Loève expansion. Moreover, we again exploit Bayesian variable selection for inference and test our work extensively via simulation studies. For the real data application, we use the cortical surface task fMRI data from the ABCD study to associate intelligence with the emotional n-back task.

4.1 Introduction

4.1.1 Motivation for Analysis of Cortical Surface Data

Functional magnetic resonance imaging measures brain activity and can be used to identify interactive relationships between brain areas. Measuring brain activity plays a crucial role in understanding the human brain, in healthy individuals as well as patient populations. fMRI works by measuring local changes in blood flow that reflects changes in brain activity, and is used with cognitive tasks to identify brain regions sub-serving those tasks (see Section 1.1.1 for an overview on fMRI). MRI modalities, such as structural or functional MRI, can be displayed in either a volumetric- or cortical surface-based representation. In Section 1.1.3, we discuss the reasons for analysing MRI applications in the surface space, rather than using traditional volumetric-based representations, when it comes to the development of statistical methods for Bayesian spatial models. In short, the main motivation for a surface-based representation is to avoid the signal contamination that occurs with 3D smoothing, such that parts of the folded cortex are blurred even if the area consists of a part of the cortex that is quite distant in the unfolded space. Smoothing on the cortical surface respects the appropriate spatial distance of points on the cortex and hence has the potential to enhance signal (Brodoehl et al., 2020).

The non-invasive nature of MRI also allows for the analysis of brain structure and function and its impact on complex human traits, such as general cognition. The relationship between human intelligence and brain features has been extensively studied with respect to cortical volume and thickness (Pol et al., 2006; Narr et al., 2007; Yu et al., 2008; Karama et al., 2011; Jung and Haier, 2007) but also in terms of assessing the influence of white matter integrity across the brain on the neurodevelopment of cognitive traits (Penke et al., 2012; Muetzel et al., 2015; Yu et al., 2008); see Oxtoby et al. (2019) for an overview. Understanding the relationship between brain features and intelligence during developmental years in preadolescence can provide insights into biomarkers of resilience; however, recent findings have challenged the reproduciblity of any studies aiming to predict cognitive traits, including any markers of intelligence, arguing that sample sizes are not large enough (Zhao et al., 2023; Marek et al., 2022; Kennedy et al., 2022; Dick et al., 2021).

Currently, massive datasets are playing a central role in neuroimaging, and the need for scalable statistical methods that can applied to thousands and in the future hundreds of thousands subjects are essential (Marek et al., 2022). Many population-based health studies, such as the Adolescent Brain Cognitive Development (ABCD) study (Casey et al., 2018), the UK Biobank (UKBB) (Miller et al., 2016), and the Human Connectome Project (HCP) (Van Essen et al., 2012), not only have larger sample sizes but also contain various other data sources, such as omics data, environmental factors, mental health questionnaires, and further clinical variables. Hence, enabling the study of more complex models which respect the spatial dependency and the multi-modality of the data and are scalable to thousands of subjects.

Several approaches have been developed specifically to analyse data from the ABCD study, a study containing over 10,000 subjects with preprocessed cortical surface data as well as multiple other imaging modalities that can aid the study of cognitive traits in preadolescents. For example, Liu et al. (2022) study the association of the emotional n-back task functional MRI on the psychopathology factor using a thresholding function and the robust Huber loss to introduce an efficient nonconvex estimation method that allows for the modelling of complex dependence structures and is robust against heavy tails and outliers in the outcome. Wu et al. (2022c) take a geometric deep learning based approach to study fluid intelligence in adults and youths by performing a scalar-on-image regression by using age dependent cortical and subcortical morphologic interactions. Another application of machine learning for scalar-on-image regression was developed by Morris et al. (2022) who study brain connectivity networks and their clinical characteristics in the ABCD study.

We want to extend the available methodology on cortical surface regression problems by studying the scalar-on-image regression, where the output is a scalar quantity and the input is an image (see Section 1.3.1 for an overview), of associating the composite intelligence score (Luciana et al., 2018) as the output with the test statistics returned from the first-level analysis of the emotional n-back task fMRI with a 2- vs. 0-back contrast (Casey et al., 2018) as the input. Note that due to the group-level analysis of fMRI data we no longer have a time component in the analysis but work with number of subjects and number of spatial locations alone. Our main objectives are the following: 1) providing scalable inference suitable for the analysis of a dataset of the size of the ABCD study via a variational approximation to the posterior and a Karhunen-Loève expansion of a Gaussian process, 2) overcoming the non-identifiability problem of scalar-on-image regressions via relaxed-thresholded Gaussian process priors, and 3) validating our results via simulation studies and a real data application to the ABCD study.

4.1.2 Thresholded Gaussian Processes

Throughout this work we use boldface to indicate a vector or a matrix. Gaussian processes (GPs) are distributions over functions which map from an input $\boldsymbol{x} \in \mathbb{R}^d$ to an one-dimensional output $f(\boldsymbol{x})$. They are able to capture flexible model structures and provide uncertainty estimates about a function given the data. GPs can also be used for prior specifications over functions in a Bayesian model. A Gaussian process is hereby defined as a collection of random variables where any finite set of unique $\{\boldsymbol{x}_i\}_{i=1}^n$ implies that the output $\boldsymbol{f} = [f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_n)]^T$ is a multivariate Gaussian distributed with a mean $\boldsymbol{m}$ and covariance $\boldsymbol{C}$.

A Gaussian process is fully specified by its mean function $m(\boldsymbol{x})$ and covariance function $C(d(\boldsymbol{x}, \boldsymbol{x}'))$ of the real process $f(\boldsymbol{x})$, where $d(\boldsymbol{x}, \boldsymbol{x}')$ defines the distance function between the two input points with $C(d(\boldsymbol{x}, \boldsymbol{x}')) \to 0$ as $d(\boldsymbol{x}, \boldsymbol{x}') \to 0$, such that

$$\begin{aligned} f(\boldsymbol{x}) &\sim \mathcal{GP}(m(\boldsymbol{x}), C(d(\boldsymbol{x}, \boldsymbol{x}'))) \\ m(\boldsymbol{x}) &= \mathbb{E}\left[f(\boldsymbol{x})\right] \\ C(d(\boldsymbol{x}, \boldsymbol{x}')) &= \mathbb{E}\left[\{f(\boldsymbol{x}) - m(\boldsymbol{x})\}\{f(\boldsymbol{x}') - m(\boldsymbol{x}')\}\right], \end{aligned}$$

where the choice of mean and covariance functions determine the properties of the GP. The mean function is responsible for defining the overall trend of the function and the covariance function describes the level of smoothness and correlation between the input points. Throughout this work we choose the zero function for the mean function, so that $m(\boldsymbol{x}) = 0$ for all input points $\boldsymbol{x}$, and we note that the stationary, positive definite correlation function depends on the hyperparameters of the kernel $\boldsymbol{\xi}$, so that $C_{\boldsymbol{\xi}}(d(\boldsymbol{x}, \boldsymbol{x}'))$. There are various options of covariance functions, such as the squared exponential kernel (Stein, 1999), the modified squared exponential kernel (Spanos et al., 2007), the Matérn kernel (Matérn, 2013), or other types of period kernels (Smola, 1998), each of which can be combined with other kernels or modified in order to exhibit certain desirable model qualities. In the following subsections we describe thresholded Gaussian processes that add sparsity and variable selection to the smoothness property of GPs by transforming the coefficients via thresholding functions. Figure 4.1 illustrates the effect of various thresholding functions, such as hard-, soft-, and relaxed-thresholded GPs, on a function draw from a standard GP.

4.1.2.1 Hard-Thresholded Gaussian Process

The hard-thresholded Gaussian process (HTGP) was first introduced as a Bayesian nonparametric prior by Shi and Kang (2015) to impose sparsity into spatially varying coefficient models and to perform variable selection in neuroimaging applications. Their proposed thresholded multi-scale Gaussian process is split into a global GP to account for the domain spatial dependence and a local GP to capture regional fluctuations in an image-on-scalar regression problem. Other approaches have since developed extended versions of the hard-thresholded Gaussian process, such as the

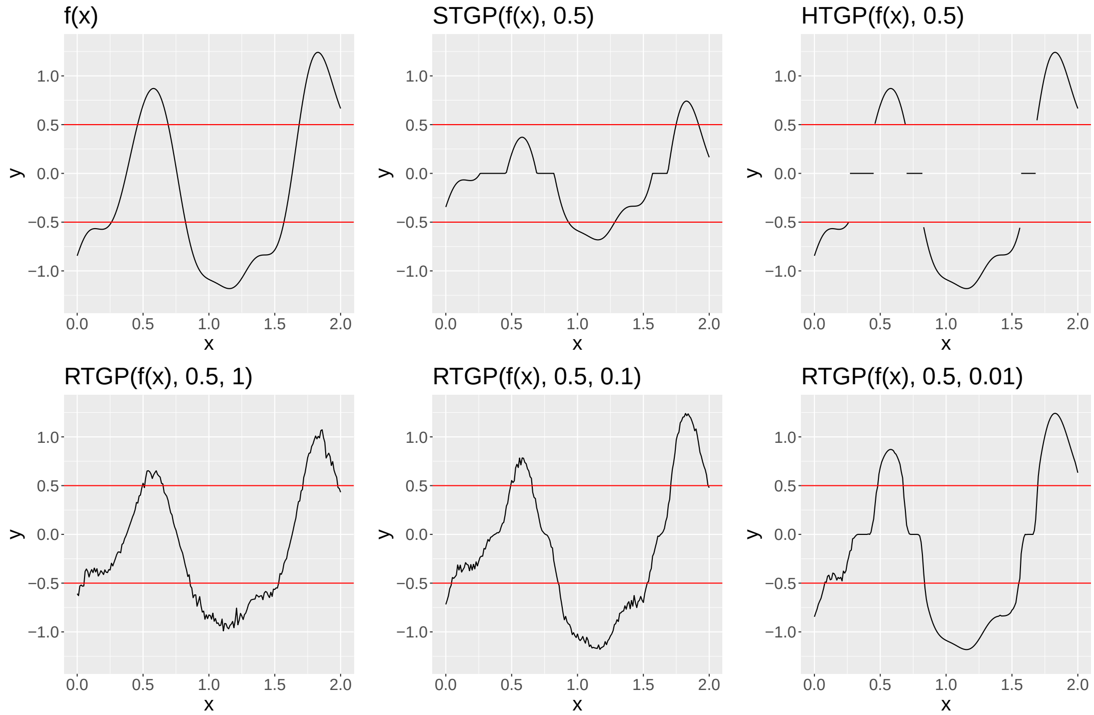


Figure 4.1: Illustration of a draw from a Gaussian Process with a squared exponential kernel $f(\boldsymbol{x})$, passing the GP through a soft-thresholding function and a hard-thresholding function with a threshold of $\delta = 0.5$ in the top row from left to right. The bottom row depicts averaged draws from a relaxed-thresholded Gaussian process with a threshold of $\delta = 0.5$ and varying relaxing parameters $\sigma_\alpha^2 = \{1, 0.5, 0.1\}$. (Adapted from Li (2023))

thresholded graph Laplacian Gaussian prior (Cai et al., 2020) or the application of the HTGP to a Bayesian spatial blind source separation problem (Wu et al., 2022a).

We define the HTGP by placing a standard GP prior over the spatially varying coefficients $\boldsymbol{\beta}(\boldsymbol{s})$, so that

$$\boldsymbol{\beta}(\boldsymbol{s}) \sim \mathcal{GP}\left\{\mathbf{0}, \sigma_\beta^2 C(d(\boldsymbol{s}, \boldsymbol{s}'))\right\},$$

where $\sigma_\beta^2 C(d(\boldsymbol{s}, \boldsymbol{s}'))$ is a stationary covariance function $\operatorname{cov}(\boldsymbol{\beta}(\boldsymbol{s}), \boldsymbol{\beta}(\boldsymbol{s}'))$. In a next step, the spatially varying coefficients $\beta(s_j)$ are transformed for all spatial locations $j = 1, \ldots, M$, where M is the number of voxels or vertices within an image depending on its representation in the volume or on the surface, by applying the hard-thresholding function

$$g_{\text{HTGP}}(x, \delta) = \begin{cases} 0, & |x| \leq \delta, \\ x, & |x| > \delta, \end{cases} \tag{4.1}$$

where the threshold $\delta > 0$ is considered a hyperparameter that controls the degree of sparsity. Hence, the thresholded spatially varying coefficients are given by passing the coefficients $\beta(s_j)$ through the hard-thresholding function $g_{\text{HTGP}}(\cdot)$ which yields $\beta_{\text{HTGP}}(s_j) = g_{\text{HTGP}}(\beta(s_j), \delta)$. Overall, this prior specification adds sparsity, accounts for spatial dependence, is piecewise-smooth, and is able to detect edge effects and jumps.

4.1.2.2 Soft-Thresholded Gaussian Process

The soft-thresholded Gaussian process (STGP) was developed by Kang et al. (2018) to introduce spatial variable selection into a Bayesian nonparamtric scalar-on-image regression model. This prior specification provides a large prior support over the class of piecewise-smooth, sparse, and continuous spatially varying regression coefficient functions, similar to trans-kriging (Cressie, 1993) or Gaussian copulas (Nelsen, 1999). Hence, enabling a gradual transition between the estimation of zero and nonzero effects in neighbouring locations. Identically to the HTGP, the spatially varying coefficients $\boldsymbol{\beta}(\boldsymbol{s})$ are drawn from a standard GP

$$\boldsymbol{\beta}(\boldsymbol{s}) \sim \mathcal{GP}\left\{\mathbf{0}, \sigma_\beta^2 C(d(\boldsymbol{s}, \boldsymbol{s}'))\right\}.$$

However, the coefficients are now passed through a soft-threshoulding function

$$g_{\text{STGP}}(x, \delta) = \begin{cases} 0, & |x| \leq \delta, \\ \operatorname{sgn}(x)(|x| - \delta), & |x| > \delta, \end{cases} \tag{4.2}$$

where $\operatorname{sgn}(x) = 1$ if $x > 0$, $\operatorname{sgn}(x) = -1$ if $x < 0$, and $\operatorname{sgn}(0) = 0$. Hence, the thresholded spatially varying coefficients are given by passing the coefficients $\beta(s_j)$ through the soft-thresholding function $g_{\text{STGP}}(\cdot)$ which yields $\beta_{\text{STGP}}(s_j) = g_{\text{STGP}}(\beta(s_j), \delta)$. The thresholding function $g_{\text{STGP}}(\cdot)$ determines the marginal distributions of $\beta_{\text{STGP}}(s_j)$ for all $j = 1, \ldots, M$ and the covariance function of $\boldsymbol{\beta}(\boldsymbol{s})$ determines the dependence structure of $\boldsymbol{\beta}_{\text{STGP}}(\boldsymbol{s})$. Equivalently to the HTGP, the STGP maps coefficients near zero to exact zero and hereby induces sparsity where the thresholding parameter $\delta > 0$ controls the level of sparsity.

4.1.2.3 Relaxed-Thresholded Gaussian Process

The relaxed-thresholded Gaussian process (RTGP) is able to capture both the HTGP and the STGP by introducing a set of latent variables $\boldsymbol{\alpha}(\boldsymbol{s}) \in \mathbb{R}^M$ with a "relaxing" parameter σ_α^2, the variance of the latent variables (Li, 2023). Thus, offering a more flexible Bayesian nonparametric prior than either the HTGP or the STGP alone, the RTGP is piecewise-smooth and sparse and models this sparsity by capturing both sparse and non-sparse patterns alike by varying the relaxing parameter.

The Gaussian process $\boldsymbol{\beta}(\boldsymbol{s})$ and the latent variables $\alpha(s_j)$ are defined by

$$\boldsymbol{\beta}(\boldsymbol{s}) \sim \mathcal{GP}\left\{\mathbf{0}, \sigma_\beta^2 C(d(\boldsymbol{s}, \boldsymbol{s}'))\right\} \tag{4.3}$$

$$\alpha(s_j) \sim \mathcal{N}\left(\beta(s_j), \sigma_\alpha^2\right), \tag{4.4}$$

where the latter is assigned for $j = 1, \ldots, M$ independently. However, compared to the previously introduced thresholding function, the relaxed-thresholding function does not threshold based on the values drawn from the Gaussian process $\boldsymbol{\beta}(\boldsymbol{s})$ but based on the values drawn from the latent distribution for $\boldsymbol{\alpha}(\boldsymbol{s})$, so that

$$g_{\text{RTGP}}(x, \tilde{x}, \delta) = \begin{cases} 0, & |\tilde{x}| \leq \delta, \\ x, & |\tilde{x}| > \delta. \end{cases} \tag{4.5}$$

Hence, the thresholded spatially varying coefficients are obtained by passing the coefficients $\beta(s_j)$ through the relaxed-thresholding function $g_{\text{RTGP}}(\cdot)$ which yields $\beta_{\text{RTGP}}(s_j) = g_{\text{RTGP}}(\beta(s_j), \alpha(s_j), \delta)$. The benefit of introducing additional latent variables is that the full conditional distribution of the spatially varying coefficients is now available in closed-form, as well as conjugate, which enables the formulation of a Gibbs sampler for efficient posterior computation.

The variance of the latent variables controls the independent noise added to the real process $\boldsymbol{\beta}(\boldsymbol{s})$. A small relaxing parameter σ_α^2 preserves the mean structure of $\boldsymbol{\beta}(\boldsymbol{s})$ fairly well and introduces only a slight jitter around the mean value. A large relaxing parameter σ_α^2 on the other hand allows for more flexibility. Figure 4.1 highlights this behaviour by showing how the RTGP with a relaxing parameter of $\sigma_\alpha^2 = 0.01$ converges to the HTGP, a RTGP with a relaxing parameter of $\sigma_\alpha^2 = 0.1$

mimics the STGP, and lastly a RTGP with a relaxing parameter of $\sigma_\alpha^2 = 1$ starts to converge towards the true signal $f(x)$.

4.1.3 Scalable Inference for Gaussian Processes

4.1.3.1 Spatial Process Approximations

Spatial process models can be computationally intensive. Neuroimaging applications in particular have thousands of spatial locations over which a spatial process is defined. Hence, the computational cost for model estimation can easily become infeasible in large-scale population-based health studies, such as the ABCD study, which contains data for thousands of subjects and various imaging modalities. The bottleneck in model fitting is the $\mathcal{O}(M^3)$ time complexity required for the inversion of some matrices when performing posterior inference, where M defines the number of spatial locations. Hence, as the number of spatial locations becomes large, computation becomes increasingly more difficult. Low-rank models or sparse models provide solutions for this problem.

The kernel convolution approximation of a spatial Gaussian process proposed by Higdon et al. (1999) is an example of a low-rank spatial model used to ease the computational burden that spatial models usually impose. Kang et al. (2018) use this method to scale their scalar-on-image regression approach to an EEG study with a large number of spatial locations and time points. The approach functions by introducing a smaller subset of locations, called “knots”, compared to the original number of spatial locations. Hence, if the number of knots r is lower than the number of spatial locations M, then computational savings are gained by reducing the time complexity from $\mathcal{O}(M^3)$ to $\mathcal{O}(Mr^2)$. However, the downside of low-rank models is that these approximate methods perform poorly if neighbouring locations are strongly correlated as the signal would dominate the noise (Stein, 2014) or the number of spatial locations is large as a large number of knots is required for a Gaussian process approximation of decent quality (Datta et al., 2016).

Sparse methods provide an alternative solution to the approximation of large spatial processes. For example, covariance tapering, as proposed by Furrer et al.

(2012), induces sparsity into the covariance kernel by using compactly supported covariance functions. On the other hand, Vecchia (1988) introduce sparsity into the precision matrix in products of lower-dimensional conditional distributions. However, the Vecchia approximation for spatial processes is computationally infeasible for large-scale applications. Hence, Katzfuss and Guinness (2021) extend the original framework to a sparse general Vecchia approximation for Gaussian processes that scale to large datasets. The main issue with sparse solutions is that inference is restricted to the estimation of the parameters of a covariance or precision function and no insight about the underlying process is gained like in the estimation of a lower-dimensional subset of realisations in the low-rank model approach (Datta et al., 2016).

Datta et al. (2016) propose a fully process-based inference approach to solve the aforementioned issues by introducing the so-called nearest-neighbour Gaussian process (NNGP). Sparsity is inserted into the model via neighbourhood sets constructed from directed acyclic graphs that extend finite dimensional models to a valid spatial process over uncountable sets. Wu et al. (2022b) extend the work by Datta et al. (2016) to a variational inference approximation of the NNGP.

4.1.3.2 Approximate Posterior Inference

The gold standard for parameter estimation and inference for Bayesian models with Gaussian process priors is MCMC sampling as it provides accurate uncertainty quantification of the spatially varying coefficients. However, image-based regression problems are extremely high-dimensional in nature and with large-scale population health studies becoming increasingly more accessible to researchers more scalable inference solutions need to be considered.

The following examples further highlight the need for scalablity in the analysis of large-scale Bayesian spatial models with thresholded GP priors. Firstly, the scalar-on-image regression model with a STGP prior proposed by Kang et al. (2018) uses Metropolis-Hastings within Gibbs sampling for posterior inference which can be computationally slow if the predefined acceptance ratio of the Metropolis-Hastings algorithm is not adequately tuned or can even lead to poor parameter estimates if the mixing of the coefficients is slow and convergence is not reached within the time limit or number of iterations set by the user. Similarly, the image-based regression model with a HTGP prior proposed by Shi and Kang (2015) suffers from non-conjugacy which requires the Metropolis-Hastings steps for posterior updates and uses the full conditionals for drawing samples from the remaining model parameters which exhibit conjugacy. Lastly, the scalar-on-image regression model with a RTGP prior proposed by Li (2023) is able to use full conditional distributions to update the model parameters with a Gibbs sampler. While the existence of closed form conditionals increases scalability, the model is still not scalable to the number of spatial locations found in MRI or large-scale population health studies, such as the ABCD study.

Hence, variational inference provides a more scalable posterior inference algorithm for image-based regression problems. Variational inference (Jordan et al., 1999) algorithms are computationally faster than MCMC methods by redefining the problem of approximating probability densities through optimisation (Blei et al., 2017). Moreover, Roy et al. (2021) provide an extensive overview on using mean-field variational inference for high-dimensional regression scenarios with sparse priors. Another example using a variational approximation for high-dimensional regression problems has been developed in our work in Section 2 and 3 for Bayesian lesion estimation with a structured spike-and-slab prior in which we use optimisation-based methods, such as variational inference and approximate posterior sampling, for posterior inference of an image-on-scalar regression problem. Variational inference hereby enables an application to a large-scale epidemiological study, the UK Biobank, and makes posterior inference feasible for approximately 40,000 subjects and 55,000 voxel locations.

The latter application highlights that variational inference can potentially suffer from some drawbacks that lead to poor variational approximations if convergence is slow, the variational family is chosen incorrectly or more generally the occurrence of underestimation of the posterior variance as the KL-divergence, used to measure the discrepancy between the posterior density and the variational density, underpenalises thin tails (Yao et al., 2018). However, approximate posterior sampling techniques, such as weighted likelihood bootstrap (Newton and Raftery, 1994), weighted Bayesian bootstrap (Newton et al., 2021), or other proposed approximate posterior sampling approaches (Lyddon et al., 2018; Fong et al., 2019; Nie and Ročková, 2022), enable the continued usage of optimisation for parameter estimation and with embarrassingly parallel implementations they recover the posterior uncertainty quantification in cases in which variational inference performs poorly. On the other hand, theoretical guarantees for approximate posterior sampling methods for high-dimensional and complex models are still nonexistent and moreover, while approximate posterior sampling techniques are parallelisable, they also incur a large computational cost due to the repeated optimisation that should mainly be considered if accurate uncertainty estimation is of interest.

4.2 Methods

Throughout this work, we assume that we can analyse the left and the right hemisphere of the cortex as separate analyses as each side of the cortex is responsible for its own functions or functions that occur in both hemispheres alike (Berlucchi, 1983). Within each analysis, we therefore make the assumption of a single hemisphere, cortical surface-based, spherical coordinate system (Fischl et al., 1999a) where we work with geodesic distances on the surface rather than Euclidean distances in the 3D volume.

Let $\mathbb{S}$ define the set of coordinates on a sphere with a known radius and $\mathcal{S} \subset \mathbb{S}$ as a set of vertices of a single hemisphere of the cortex at which MRI data is observed. In our real data application to the ABCD study, we work with data that has been mapped from the native patient's brain space with approximately 150,000 vertices to a normalised template brain space with approximately 30,000 vertices in $\mathcal{S}$. For any two spatial locations $\boldsymbol{s}, \boldsymbol{s'} \in \mathbb{S}$, we measure their distance $d(\boldsymbol{s}, \boldsymbol{s'})$ with the great-circle distance between $\boldsymbol{s}$ and $\boldsymbol{s'}$. We note that in theory we can let $\mathbb{S}$ represent any topological surface or even a volume in a volumetric-based analysis and $d(\cdot,\cdot)$ can also represent any appropriate distance metric. The distance between $\boldsymbol{s}$ and $\boldsymbol{s'}$ is used in $C_{\boldsymbol{\xi}}(d(\boldsymbol{s},\boldsymbol{s'}))$ which represents the stationary spatial correlation function defined on $\mathbb{S}$ with the kernel hyperparameters $\boldsymbol{\xi}$. Generally, the correlation function $C(\cdot)$ can be any positive definite kernel function defined so that $C(0) = 1$ and $C(d) \leq 1$ for all $d > 0$.

4.2.1 Model

We propose a Bayesian scalar-on-image regression model with a relaxed-thresholded Gaussian process prior. Firstly, the scalar output y_i for every subject $i = 1, \ldots, N$ is modelled with a Gaussian random variable

$$\left[y_i|\boldsymbol{x}_i, \tilde{\boldsymbol{\beta}}, \boldsymbol{\alpha}(\boldsymbol{s})\sigma_\epsilon^2\right] \sim \mathcal{N}\left\{\beta_0 + \sum_{j=1}^{M}\tilde{\beta}(s_j)I(|\alpha(s_j)| > \delta)x_{i,j}, \sigma_\epsilon^2\right\} \tag{4.6}$$

with the mean expressed by the linear predictor and the variance defined as the residual noise σ_ϵ^2. The linear predictor contains the sum of the intercept β_0 and the linear combination of imaging covariates $x_{i,j}$ for every vertex location $j = 1, \ldots, M$ and thresholded spatially varying coefficients $\beta_{\text{RTGP}}(s_j) = \tilde{\beta}(s_j)I(|\alpha(s_j)| > \delta)$, where $\tilde{\beta}(s_j)$ are spatially varying coefficients, $\alpha(s_j)$ are latent variables, and δ is the threshold parameter that determines the degree of sparsity in the model. Here we only consider one imaging modality, but this approach can be extended to consider multiple types of images to assist in the prediction of the scalar outcome.

Sparsity and smoothness is incorporated in the modelling of the spatially varying coefficients by placing a relaxed-thresholded Gaussian process on the coefficients. We put a Gaussian process prior over the spatially varying coefficients

$$\tilde{\boldsymbol{\beta}}(\boldsymbol{s}) \sim \mathcal{GP}\left\{\boldsymbol{0}, \sigma_\beta^2 C(d(\boldsymbol{s},\boldsymbol{s}'))\right\}, \tag{4.7}$$

with mean 0, stationary covariance kernel $C(\cdot)$ and parameter σ_β^2 which controls the maximum marginal variance of $\boldsymbol{\beta}(\boldsymbol{s})$. For the covariance function of the GP, we choose a two parameter exponential radial basis function

$$C(d(\boldsymbol{s},\boldsymbol{s}')) = \exp\left\{-\phi d(\boldsymbol{s},\boldsymbol{s}')^\nu\right\}, \tag{4.8}$$

where $\boldsymbol{\xi} = (\phi, \nu)^T$ are the kernel hyperparameters with $\phi > 0$ and $\nu \in (0, 2]$ (Jousse et al., 2021). The bandwidth or inverse scale parameter ϕ hereby controls how rapidly the correlation decays. On the other hand, the kernel exponent ν is responsible for how much smoothness is introduced. If the kernel exponent is $\nu = 2$, then the kernel is synonymous to the stationary and isotropic Gaussian kernel. Other covariance functions can also be considered; however, we choose the two parameter exponential radial basis function as Gaussian smoothing has a long history in MRI applications (Whiteman et al., 2023).

The latent variables $\alpha(s_j)$ are drawn from a Gaussian distribution of the form

$$\alpha(s_j) \sim \mathcal{N}(\tilde{\beta}(s_j), \sigma_\alpha^2) \qquad \text{for all } j = 1, \dots, M, \tag{4.9}$$

where σ_α^2 defines the relaxing parameter of the latent variable $\alpha(s_j)$ that determines if the thresholding property is closer to a HTGP or a STGP. Furthermore, we use a Karhunen-Loève expansion which represents a standard GP with a infinite number of basis functions using Mercer's theorem

$$\tilde{\beta}(s_j) = \sum_{l=1}^{\infty} \theta_l \sqrt{\lambda_l} \psi_l(s_j) \qquad \theta_l \sim \mathcal{N}(0, \sigma_\beta^2), \tag{4.10}$$

where θ_l are basis coefficients, $\{\lambda_l\}_{l=1}^{\infty}$ are eigenvalues in descending order and $\{\psi_l(s_j)\}_{l=1}^{\infty}$ are the corresponding orthonormal eigenfunctions. We truncate the basis expansion to approximate the Gaussian process with a finite number of parameters, so that

$$\tilde{\beta}(s_j) \approx \sum_{l=1}^{L} \theta_l \sqrt{\lambda_l} \psi_l(s_j), \tag{4.11}$$

where L defines the number of basis functions. The number of eigenfunctions can be determined by performing a principal components analysis where L is chosen so that a certain percentage of total variation is captured, so that

$$\min \left\{ l : \left(\sum_{k=1}^{l} \lambda_l \right) / \left(\sum_{k=1}^{\infty} \lambda_l \right) \geq \kappa_L \right\}, \tag{4.12}$$

for $\kappa_L \in (0, 1)$ where a larger κ_L provides a better approximation of the underlying GP.

For a full Bayesian hierarchical model specification, we place a Normal prior on the intercept $\beta_0 \sim \mathcal{N}(0, \sigma^2_{\beta_0})$, a discrete Uniform prior on the threshold parameter $\delta \sim \text{Uniform}(t_{\min}, t_{\max})$ with n_δ options equally spaced values between $t_{\min}$ and $t_{\max}$, and Half-Cauchy priors on the standard deviation parameters, specifically $\sigma_\beta \sim \text{Half-Cauchy}(0, s_\beta)$, $\sigma_\epsilon \sim \text{Half-Cauchy}(0, s_\epsilon)$, and $\sigma_\alpha \sim \text{Half-Cauchy}(0, s_\alpha)$. Lastly, we utilise the scale mixture representation of the Half-Cauchy priors which re-expresses the prior on the standard deviations as Inverse-Gamma priors by introducing the additional latent variables a_β, a_ϵ, and a_α, so that

$$\sigma^2_\beta \sim \text{Inverse-Gamma}(1/2,\ 1/a_\beta), \quad a_\beta \sim \text{Inverse-Gamma}(1/2,\ 1/s^2_\beta), \tag{4.13}$$

$$\sigma^2_\epsilon \sim \text{Inverse-Gamma}(1/2,\ 1/a_\epsilon), \quad a_\epsilon \sim \text{Inverse-Gamma}(1/2,\ 1/s^2_\epsilon), \tag{4.14}$$

$$\sigma^2_\alpha \sim \text{Inverse-Gamma}(1/2,\ 1/a_\alpha), \quad a_\alpha \sim \text{Inverse-Gamma}(1/2,\ 1/s^2_\alpha), \tag{4.15}$$

where s^2_β, s^2_ϵ, and s^2_α are considered hyperparameters of the model. We use scale-rate parameterisation of the Inverse-Gamma distribution.

4.2.2 Posterior Computation

Previous scalar-on-image regression models with HTGP (Shi and Kang, 2015) or STGP (Kang et al., 2018) prior distributions on the spatially varying coefficients require MCMC sampling for posterior computation which does not scale well to large-scale studies and high-dimensional voxel- or vertex-wise analyses commonly performed in neuroimaging applications. Specifically, both thresholded coefficient models rely on Metropolis-Hastings within Gibbs sampling with a kernel convolution approximation for spatial GPs for estimating model parameters. While this approach is computationally more efficient than other MCMC alternatives, it is nonetheless an approximation and moreover the quality of parameter estimation is sensitive to the choice of proposal distributions. Our contributions to increase the scalability for posterior computation are two-fold: 1) by deriving a Gibbs sampler to perform more efficient MCMC sampling by utilising the conjugacy within the RTGP model, and 2) by using variational optimisation algorithms rather than MCMC sampling to acquire posterior estimates.

4.2.2.1 Gibbs Sampler

For the Gibbs sampler of the Bayesian spatial regression model, we need the full conditional distributions of all parameters within the model. The set of model parameters is given by

$$\boldsymbol{\Omega} = \left\{\beta_0, \boldsymbol{\theta}, \boldsymbol{\alpha}(\boldsymbol{s}), \delta, \sigma_\epsilon^2, \sigma_\beta^2, \sigma_\alpha^2, a_\epsilon, a_\beta, a_\alpha\right\}. \tag{4.16}$$

The full set of derivations can be found in Appendix Section C.1. However, within this section we will introduce the posterior conditional distributions of the most crucial variables, such as the basis coefficient parameters $\boldsymbol{\theta}$, the latent parameters $\boldsymbol{\alpha}(\boldsymbol{s})$, and the threshold parameter δ.

Due to the Karhunen-Loève expansion, we no longer estimate the spatially varying coefficients themselves but rather acquire the posterior conditional distribution of the basis coefficients

$$[\boldsymbol{\theta} \mid \beta_0, \delta, \boldsymbol{\alpha}(\boldsymbol{s}), \sigma_\epsilon, \sigma_\beta, \sigma_\alpha, \boldsymbol{X}, \boldsymbol{y}] \sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right) \tag{4.17}$$

$$\begin{aligned}\boldsymbol{\mu}_{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \Bigg[\sigma_\epsilon^{-2} \left[\boldsymbol{y} - \beta_0 \mathbb{1}_n\right]^T \left[\boldsymbol{X} \operatorname{diag}\left\{I(|\alpha(s_j)| > \delta)\right\}_{j=1}^M \boldsymbol{\Psi}^T \operatorname{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^L\right] + \\ \sigma_\alpha^{-2} \boldsymbol{\alpha}(\boldsymbol{s})^T \boldsymbol{\Psi}^T \operatorname{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^L\Bigg]^T\end{aligned} \tag{4.18}$$

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \Bigg[\sigma_\epsilon^{-2} \left[\boldsymbol{X} \operatorname{diag}\left\{I(|\alpha(s_j)| > \delta)\right\}_{j=1}^M \boldsymbol{\Psi}^T \operatorname{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^L\right]^T \\ \left[\boldsymbol{X} \operatorname{diag}\left\{I(|\alpha(s_j)| > \delta)\right\}_{j=1}^M \boldsymbol{\Psi}^T \operatorname{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^L\right] + \sigma_\beta^{-2} \boldsymbol{I} + \\ \sigma_\alpha^{-2} \left[\boldsymbol{\Psi}^T \operatorname{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^L\right]^T \left[\boldsymbol{\Psi}^T \operatorname{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^L\right]\Bigg]^{-1}\end{aligned} \tag{4.19}$$

where the regularisation of the basis coefficients comes in through prior on the basis coefficients $\boldsymbol{\theta}$ as well as the prior on the latent variables $\boldsymbol{\alpha}(\boldsymbol{s})$. The posterior estimates of $\boldsymbol{\theta}$ then provide information on the spatially varying coefficients $\tilde{\boldsymbol{\beta}}(\boldsymbol{s})$ by plugging the posterior quantity of interest in Equation 4.10, so that the spatially varying coefficients $\tilde{\beta}(s_j) \approx \sum_{l=1}^L \theta_l \sqrt{\lambda_l} \Psi_l(s_j)$ are obtained.

Secondly, the posterior conditional distribution of the latent variables $\alpha(s_j)$ for each $j = 1, \ldots, M$ is acquired by a mixture of truncated normal distributions

$$\begin{aligned}[\alpha(s_j)|\boldsymbol{\theta}, \sigma_\alpha] \sim w_{-1,j} \text{ Truncated-Normal}_{(-\infty,-\delta)} &\left(\sum_{l=1}^{L} \theta_l \sqrt{\lambda_l} \psi_l(s_j), \sigma_\alpha^2\right) + \\ w_{0,j} \text{ Truncated-Normal}_{(-\delta,\delta)} &\left(\sum_{l=1}^{L} \theta_l \sqrt{\lambda_l} \psi_l(s_j), \sigma_\alpha^2\right) + \\ w_{1,j} \text{ Truncated-Normal}_{(\delta,\infty)} &\left(\sum_{l=1}^{L} \theta_l \sqrt{\lambda_l} \psi_l(s_j), \sigma_\alpha^2\right)\end{aligned}$$

where the mean is given by the spatially varying coefficients $\tilde{\beta}(s_j) = \sum_{l=1}^{L} \theta_l \sqrt{\lambda_l} \Psi_l(s_j)$ and the variance is given by the prior variance of the latent variables $\alpha(s_j)$. The mixture weights $\{w_{-1,j}, w_{0,j}, w_{1,j}\}$ are required to sum to 1 and are given by the following set of equations

$$\begin{aligned}\tilde{w}_{-1,j} &\propto \Phi\left(-\frac{\left[\delta + \sum_{l=1}^{L} \theta_l \sqrt{\lambda_l} \psi_l(s_j)\right]}{\sigma_\alpha}\right) \; c_j \\ \tilde{w}_{0,j} &\propto \left[\Phi\left(\frac{\delta - \sum_{l=1}^{L} \theta_l \sqrt{\lambda_l} \psi_l(s_j)}{\sigma_\alpha}\right) - \Phi\left(-\frac{\left[\delta + \sum_{l=1}^{L} \theta_l \sqrt{\lambda_l} \psi_l(s_j)\right]}{\sigma_\alpha}\right)\right] \\ \tilde{w}_{1,j} &\propto \left[1 - \Phi\left(\frac{\delta - \sum_{l=1}^{L} \theta_l \sqrt{\lambda_l} \psi_l(s_j)}{\sigma_\alpha}\right)\right] \; c_j\end{aligned}$$

alongside the identical normalising constant c_j within the non-thresholded mixture weights

$$\begin{aligned}c_j = \exp\Bigg\{-\frac{1}{2\sigma_\epsilon^2} &\left[\sum_{l=1}^{L} \theta_l \sqrt{\lambda_l} \psi_l(s_j)\right] \left[\sum_{l=1}^{L} \theta_l \sqrt{\lambda_l} \psi_l(s_j) \sum_{i=1}^{n} x_{i,j}^2 - \right. \\ &\left. 2\sum_{i=1}^{n} \left(y_i - \sum_{j' \neq j} \sum_{l=1}^{L} \theta_l \sqrt{\lambda_l} \psi_l(s_{j'}) I(|\alpha_{j'}| > \delta) x_{i,j'} - \beta_0\right) x_{i,j} \left(\sum_{l=1} \theta_l \sqrt{\lambda_l} \psi_l(s_j)\right)\right]\Bigg\}.\end{aligned}$$

For the constraint $\sum_{k \in \{-1,0,1\}} w_{k,j} = 1$ to hold, the mixture weights need to be renormalised by defining $w_{-1,j} = \tilde{w}_{-1,j} / \sum_{k \in \{-1,0,1\}} \tilde{w}_{k,j}$, $\tilde{w}_{0,j} = \tilde{w}_{0,j} / \sum_{k \in \{-1,0,1\}} \tilde{w}_{k,j}$, and $\tilde{w}_{1,j} = \tilde{w}_{1,j} / \sum_{k \in \{-1,0,1\}} \tilde{w}_{k,j}$.

Thirdly, the threshold δ is drawn from a discrete Uniform prior from the equispaced threshold options δ_t for $t = 1, \ldots, n_\delta$ between the prior bounds $t_{\min}$ and $t_{\max}$, where n_δ is the number of prior threshold options, with a posterior conditional probability of the threshold $\delta = \delta_t$, so that

$$\Pr(\delta = \delta_t \mid \cdot) \propto \frac{1}{n_\delta} \exp\left(-\frac{1}{2\sigma_\epsilon^2} \sum_{i=1}^{N} \left\{ y_i - \beta_0 - \sum_{j=1}^{M} \tilde{\beta}(s_j) I(|\alpha(s_j)| > \delta_t) x_{i,j} \right\}^2 \right).$$

Lastly, all further posterior conditional distributions are of standard form and can be found in Appendix Section C.1.

4.2.2.2 Variational Inference

The derivation of a Gibbs sampler for our Bayesian scalar-on-image regression model with a RTGP prior provides the foundation for more scalable posterior inference solutions. Specifically, we propose the usage of mean-field variational inference (Blei et al., 2017) rather than MCMC sampling to reduce the computational cost of this modelling framework, see Section 1.5 for a generic review on variational inference.

Variational inference requires the full joint distribution of the Bayesian spatial regression model which consists of the likelihood $p(\boldsymbol{y}|\boldsymbol{X}, \tilde{\boldsymbol{\beta}}(\boldsymbol{s}), \boldsymbol{\alpha}(\boldsymbol{s}), \delta, \beta_0, \sigma_\epsilon^2)$ and the joint prior distribution $p(\beta_0, \boldsymbol{\theta}, \boldsymbol{\alpha}(\boldsymbol{s}), \delta, \sigma_\epsilon^2, \sigma_\beta^2, \sigma_\alpha^2, a_\epsilon, a_\beta, a_\alpha)$ which is expressed through its prior conditional distributions. We use mean-field coordinate ascent variational inference (CAVI) (Bishop, 2006) where we update each variational posterior by specifying an appropriate variational distribution q on a model parameter $\boldsymbol{\omega} \in \boldsymbol{\Omega}$, where the set of model parameters $\boldsymbol{\Omega}$ is identical to Equation 4.16, and then by updating the model parameters while keeping all other parameters fixed.

In our model, the factorisation of the variational distributions for MFVI is expressed by

$$\begin{aligned} &q(\beta_0, \boldsymbol{\theta}, \boldsymbol{\alpha}(\boldsymbol{s}), \delta, \sigma_\epsilon^2, \sigma_\beta^2, \sigma_\alpha^2, a_\epsilon, a_\beta, a_\alpha) \\ &= q(\beta_0) \times q(\boldsymbol{\theta}) \times \prod_{j=1}^{M} q(\alpha(s_j)) \times q(\delta) \times q(\sigma_\epsilon^2) \times q(\sigma_\beta^2) \times q(\sigma_\alpha^2) \times q(a_\epsilon) \times q(a_\beta) \times q(a_\alpha). \end{aligned}$$

The CAVI algorithm iterates through the steps of determining a suitable variational distribution $q(\boldsymbol{\omega})$ and of updating its variational parameters, while keeping the other parameters in $\boldsymbol{\Omega}$ fixed, until a convergence criterion is satisfied. All derivation details and information on initialisation of the variational inference algorithm can be found in Appendix Section C.2.

4.3 Results

In this section, we perform both simulation studies and a real data application for a cortical surface-based scalar-on-image regression problem. We choose to compare our model, which is denoted with RTGP estimated via variational inference, to simple baseline models, such as a Bayesian regression model with a Normal prior (BR + Normal) and a Horseshoe prior (BR + Horseshoe) on the image coefficients $\beta(s_j)$ for each $j = 1, \dots, M$, and more complex baseline models, such as performing a Gaussian process regression by transforming the input image data with the bases, consisting of the eigenfunctions and eigenvalues of the kernel decomposition, and placing a Normal prior (GPR + Normal) and a Horseshoe prior (GPR + Horseshoe) on the basis coefficients θ_l for each $l = 1, \dots, L$. We also compare our model against baseline frequentist approaches, such as Ridge and LASSO regression. We provide a more detailed overview of the model setup of the baseline models in Appendix Section C.5.1 and a comparison of the performance of RTGP-VI (= RTGP estimated via variational inference) and RTGP-Gibbs (= RTGP estimated via Gibbs sampler) in Appendix Section C.5.2. For the baseline Bayesian models, we use the R package `fastBayesReg`[1] which implements MCMC via a Gibbs sampler. We run each chain for $n_{\text{iter}} = 5,000$ and apply a burn-in of $n_{\text{burn-in}} = 2,000$ to those samples. For the frequentist models, we use the R package "`glmnet`".

The distance function $d(\boldsymbol{s}, \boldsymbol{s}')$ is calculated for each combination of spatial locations on the cortical surface through the Python package `BrainSMASH`[2] (Burt et al., 2020). Note that we acquire this geodesic distance matrix for each cortical hemisphere separately by using the template mesh atlases provided in the HCP pipeline[3], specifically for the real data analysis we use the file "`L.sphere.32k_fs_LR.surf.gii`" for the left hemisphere and the file "`R.sphere.32k_fs_LR.surf.gii`" for the right hemisphere. For the simulation study, we use the subsampled version of the above

---

[1] `https://github.com/kangjian2016/fastBayesReg/tree/master`

[2] `https://brainsmash.readthedocs.io/en/latest/index.html`

[3] `https://github.com/Washington-University/HCPpipelines/tree/master/global/templates/standard_mesh_atlases`

surface template to 2,000 spatial locations via the R package `ciftiTools`[4] (Pham et al., 2022) command “`resample_cifti`”. We note that we use `ciftiTools` for all data manipulation and display of cortical surface data in our analysis. In the simulation study, we only focus on an analysis of the left hemisphere of the brain in order to lower the computational cost. However, in the real data analysis we perform a separate analysis for both hemispheres and hence any plots generated via `ciftiTools` display the left hemisphere on the left and and right hemisphere on the right whereas the difference between the top and bottom row plots is the perspective of the 3D cortex. For the Karhunen-Loève expansion we then require the eigenvalue decomposition of our kernel matrix $C(d(\boldsymbol{s}, \boldsymbol{s}'))$ and for this we use the R package `RSpectra`[5] which yields orthonormal eigenfunctions $\boldsymbol{\Psi}$ and their corresponding eigenvalues $\boldsymbol{\lambda}$. For the simulation study we use a number of basis functions $L = 100$, and for the real data application we use $L = 800$ which captures approximately 30% of the total variation determined through Equation (4.12), see Appendix Section C.6.2.

For a comprehensive evaluation of the performance of RTGP compared to baseline models, BR + Normal, BR + Horseshoe, GPR + Normal, GPR + Horseshoe, Ridge and LASSO regression, we choose to evaluate the quality of the parameter estimates via absolute bias $\beta_{\text{bias}} = \frac{1}{M}\sum_{j=1}^{M}\{|\beta(s_j) - \hat{\beta}(s_j)|\}$ and mean squared error (MSE) $\beta_{\text{MSE}} = \frac{1}{M}\sum_{j=1}^{M}\{\beta(s_j) - \hat{\beta}(s_j)\}^2$, the predictive results via coefficient of determination $R^2 = \text{cor}(y, \hat{y})^2$ and the predictive MSE $y_{\text{MSE}} = \frac{1}{N}\sum_{j=1}^{N}\{y - \hat{y}\}^2$, and the inference results by assessing true positive (TP), false positive (FP), true negative (TN), and false negative (FN) discoveries in the following measures: (1) sensitivity/true positive rate (TPR $= \frac{\text{TP}}{\text{TP+FN}}$), (2) true discovery rate (TDR $= \frac{\text{TP}}{\text{TP+FP}}$), (3) specificity/1 - false positive rate (FPR $= \frac{\text{FP}}{\text{FP+TN}}$), and (4) false discovery rate (FDR $= \frac{\text{FP}}{\text{FP+TP}}$). Binary significance is determined for RTGP by thresholding the expected value $\mathbb{E}[I(|\alpha(s_j)| > \delta)]$ for each vertex $j = 1, \ldots, M$ according to the median probability model (Barbieri and Berger, 2004) and for the other Bayesian approaches (BR + Normal, BR + Horseshoe, GPR + Normal,

---

[4] `https://cran.r-project.org/web/packages/ciftiTools/ciftiTools.pdf`

[5] `https://cran.r-project.org/web/packages/RSpectra/vignettes/introduction.html`

GPR + Horseshoe) by determining whether or nor 0 is included in the 95%-HPDI (highest posterior density interval). We use the R package "`bayestestR`"[6] to calculate the HPDI for each model.

4.3.1 Simulation Study

In the simulation study, we evaluate the performance of our model RTGP compared to baseline approaches with different simulation study settings, such as varying sample sizes ($N = \{500; 1,000; 2,000\}$). The test dataset in the respective simulation studies always consists of 1,000 subjects. In order to ensure the robustness of our results, we evaluate all results on 10 replicated datasets and average their results. While our focus is on the analysis of cortical surface data, we do provide simulation studies for the analysis of volumetric data where we use a modified squared exponential kernel with Euclidean distance as a distance measure and the R package "`BayesGPfit`"[7] to perform the eigendecomposition of the kernel matrix, see Appendix Section C.3. In the following subsections, we will first explain the data generating process, which is deliberately different from the model in order to provide a fair comparison between our approach RTGP and the other baseline approaches, and then interpret the results of our surface-based simulation study.

4.3.1.1 Data Generating Process

The data generating process for this surface-based simulation study is based on a realistic effect of interest. Firstly, we derive the true beta map by performing a mass-univariate image-on-scalar regression associating the 2- vs 0-back contrast of the emotional n-back task fMRI on intelligence scores and confounding variables (age, sex, family income, parental education level, marital status, race/ethnicity, and site scanner) on the ABCD study dataset described in Section 4.3.2.1 as part of our real data application. Secondly, we hard-threshold the acquired test statistic map from the previous mass-univariate image-on-scalar analysis by setting all vertex values below 10 to 0 and re-scale the true beta map by dividing it by 10. Thirdly, we lower

---

[6] `https://easystats.github.io/bayestestR/reference/hdi.html`

[7] `https://github.com/kangjian2016/BayesGPfit`

the computational cost of our simulation study by sub-sampling the true beta map, shown in Figure 4.2 (c), with the function provided by the R package `ciftiTools` from the original approximately 30,000 vertices to approximately 2,000 vertices. The input image $\boldsymbol{x}_i$ for every subject i is drawn from a GP with the kernel function defined in Equation (4.8) with the kernel hyperparmeters $\boldsymbol{\xi} = (\phi = 0.077, \nu = 2)^T$ where two examples are shown in Figure 4.2 (a) and (b). The scalar output y_i is generated for every subject i by plugging in the above quantities in the equation $y_i = \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}(\boldsymbol{s}) + \epsilon_i$, where the true intercept is $\beta_0 = 2$ and the random error is drawn from a Normal distribution $\epsilon_i \sim \mathcal{N}(0, \sigma^2 = 0.2)$.

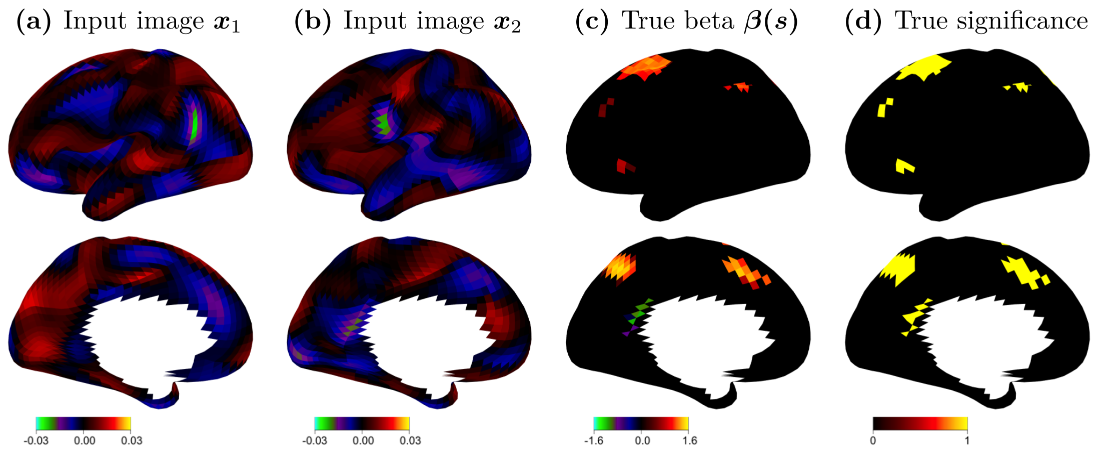

(a) Input image $\boldsymbol{x}_1$ **(b)** Input image $\boldsymbol{x}_2$ **(c)** True beta $\boldsymbol{\beta}(\boldsymbol{s})$ **(d)** True significance

Figure 4.2: Overview of data generating process with (a) and (b) showing examples of input images generated with a random draw from a GP, (c) specifies the true subsampled and hard-thresholded beta map from the mass-univariate analysis in Appendix Section C.4, and (d) the corresponding binary significance map.

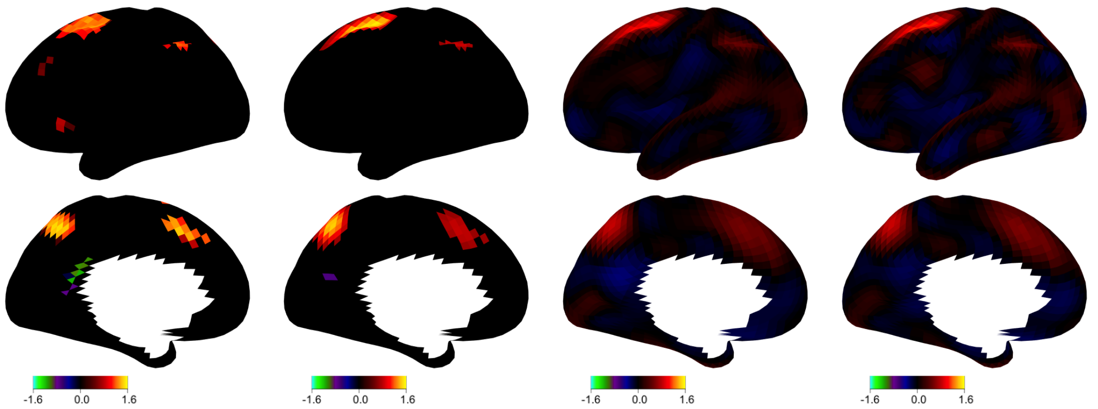
4.3.1.2 Results Interpretation

Firstly, we evaluate the quality of the parameter estimates and find that our model RTGP has consistently lower absolute bias and MSE for all estimated sample sizes than any of the other baseline approaches as shown in Table 4.1. Figure 4.6 supports that RTGP outperforms the other baseline approaches with respect to parameter estimation where RTGP is able to not only capture the true effect size magnitude much better than the other models but also is able to pick up where the spatial signal is located far more accurately than the other approaches which suffer from oversmoothing in the case of GPR + Normal, GPR + Horseshoe, BR + Normal, and Ridge regression. On the other hand, BR + Horseshoe and LASSO regression exhibit very localised, speckled and extreme estimated parameter values as these models ignore the spatial dependency between neighbouring vertices and put extreme weight on a single vertex rather than smoothing the effect. In doing so both models favour solutions which are sparse in nature and aim at preventing overfitting; however, in doing so accurate parameter estimation and inference is jeopardised while out-of-sample prediction still yields comparable or superior results (see Table 4.2).

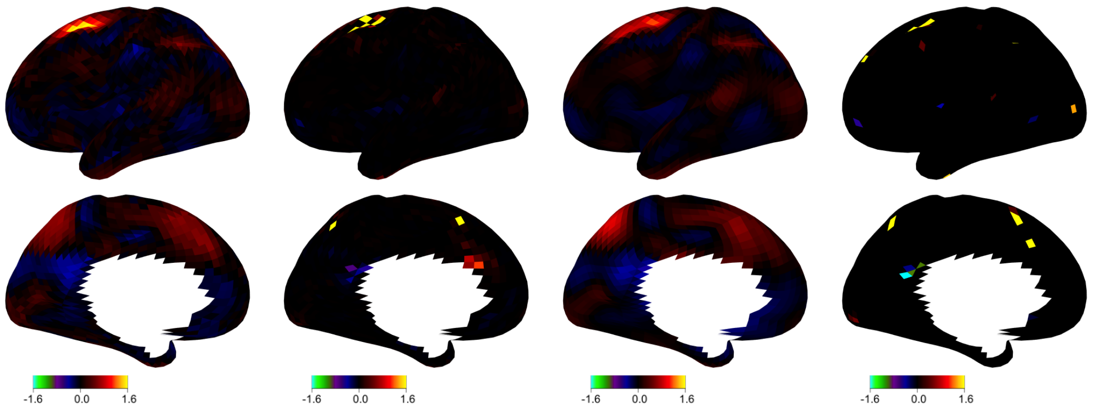
N=500	**Bias**	**MSE**
RTGP-VI	**6.56** (1.97)	**3.23** (0.69)
GPR + Normal	11.77 (0.40)	3.37 (0.14)
GPR + Horseshoe	12.48 (1.67)	3.74 (0.58)
BR + Normal	10.81 (0.19)	3.28 (0.11)
BR + Horseshoe	8.11 (0.31)	19.14 (4.16)
Ridge	11.50 (0.33)	3.47 (0.16)
LASSO	7.89 (0.43)	22.24 (4.74)
N=1,000	**Bias**	**MSE**
RTGP-VI	**3.74** (0.52)	**2.45** (0.33)
GPR + Normal	10.31 (0.50)	3.04 (0.09)
GPR + Horseshoe	11.04 (0.60)	3.21 (0.16)
BR + Normal	9.78 (0.29)	2.97 (0.06)
BR + Horseshoe	7.79 (0.85)	22.38 (15.89)
Ridge	10.51 (0.35)	3.17 (0.10)
LASSO	7.89 (0.43)	19.11 (3.80)
N=2,000	**Bias**	**MSE**
RTGP-VI	**3.22** (0.26)	**2.09** (0.18)
GPR + Normal	9.66 (0.22)	2.83 (0.05)
GPR + Horseshoe	11.44 (1.13)	3.18 (0.33)
BR + Normal	9.33 (0.20)	2.80 (0.03)
BR + Horseshoe	7.75 (0.41)	21.33 (6.57)
Ridge	10.22 (0.23)	3.04 (0.06)
LASSO	7.17 (0.28)	17.31 (3.85)

Table 4.1: Evaluation of parameter estimate results with bias and MSE mean (standard error) for a simulation study setting with varying sample sizes $N = \{500; 1,000; 2,000\}$ for our model RTGP-VI and the baseline models GPR + Normal, GPR + Horseshoe, BR + Normal, BR + Horseshoe, Ridge and LASSO. The values are multiplied by a scaling factor of 10^2 for clarity.

Secondly, we find that the out-of-sample prediction results are comparable across all models which is common in scalar-on-image regression problems which suffer



(a) True beta **(b)** RTGP-VI **(c)** GPR + Normal **(d)** GPR + Horseshoe



(e) BR + Normal **(f)** BR + Horseshoe **(g)** Ridge **(h)** LASSO

Figure 4.3: Comparison of the (a) true parameter map with the estimated parameter maps from (b) our RTGP-VI model and (c)-(h) the other baseline approaches (simulation study setting: N = 500). Note that the we limit the colourbar to the range of the true parameter values which lie between -1.6 to 1.6 for a fair comparison of models. However, the (f) BR + Horseshoe model and (h) the frequentist LASSO regression both exhibit far more extreme values which lie outside the plotted range.

from non-identifiability if no model constraints are imposed. The lowest possible MSE for example is given by the residual noise which is known in the simulation study with $\sigma_\epsilon^2 = 0.04$ (multiplied with scaling factor the true residual noise variance is $\sigma_\epsilon^2 = 4$). In the simulation studies, we also observe that approaches which have poor inference results, such as BR + Horseshoe for example with no observed activation, exhibit better predictive performance with a R^2 (test) = 41.07% and MSE (test) = 4.37 for a sample size of $N = 500$ whereas our RTGP model has a R^2 (test) = 39.38% and MSE (test) = 4.86, see Table 4.2 for further results. Hence, showcasing that many combinations of spatial locations can be predictive of our outcome of interest; however, depending on what a researcher is interested in it can be favourable to choose a model which has slightly worse out-of-sample predictive results with respect to R^2 and MSE but has superior inference results with lower false positive activation compared to the other approaches.

N=500	**$\mathbf{R}^2$ (train)**	**MSE (train)**	**$\mathbf{R}^2$ (test)**	**MSE (test)**
RTGP-VI	49.89 (1.41)	3.92 (0.31)	39.38 (2.33)	4.86 (0.30)
GPR + Normal	55.65 (2.23)	3.37 (0.30)	35.32 (2.21)	4.82 (0.23)
GPR + Horseshoe	53.66 (3.00)	3.58 (0.36)	34.21 (2.37)	4.88 (0.22)
BR + Normal	**57.18** (2.65)	**3.32** (0.33)	35.60 (2.41)	4.78 (0.21)
BR + Horseshoe	53.03 (3.04)	3.54 (0.34)	**41.07** (2.56)	**4.37** (0.19)
Ridge	56.95 (2.69)	3.35 (0.32)	35.48 (2.35)	4.79 (0.22)
LASSO	52.50 (3.35)	3.67 (0.41)	40.22 (2.51)	4.45 (0.15)
N=1,000	**$\mathbf{R}^2$ (train)**	**MSE (train)**	**$\mathbf{R}^2$ (test)**	**MSE (test)**
RTGP-VI	44.05 (2.30)	4.08 (0.19)	41.75 (2.23)	4.32 (0.17)
GPR + Normal	49.06 (2.20)	3.71 (0.17)	39.15 (1.69)	4.46 (0.16)
GPR + Horseshoe	48.59 (2.42)	3.74 (0.16)	39.03 (1.35)	4.47 (0.15)
BR + Normal	**50.20** (2.15)	**3.63** (0.15)	39.79 (1.58)	4.42 (0.17)
BR + Horseshoe	48.22 (2.29)	3.74 (0.15)	**42.70** (1.62)	**4.20** (0.15)
Ridge	50.13 (2.15)	3.64 (0.14)	39.65 (1.53)	4.43 (0.18)
LASSO	47.72 (2.29)	3.83 (0.16)	42.54 (1.66)	4.25 (0.16)
N=2,000	**$\mathbf{R}^2$ (train)**	**MSE (train)**	**$\mathbf{R}^2$ (test)**	**MSE (test)**
RTGP-VI	42.22 (1.60)	4.20 (0.14)	42.66 (2.32)	4.26 (0.20)
GPR + Normal	45.86 (1.71)	3.95 (0.14)	41.74 (2.19)	4.32 (0.15)
GPR + Horseshoe	46.17 (1.68)	3.91 (0.14)	41.93 (2.22)	4.29 (0.15)
BR + Normal	**46.82** (1.60)	**3.86** (0.14)	42.46 (2.06)	4.25 (0.15)
BR + Horseshoe	45.95 (1.70)	3.92 (0.14)	**43.85** (1.89)	**4.15** (0.15)
Ridge	46.81 (1.60)	**3.86** (0.14)	42.46 (2.07)	4.25 (0.15)
LASSO	45.76 (1.53)	3.95 (0.13)	43.65 (1.86)	4.17 (0.16)

Table 4.2: Evaluation of prediction results with R^2 (in%) and predictive MSE (calculated for the training and test set) mean (standard error) for a simulation study setting with varying sample sizes $N = \{500; 1,000; 2,000\}$ for our model RTGP-VI and the baseline models GPR + Normal, GPR + Horseshoe, BR + Normal, BR + Horseshoe, Ridge and LASSO. The MSE values are multiplied by a scaling factor of 10^2 for clarity.

Thirdly, the evaluation of inference results can only be performed for a comparison of RTGP to the baseline models that involve a Gaussian process regression (GPR + Normal, GPR + Horseshoe). For the other approaches, significance can either not be obtained as in the case of the frequentist approach (Ridge), significance is too speckled (LASSO), or no significant result is observed for the Bayesian regression models which place a Normal or Horseshoe prior over the image coefficients (BR +

Normal, BR + Horseshoe). Table 4.3 shows that GPR + Normal / Horseshoe has higher sensitivity across all sample sizes than RTGP, e.g. with TPR = 87.00 for GPR + Horseshoe for N=500, but suffers from identifying many false positive clusters, see Figure 4.4, leading to low specificity, e.g. with FPR = 13.93. On the other hand, our RTGP has only comparable sensitivity results but otherwise outperforms the baseline Gaussian process regressions with respect to TDR, FPR, and FDR for all sample sizes. For example, in the case of sample size N=500 the sensitivity of RTGP is TPR = 75.22 but the specificity is FPR = 6.23 where the results are even more pronounced for increasing sample sizes. Figure 4.4 support the results described above where we can observe accurate spatially localised activation compared to the truth for RTGP whereas both GPR + Normal and GPR + Horseshoe are able to localise the major clusters of significance but significance is determined far beyond the borders of the true cluster of significance leading to false positive activation. Additionally, GPR + Horseshoe suffers from false positive activation across the cortex and seems to identify far more clusters as significant than are actually present in the true significance map.

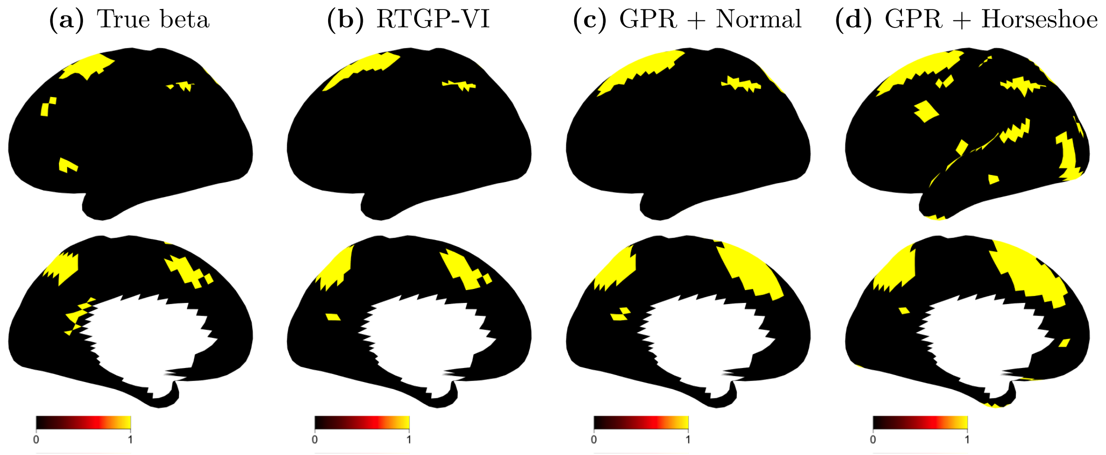


Figure 4.4: Comparison of the (a) true binary significance map with the estimated parameter maps from (b) our RTGP-VI model and (c)-(d) the baseline Gaussian process regressions with Normal and Horseshoe priors on the basis coefficients simulation study setting: N = 500). We omit the significance maps of the other maps because they either show now significance across the cortex (BR + Normal, BR + Horseshoe), significance cannot be obtained (Ridge), or significance is too speckled (LASSO).

N=500	**TPR**	**TDR**	**FPR**	**FDR**
RTGP-VI	75.22 (7.03)	**56.85** (17.96)	**6.23** (5.78)	**44.15** (17.96)
GPR + Normal	**89.00** (2.98)	34.57 (8.68)	9.80 (4.22)	65.43 (8.68)
GPR + Horseshoe	87.00 (3.40)	26.98 (7.69)	13.93 (6.35)	73.02 (7.69)
N=1,000	**TPR**	**TDR**	**FPR**	**FDR**
RTGP-VI	79.56 (5.81)	**60.72** (8.05)	**2.86** (1.08)	**39.28** (8.05)
GPR + Normal	91.11 (2.51)	34.56 (4.21)	9.27 (1.67)	65.44 (4.21)
GPR + Horseshoe	**92.22** (2.51)	23.30 (2.41)	16.19 (2.28)	76.70 (2.41)
N=2,000	**TPR**	**TDR**	**FPR**	**FDR**
RTGP-VI	79.33 (3.19)	**66.70** (3.20)	**2.10** (0.33)	**33.30** (3.20)
GPR + Normal	92.11 (1.33)	35.12 (3.61)	9.10 (1.45)	64.88 (3.61)
GPR + Horseshoe	**93.44** (0.97)	15.25 (2.55)	28.37 (6.89)	84.75 (2.55)

Table 4.3: Evaluation of inference results with TPR, TDR, FPR, and FDR mean (standard error) for a simulation study setting with varying sample sizes $N = \{500; 1,000; 2,000\}$ for our model RTGP-VI and the baseline models GPR + Normal and GPR + Horseshoe. We omit the inference comparisons of the other models because they either show now significance across the cortex (BR + Normal, BR + Horseshoe), significance cannot be obtained (Ridge), or significance is too speckled (LASSO). The values are multiplied by a scaling factor of 10^2 for clarity.

4.3.2 ABCD Study Application

In the real data application on the ABCD study, we now showcase how our model RTGP behaves in comparison to the baseline approaches introduced in the previous section. We highlight how our model scales to far more complex problems with larger sample sizes and number of image locations than previously exposed to in the simulation study setting.

4.3.2.1 Dataset and Preprocessing

In this work, we are utilising the emotional n-back task-based functional MRI data (Casey et al., 2018) as the input image $\boldsymbol{x}_i \in \mathbb{R}^M$ in addition to the total composite score from the NIH toolbox cognition battery (Luciana et al., 2018) as the scalar output y_i to perform an scalar-on-image regression on real data. Specifically, we are utilising the test statistics acquired in a first-level fMRI analysis of the emotional n-back task with the 2- vs. 0-back contrast for each subject and each vertex, see Section 1.1.1.2 for an overview of first-level fMRI analysis, which outputs a test statistic map for each subject for the group scalar-on-image regression we are

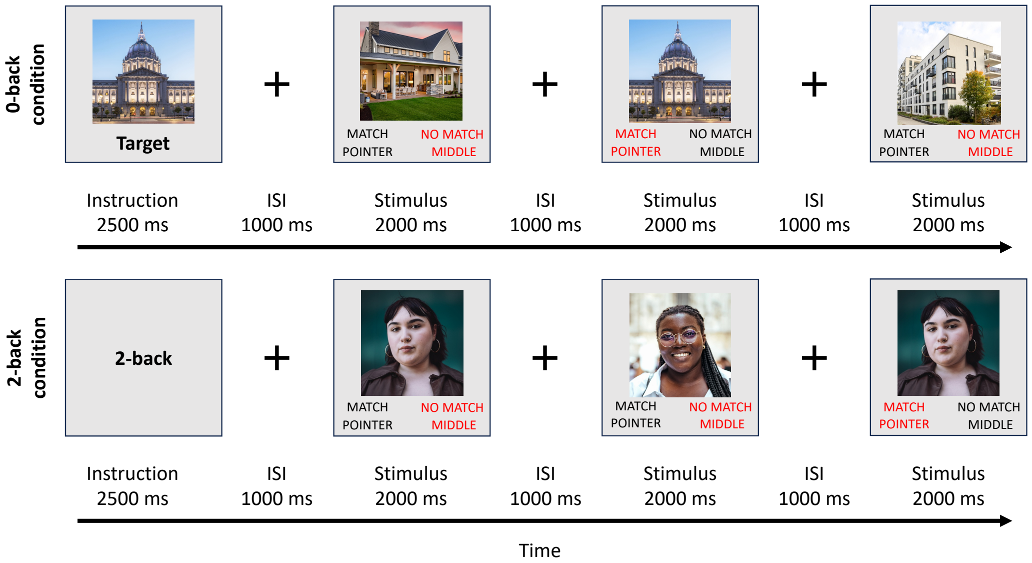


Figure 4.5: Emotional n-back task within the ABCD study with 0-back and 2-back condition including images of faces and locations. (Adapted from Barch et al. (2013) and Casey et al. (2018))

aiming to perform. The ABCD study contains the following three functional MRI tasks: the monetary incentive delay task, the stop signal task, and the emotional n-back task. Due to recent reports by Makowski et al. (2023) we use the data from the emotional n-back task with the 2- vs. 0-back contrast as the input of our regression as it contains the strongest signal for detecting any association between cognition and imaging modalities, see Makowski et al. (2023) for an overview of mass-univariate analyses between intelligence scores and various MRI modalities and Appendix Section C.6.1 and C.4 for showcasing the signal strength of the emotional n-back 2- vs 0-back contrast.

The emotional n-back task is a recognition memory test which contains a low and high memory load measure via the 0-back 2-back condition respectively. The former is capable of assessing working memory and the latter allows for examining the cognitive abilities of the participants. The task includes 48 old and 48 new stimuli with equal numbers of each stimuli type in the old and new stimulus set. Moreover, each set contains happy, fearful, and neutral facial expressions as well as places. Each task includes two runs consisting of eight blocks each. The 0-back condition shows a target image where the participants are then asked to identify at each consecutively shown image if it matches the target. The 2-back condition on the other hand asks the participants to respond with “match” if the current stimulus is identical to the stimulus shown two trials back. The exact timings of the emotional n-back task for both of the conditions is explained in Figure 4.5. For further detail on the imaging setup as well as any preprocessing performed on the data we refer to the work performed by Casey et al. (2018), Hagler et al. (2019) and Feczko et al. (2021).

In the following real data analysis, we utilise the population subset of the ABCD study which has the emotional n-back task fMRI data available without any missing values alongside several confounding variables, such as sex, age, race / ethnicity, parents’ educational level, family income, parents’ marital status, and imaging site ID, listed in Table 4.4. We also split our dataset in training, validation and test data with a ratio of 8/1/1 and absolute sample sizes of $N_{train} = 3,136$, $N_{val} = 517$, and $N_{test} = 517$. The validation dataset is used for the RTGP-VI and GPR models in order to determine the kernel hyperparameters of the GP, such as the bandwidth and kernel exponent parameter, by evaluating the ELBO and the out-of-sample prediction performance.

Sex	51.74% (female) / 48.26% (male)
Age	119.80 (7.41) (in months)
Race / Ethnicity	81.16% (White) / 15.80% (Black) / 6.41% (Asian) / 17.79% (Hispanic)
Parents’ educational level	8.56% (High school or GED) / 14.89% (some college) / 31.39% (Bachelor degree) / 28.21% (postgraduate degree)
Family income	29.93% (lower than 50k) / 27.38% (between 50k and 100k) / 42.69% (more than 100k)
Marital status	73.84% (married) / 26.16% (single / divorced / widowed / living with partner)
Intelligence score	103.43 (16.99)

Table 4.4: Characteristics of ABCD emotional n-back task fMRI population.

4.3.2.2 Results Interpretation

The real data application on the ABCD study reveals similar trends as the simulation studies for the parameter estimation of the spatially varying regression coefficients in the scalar-on-image regression associating composite intelligence scores with the emotional n-back task fMRI test statistic values while accounting for confounding variables. Figure 4.6 shows that the Bayesian models that directly place a prior on the image coefficients, BR + Normal and BR + Horseshoe, or their frequentist counterparts, Ridge and LASSO regression, do not account for any spatial dependence in the model specification and therefore yield speckled results. Moreover, BR + Horseshoe and LASSO models do exhibit a certain level of shrinkage behaviour; however, neither models are able to recover any spatial clusters and exhibit speckled, extreme parameter estimation across the cortex. The GPR + Normal model exhibits more spatially smooth parameter estimates but does not seem to be sensitive enough to pick up any signal. For the other Gaussian process regression with a Horseshoe prior on the basis coefficients the effect size map shown in Figure 4.6(d) depicts smooth parameter estimates across the cortex but exhibits a far smaller ranger of estimated parameter values than the other approaches with values between -0.004 and 0.004 which suggests that the global variance parameter of the Horseshoe prior biases the parameter estimates that should have a larger signal to a smaller effect size. On the other hand, our model RTGP showcases the benefit of thresholding the Gaussian process prior which is placed over the spatially varying image coefficients where the image coefficients drawn from a GP prior are thresholded to 0 if the latent variable placed at each vertex does not surpass a estimated threshold. The benefit being that parameter estimation and inference occurs separately. Hence, the parameter map in Figure 4.6(b) of our RTGP model does not suffer from the same problem as the GPR + Horseshoe model in Figure 4.6(d) where the global variance has the power to bias all coefficients in favour of stronger shrinkage.

For inference, we determine binary significance in the same manner as described in Section 4.3 where we threshold the expected value $\mathbb{E}[I(|\alpha(s_j)| > \delta)]$ for every vertex $j = 1, \dots, M$ with 0.5 for the RTGP model and in the GPR + Normal and GPR + Horseshoe model we determine whether or not 0 is included in the 95%-HPDI interval. We acknowledge that presently there is no principled way to find a significance threshold that corresponds to the same level of sensitivity versus specificity trade-off across models; however, we show that for the RTGP model varying the significance threshold yields comparable inference results and a similar level of out-of-sample predictive MSE across the thresholds, see Appendix Section C.6.5. We also find that varying the size of the HPDI for the GPR baseline models does not drastically change the significance results. Figure 4.7 suggests that our RTGP model potentially produces more sensitive results than the GPR + Horseshoe baseline approach which yields significant results at the same vertices but has far less spatial activation than the RTGP approach, see Figure 4.7(d) where we plot the overlap between the two binary significance maps and display the RTGP activation in red and any overlap between the approaches in yellow. Our RTGP model finds that 2,310 vertices are significant, the GPR + Normal model results in only 9 activations and the GPR + Horseshoe approach yields 247 significant vertices.

Unfortunately, the problem of associating intelligence scores with the emotional n-back task fMRI has only been studied as an image-on-scalar problem which does not provide us with an indication of where we might expect true significance. We therefore can only suggest that our RTGP model has the potential to uncover significance of smaller effects yielding more sensitive results in data with complex spatial structures whereas the baseline approach GPR + Horseshoe provides potentially more conservative results in this real data application. Hence, the researcher can use either approach for prediction but should carefully evaluate whether or not they desire small areas of strong effect as shown by GPR + Horseshoe or potentially also consider subtler effects that can be relevant towards the prediction as additionally discovered by RTGP. In the case of RTGP, approximately the same areas of the brain in the dorsolateral and medial prefrontal cortices and precuneus bilaterally were found to have strong positive associations relevant towards the prediction of intelligence scores as the image-on-scalar regression studying the reverse problem (Makowski et al., 2023). In Appendix Section C.6.6, we show that this result is robust to removing 10% of the training data and performing 10-fold cross-validation on various splits of the training data. Lastly, Table 4.5 also shows that the

Left Hemisphere	R^2 (train)	MSE (train)	R^2 (test)	MSE (test)
Ridge	66.66	110.9729	18.33	252.4871
LASSO	35.45	188.6745	28.34	216.8615
BR + Normal	**100.00**	**0.0201**	4.42	532.4117
BR + Horseshoe	49.95	149.9042	27.48	218.8589
GPR + Normal	26.25	222.9945	16.01	286.7683
GPR + Horseshoe	31.09	198.3787	**29.97**	**211.6691**
RTGP-VI	27.87	208.9946	29.96	211.8090
Right Hemisphere	R^2 (train)	MSE (train)	R^2 (test)	MSE (test)
Ridge	66.74	109.7410	19.35	246.4008
LASSO	36.18	186.8528	28.43	216.6652
BR + Normal	**100.00**	**0.0067**	2.63	526.8691
BR + Horseshoe	52.00	144.3401	24.88	227.0925
GPR + Normal	26.74	222.8687	15.18	292.7009
GPR + Horseshoe	31.59	196.8422	28.39	215.9624
RTGP-VI	29.91	203.1902	**29.27**	**215.5252**

Table 4.5: Comparison of prediction results of ABCD study analysis evaluated with training and test R^2 (in %) and predictive MSE for the left and the right hemisphere of the brain. The models which do not take any spatial dependence between the vertices into account exhibit overfitting and perform poorly in the test dataset. The GPR + Horseshoe and RTGP-VI possess on par predictive performance.

only two models that have reasonable inference results in addition to low predictive MSE and high R^2 are GPR + Horseshoe and our suggested approach RTGP.

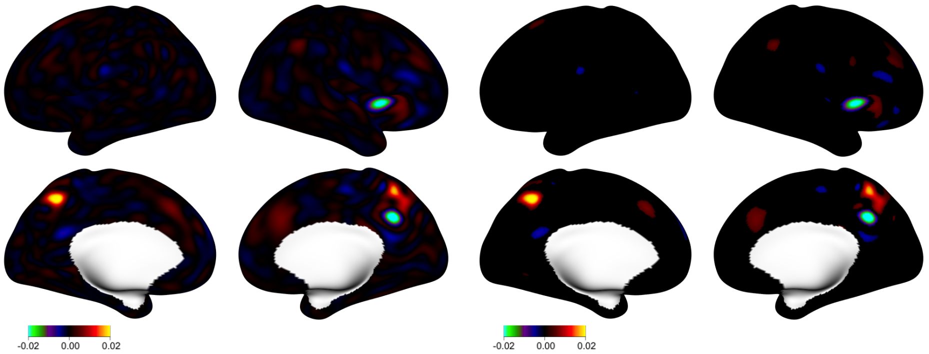
4.4 Discussion and Future Work

Within this work we have proposed a Bayesian nonparameteric scalar-on-image regression model with a relaxed-thresholded Gaussian process prior. Our contributions are three-fold: 1) we introduce a more flexible thresholding prior that is able to perform hard- and soft-thresholding of spatially varying coefficients alike, 2) we increase the scalability of our scalar-on-image regression models with a thresholded GP prior to large-scale applications by using a variational approximation to the posterior instead of MCMC sampling, and 3) to the best of our knowledge the Karhunen-Lòeve expansion has not been used on kernels with a correlation function that measures distance on the surface.

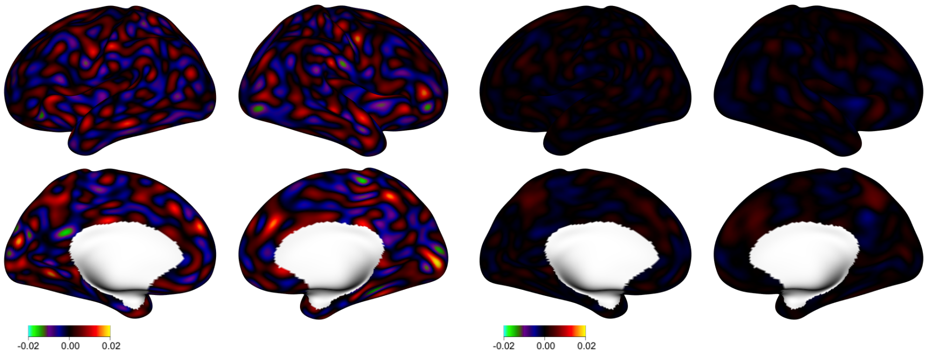
We do note that our model relies on an approximation of GPs, specifically the Karhunen-Lòeve expansion, which is reliant on determining a number of basis functions that is far lower than the number of vertices of the image. The goal of the approximation is to yield an adequate approximation while being computationally much faster. We determine this number by satisfying a certain percentage of total variation, see Equation 4.12 and Appendix Section C.15, and find that capturing only 30% of the total variation yields adequate results. We do not claim that this holds for other applications and hence our VI approach would suffer from the same computational bottleneck as in MCMC sampling, specifically the inversion of the precision matrix, if the number of basis functions is increased to capture a value closer to 100% of the total variation.

Regarding future work, we are interested in extending the current scalar-on-image regression problem to multiple data modalities, such as structural, diffusion-weighted or resting-state functional MRI. Current approaches often summarise the results of different image modalities (Kang et al., 2018; Guo et al., 2022); however, introducing smoothness and sparsity for each modality separately would require a separate thresholded GP prior with its own thresholding parameter. The advantage of introducing a separate prior for each spatially varying coefficient set is that the image dimension would not be required to match and hence, enabling the combination of various image modalities as input sources. Furthermore, introducing multiple image modalities poses the question of how interaction effects impact the outcome of interest.

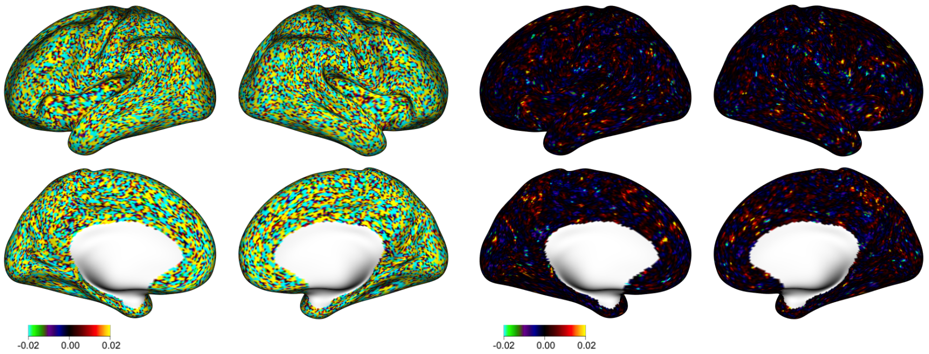
As another area of future work, we are interested in exploring other potential kernels for the covariance function of the GP prior on the spatially varying coefficients. Our simulation studies and real data application revealed that the GPR + Horseshoe model has a lower predictive MSE than RTGP in both scenarios. An interesting area of future work would hence be to test if a Horseshoe kernel can be used in lieu of the two parameter exponential radial basis function that we currently employ and achieve better performance with respect to prediction in addition to inference while still retaining a scalable inference algorithm through variational inference.



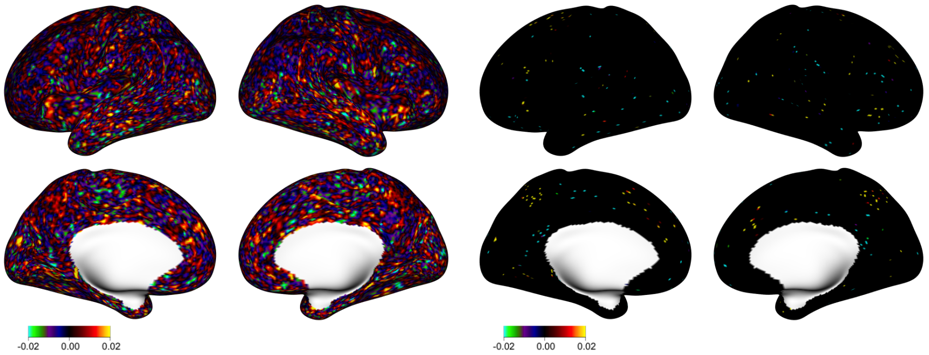
(a) RTGP-VI (unthresholded) **(b)** RTGP-VI (thresholded)



(c) GPR + Normal **(d)** GPR + Horseshoe

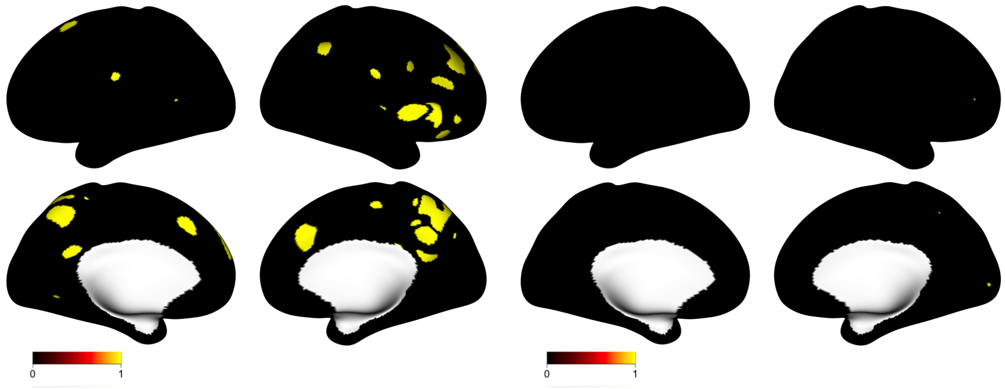


(e) BR + Normal **(f)** BR + Horseshoe



(g) Ridge **(h)** LASSO

Figure 4.6: Comparison of the estimated parameter maps from (a)-(b) our RTGP-VI model (unthresholded and thresholded) and (c)-(h) the other baseline approaches. Note that the we limit the colourbar to the range of the RTGP-VI parameter values which lie between -0.02 to 0.02 for a fair comparison of models. However, the (f) BR + Horseshoe model and (h) the frequentist LASSO regression both exhibit far more extreme values which lie outside the plotted range.



(a) RTGP-VI

(b) GPR + Normal

(c) GPR + Horseshoe

(d) Overlap: RTGP & GPR + Horseshoe

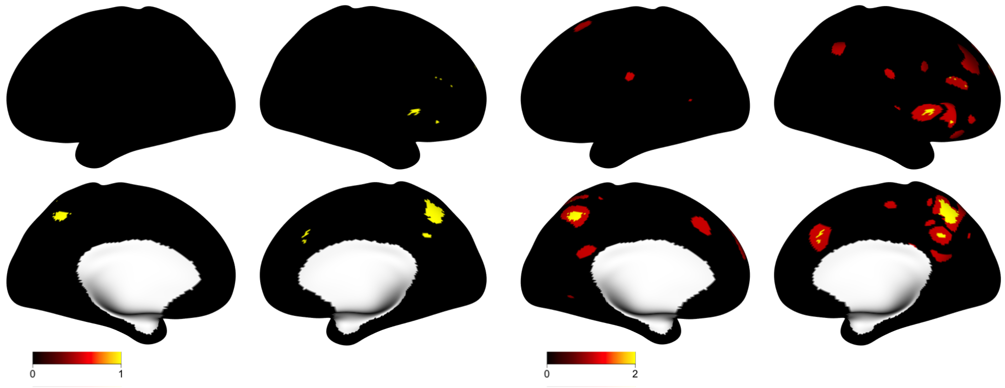


Figure 4.7: Comparison of the estimated binary significance maps from (a) our RTGP-VI model to the other baseline approaches, (b) GPR + Normal and (c) GPR + Horseshoe, and (d) overlapping the significance results of the RTGP-VI model (red) and GPR + Horseshoe model (yellow). Note that the other models are omitted as significance can either not be determined (Ridge), significance is too speckled (LASSO), or no vertex is significant in the estimated map (BR + Normal, BR + Horseshoe).

5

Final Remarks

Contents

5.1 Conclusions

Each contribution of this thesis has been inspired by the work of many other researchers and builds on their work. For instance, the model BLESS from Chapter 2 is derived based on the Bayesian spatial model by Ge et al. (2014), the extension to account for uncertainty quantification in Chapter 3 is based on the work by Nie and Ročková (2022), and lastly my contribution to thresholded Gaussian processes started with the work by Kang et al. (2018) on soft-thresholded Gaussian process priors. Overall, the aim of this thesis is to contribute to the development of statistical models for neuroimaging applications beyond the advancements made by the other researchers listed above by accounting for the spatial dependence between neighbouring locations within the brain, utilising Bayesian variable selection to drive negligible effect locations to zero, and performing parameter estimation, inference, and prediction through scalable algorithms that scale to large sample sizes and number of coefficients.

In Chapter 2, I have developed a Bayesian spatial image-on-scalar regression model, called BLESS, where the dependent variables are lesion masks, which are binary-valued 3D images, and the independent variables are scalar quantities of interest, such as age. The main contribution of this work is two-fold: The development of 1) structured spike-and-slab priors on spatially varying coefficients, specifically defined by independent mixture of normal distributions with different variances and a MCAR prior on the sparsity parameters within the inclusion probabilities of the indicator variables to induce spatial structure, and 2) a scalable inference algorithm through a variational approximation to the posterior in combination with dynamic posterior exploration, an annealing-like strategy. Our UK Biobank application of associating age with lesion incidence demonstrated the scalability of our image-on-scalar method to sample sizes with thousands of subjects and a large number of parameters.

In Chapter 3, I extend the work on BLESS to provide accurate uncertainty quantification, a property previously lost through the usage of variational approximations which underestimate the posterior variance, and to provide cluster size-based imaging statistics, a quantity often preferred by neuroimaging statisticians for the evaluation of inference. We propose an approximate posterior sampling technique, derived by re-weighting the likelihood function through Dirichlet weights, mimicking the WLB, and perturbing the prior mean of the spike-and-slab distribution on the spatially varying coefficients. Our approach retains the computational scalability of the simple VI algorithm used for BLESS-VI in Chapter 2 by estimating each bootstrap iteration in an embarrassingly parallel manner with a variational approximation to the posterior. Besides allowing for better uncertainty quantification, we found that the approximate posterior sampling algorithm naturally lends itself to the derivation of cluster size-based imaging statistics where we provide credible intervals of cluster size and prevalence maps measuring the reliability of spatial cluster occurrence across the brain.

In Chapter 4, I switch from the development of statistical methods for volumetric-based representations of MRI scans to advancing Bayesian spatial models for cortical surface-based data representations. Analysing MRI on the surface has only increased in popularity across the recent years due to better analysis, preprocessing and display tools becoming available, see Mejia et al. (2020). In Section 1.1.3, I also provide an overview on the benefits of cortical surface-based representations with respect to Bayesian spatial modelling applications. Hence, my last contribution in this thesis is the development of a general class of thresholded Gaussian process priors, that encompass both hard- and soft-thresholding properties, for a Bayesian scalar-on-image regression problem. Again providing a variational inference algorithm in combination with the Karhunen-Loève expansion for GPs for scalable posterior estimation and inference. We also apply our work on a large-scale imaging study, the ABCD study, where we identified spatial regions across the surface of the brain in the emotional n-back task that we determine as relevant towards making a prediction of intelligence scores in a cohort of children aged between 9 to 10 years old.

5.2 Future Directions

My thesis increases the scalability of posterior parameter estimation and inference for some Bayesian spatial models. As a future direction, applying BLESS to other image-on-scalar regression problems or RTGP to further scalar-on-image regression problems in large-scale population health-based studies has the potential to uncover smaller and subtler effects of interest across the brain. For example, BLESS can be used to identify associations between cerebrovascular risk factors and lesion incidence in order to explain their subtle impact on disease development and therefore the cognitive impact of cerebrovascular risk-related white matter lesions (Veldsman et al., 2020). Another interesting direction for the application of BLESS is the Novartis-Oxford multiple sclerosis dataset (Dahlke et al., 2021) which contains lesions masks for approximately 8,000 subjects with corresponding subject-specific data, such as age, sex, disease severity scores, and multiple sclerosis sub-type. For example, BLESS can provide insights on where in the brain a certain disease severity is associated with lesion incidence.

While I find empirically that BB-BLESS is able to quantify uncertainty better than its sole variational counterpart BLESS-VI, I have not derived any theoretical justifications for using variational inference in lieu of MAP estimation for Bayesian bootstrap methods, using non-separable mixture of normal distribution formulation for the spike-and-slab prior on the coefficients in lieu of the SS-LASSO priors used in the work by Nie and Ročková (2022); further, we do not provide a theoretical guarantee for the appropriate selection of a jittering distribution to draw mean perturbations from, specifically we do not justify why the spike variance determined by DPE as the variance in the spike distribution is the most appropriate jittering distribution for re-weighting the prior in the pseudo-posterior of BB-BLESS. However, we pose these questions on theoretical guarantees as an interesting question for future research.

The suggestions for extending our work on RTGP priors are two-fold. Firstly, we found that the GPR + Horseshoe model which originally was formulated as a baseline model, performs surprisingly well with respect to prediction. Extending the covariance kernel function of the GP to a Horseshoe kernel in lieu of a radial basis kernel can potentially increase the predictive performance of RTGP priors. Secondly, we use only one data modality within our simulation studies and real data application. It poses therefore an interesting future direction to incorporate other data modalities, such as structural or diffusion-weighted MRI, into our model and study its interaction effects between such highly correlated data sources.

Lastly, I want to encourage the development of toolboxes that can be easily applied by clinicians and researchers with a non-statistical background. While I do provide the code for all methods on my Github[1] profile, I would like to provide easier to use packages for researchers in `R` or `Python` to enable the usage of complex Bayesian spatial models for a wider audience. Moreover, I would like to provide

---

[1] `https://github.com/annamenacher`

faster model estimation by implementing parts of my code in `C++` through using the R package "`Rcpp`" which integrates faster to execute `C++` code into `R`.

Appendices

A

Appendix: Scalable Image-on-Scalar Regression with a Structured Spike-and-Slab Prior

Contents

A.1 Derivation of Variational Inference Algorithm

A.1.1 Joint Distribution

$$
\begin{aligned}
Q &= p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\Sigma}) \\
&= p(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{z}|\boldsymbol{\beta}, \boldsymbol{\beta}_0)p(\boldsymbol{\beta}_0)p(\boldsymbol{\beta}|\boldsymbol{\gamma})p(\boldsymbol{\gamma}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\Sigma})p(\boldsymbol{\Sigma}^{-1}) \\
&= \prod_{i=1}^{N}\prod_{j=1}^{M} \Pr(y_i(s_j)|z_i(s_j)) \\
&\times \prod_{i=1}^{N}\prod_{j=1}^{M} \mathcal{N}(z_i(s_j); \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(s_j) + \beta_0(s_j), 1) \\
&\times \prod_{j=1}^{M} \mathcal{N}(\beta_0(s_j); \mu_{\beta_0}, \sigma_{\beta_0}^2) \\
&\times \prod_{j=1}^{M} \mathcal{N}(\boldsymbol{\beta}(s_j); \boldsymbol{0}, \mathrm{diag}\{\nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j)\}_{p=1}^{P}) \\
&\times \prod_{p=1}^{P}\prod_{j=1}^{M} Bernoulli(\gamma_p(s_j); \sigma(\theta_p(s_j))) \\
&\times \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{0}, \boldsymbol{\Sigma}(\boldsymbol{D}-\boldsymbol{W})^{-1}) \\
&\times Wishart(\boldsymbol{\Sigma}^{-1}; P, \boldsymbol{I})
\end{aligned}
$$

$$
\begin{aligned}
Q^* \propto\ & \mathbb{E}_{q(z,\beta,\beta_0,\gamma,\theta,\Sigma^{-1})} \Bigg[-\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{M} (z_i(s_j) - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}(s_j) - \beta_0(s_j))^2 - \frac{1}{2\sigma_{\beta_0}^2} \sum_{j=1}^{M} (\beta_0(s_j) - \mu_{\beta_0})^2 \\
& - \frac{1}{2} \sum_{j=1}^{M} \log(|\mathrm{diag}\{\nu_0(1-\gamma_p(s_j)) + \nu_1 \gamma_p(s_j)\}_{p=1}^{P}|) \\
& - \frac{1}{2} \sum_{j=1}^{M} \boldsymbol{\beta}(s_j)^{\mathrm{T}} \mathrm{diag}\{\nu_0(1-\gamma_p(s_j)) + \nu_1 \gamma_p(s_j)\}_{p=1}^{P} \boldsymbol{\beta}(s_j) \\
& + \sum_{p=1}^{P} \sum_{j=1}^{M} \gamma_p(s_j) \log(\sigma(\theta_p(s_j))) + \sum_{p=1}^{P} \sum_{j=1}^{M} (1 - \gamma_p(s_j) \log(1 - \sigma(\theta_p(s_j)))) \\
& + \frac{1}{2} \sum_{j=1}^{M} \log \left\{ |\boldsymbol{\Sigma}^{-1}| \right\} - \frac{1}{2} \sum_{s_i \sim s_j} \left[\boldsymbol{\theta}(s_i) - \boldsymbol{\theta}(s_j) \right]^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \left[\boldsymbol{\theta}(s_i) - \boldsymbol{\theta}(s_j) \right] \\
& + \frac{1}{2} \left\{ (\nu - P - 1) \log\{|\boldsymbol{\Sigma}^{-1}|\} - \mathrm{tr}(\boldsymbol{\Sigma}^{-1}) \right\} \Bigg]. \\
\geq\ & \mathbb{E}_{q(z,\beta,\beta_0,\gamma,\theta,\Sigma^{-1})} \Bigg[-\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{M} (z_i(s_j) - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}(s_j) - \beta_0(s_j))^2 - \frac{1}{2\sigma_{\beta_0}^2} \sum_{j=1}^{M} (\beta_0(s_j) - \mu_{\beta_0})^2 \\
& - \frac{1}{2} \sum_{j=1}^{M} \log(|\mathrm{diag}\{\nu_0(1-\gamma_p(s_j)) + \nu_1 \gamma_p(s_j)\}_{p=1}^{P}|) \\
& - \frac{1}{2} \sum_{j=1}^{M} \boldsymbol{\beta}(s_j)^{\mathrm{T}} \mathrm{diag}\{\nu_0(1-\gamma_p(s_j)) + \nu_1 \gamma_p(s_j)\}_{p=1}^{P} \boldsymbol{\beta}(s_j) \\
& + \sum_{p=1}^{P} \sum_{j=1}^{J} \left[\log(\sigma(\xi_p(s_j))) + \theta_p(s_j)\gamma_p(s_j) - \frac{(\theta_p(s_j) + \xi_p(s_j))}{2} - \lambda(\xi_p(s_j))(\theta_p(s_j)^2 - \xi_p(s_j)^2) \right] \\
& + \frac{1}{2} \sum_{j=1}^{M} \log \left\{ |\boldsymbol{\Sigma}^{-1}| \right\} - \frac{1}{2} \sum_{s_i \sim s_j} \left[\boldsymbol{\theta}(s_i) - \boldsymbol{\theta}(s_j) \right]^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \left[\boldsymbol{\theta}(s_i) - \boldsymbol{\theta}(s_j) \right] \\
& + \frac{1}{2} \left\{ (\nu - P - 1) \log\{|\boldsymbol{\Sigma}^{-1}|\} - \mathrm{tr}(\boldsymbol{\Sigma}^{-1}) \right\} \Bigg].
\end{aligned}
$$

A.1.2 Variational Approximations

A.1.2.1 Update $z_i(s_j)$

Assume $y_i(s_j) = 1$ and $\boldsymbol{\eta}(s_j) = \boldsymbol{x}_i^T \mathbb{E}_{q(\beta)}[\boldsymbol{\beta}(s_j)] + \mathbb{E}_{q(\beta_0)}[\beta_0(s_j)]$:

$$
\begin{aligned}
\ln(q^*(z_i(s_j))) &\propto \mathbb{E}_{q(\beta,\beta_0,\gamma,\theta,\Sigma)}[\ln\{p(y_i(s_j), z_i(s_j), \boldsymbol{\beta}(s_j), \beta_0(s_j), \boldsymbol{\gamma}(s_j), \boldsymbol{\theta}(s_j), \boldsymbol{\Sigma}|\boldsymbol{x}_i)\}] \\
&\propto \ln\{p(y_i(s_j)|z_i(s_j))\} + \mathbb{E}_{q(\beta,\beta_0)}\left[\ln\{p(z_i(s_j)|\boldsymbol{x}_i, \boldsymbol{\beta}(s_j), \beta_0(s_j))\}\right] \\
&\propto y_i(s_j)\ln\{\mathbb{1}(z_i(s_j) > 0)\} + (1 - y_i(s_j))\ln\{\mathbb{1}(z_i(s_j) \leq 0)\} \\
&- \frac{1}{2}\mathbb{E}_{q(\beta,\beta_0)}[(z_i(s_j) - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(s_j) - \beta_0(s_j))^2] \\
&\propto \ln\{\mathbb{1}(z_i(s_j) > 0)\} - \frac{1}{2}z_i(s_j)^2 + z_i(s_j)[\mathbb{E}_{q(\beta)}[\boldsymbol{\beta}(s_j)]^{\mathrm{T}}\boldsymbol{x}_i + \mathbb{E}_{q(\beta_0)}[\beta_0(s_j)]] \\
&\propto \ln\{\mathbb{1}(z_i(s_j) > 0)\} - \frac{1}{2}z_i(s_j)^2 + z_i(s_j)\eta_i(s_j)
\end{aligned}
$$

$$
q^*(z_i(s_j)) = \begin{cases} \mathcal{TN}_+(z_i(s_j); \eta_i(s_j), 1), & \text{if } y_i(s_j) = 1, \\ \mathcal{TN}_-(z_i(s_j); \eta_i(s_j), 1), & \text{if } y_i(s_j) = 0. \end{cases}
$$

A.1.2.2 Update $\beta(s_j)$

$$
\begin{aligned}
\ln(q^*(\boldsymbol{\beta}(s_j))) &\propto \mathbb{E}_{q(z,\beta_0,\gamma,\theta,\Sigma)}[\ln\{p(y_i(s_j), z_i(s_j), \boldsymbol{\beta}(s_j), \beta_0(s_j), \boldsymbol{\gamma}(s_j), \boldsymbol{\theta}(s_j), \boldsymbol{\Sigma}|\boldsymbol{x}_i)\}] \\
&\propto \mathbb{E}_{q(\gamma)}[\ln\{p(\boldsymbol{\beta}(s_j)|\boldsymbol{\gamma}(s_j))\}] + \mathbb{E}_{q(z,\beta_0)}[\ln\{p(\boldsymbol{z}(s_j)|X, \boldsymbol{\beta}(s_j), \beta_0(s_j))\}] \\
&\propto -\frac{1}{2}\boldsymbol{\beta}(s_j)^{\mathrm{T}}\mathbb{E}_{q(\gamma)}[\mathrm{diag}\{\nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j)\}_{p=1}^{P}]^{-1}\boldsymbol{\beta}(s_j) \\
&- \frac{1}{2}\mathbb{E}_{q(z,\beta_0)}\left[[\boldsymbol{z}(s_j) - X\boldsymbol{\beta}(s_j) - \beta_0(s_j)]^{\mathrm{T}}[\boldsymbol{z}(s_j) - X\boldsymbol{\beta}(s_j) - \beta_0(s_j)]\right] \\
&\propto \mathbb{E}_{q(z,\beta_0)}\left[\boldsymbol{\beta}(s_j)^T X^T[\boldsymbol{z}(s_j) - \beta_0(s_j)] - \frac{1}{2}\boldsymbol{\beta}(s_j)^T X^T X\boldsymbol{\beta}(s_j)\right] \\
&- \frac{1}{2}\mathrm{tr}(\mathbb{E}_{q(\gamma)}[\mathrm{diag}\{\nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j)\}_{p=1}^{P}]^{-1}\boldsymbol{\beta}(s_j)\boldsymbol{\beta}(s_j)^T) \\
&\propto \boldsymbol{\beta}(s_j)^T X^T[\mathbb{E}_{q(z}[\boldsymbol{z}(s_j)] - \mathbb{E}_{q(\beta_0)}[\beta_0(s_j)]] \\
&- \frac{1}{2}\mathrm{tr}([X^T X + \mathbb{E}_{q(\gamma)}[\mathrm{diag}\{\nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j)\}_{p=1}^{P}]^{-1}]\boldsymbol{\beta}(s_j)\boldsymbol{\beta}(s_j)^T)
\end{aligned}
$$

$$
\begin{aligned}
q^*(\boldsymbol{\beta}(s_j)) &= \mathcal{N}(\boldsymbol{\beta}(s_j); \mu_{\beta(s_j)}, \Sigma_{\beta(s_j)}) \\
\mu_{\beta(s_j)} &= \Sigma_{\beta(s_j)} X^{\mathrm{T}}\left[\mathbb{E}_{q(z)}[\boldsymbol{z}(s_j)] - \mathbb{E}_{q(\beta_0)}[\beta_0(s_j)]\right] \\
\Sigma_{\beta(s_j)} &= \left[X^{\mathrm{T}}X + \mathbb{E}_{q(\gamma)}[\mathrm{diag}\{\nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j)\}_{p=1}^{P}]^{-1}\right]^{-1}
\end{aligned}
$$

A.1.2.3 Update $\beta_0(s_j)$

$$\begin{aligned}
\ln(q^*(\beta_0(s_j))) &\propto \mathbb{E}_{q(z,\beta,\gamma,\theta,\Sigma)}[\ln\{p(y_i(s_j), z_i(s_j), \boldsymbol{\beta}(s_j), \beta_0(s_j), \boldsymbol{\gamma}(s_j), \boldsymbol{\theta}(s_j), \boldsymbol{\Sigma}|\boldsymbol{x}_i)\}] \\
&\propto \mathbb{E}_{q(z,\beta)}[\ln p(\boldsymbol{z}(s_j)|X, \beta_0(s_j), \boldsymbol{\beta}(s_j))] + \ln p(\beta_0(s_j)) \\
&\propto \mathbb{E}_{q(z,\beta)}[\ln \mathcal{N}(\boldsymbol{z}(s_j); X\boldsymbol{\beta}(s_j) + \beta_0(s_j), I)] + \ln \mathcal{N}(\beta_0(s_j); 0, \sigma^2_{\beta_0}) \\
&\propto \sum_{i=1}^{N} \beta_0(s_j)[\mathbb{E}_{q(z)}[z_i(s_j)] - \boldsymbol{x}_i^T \mathbb{E}_{q(\beta)}[\boldsymbol{\beta}(s_j)]] - \frac{N}{2}\beta_0(s_j)^2 - \frac{1}{2\sigma^2_{\beta_0}}\beta_0(s_j)^2
\end{aligned}$$

$$\begin{aligned}
q^*(\beta_0(s_j)) &= \mathcal{N}(\beta_0(s_j); \mu_{\beta_0(s_j)}, \sigma^2_{\beta_0(s_j)}) \\
\mu_{\beta_0(s_j)} &= [N + \frac{1}{\sigma^2_{\beta_0}}]^{-1}[\sum_{i=1}^{N} \mathbb{E}_{q(z)}[z_i(s_j)] - \boldsymbol{x}_i^T \mathbb{E}_{q(\beta)}[\boldsymbol{\beta}(s_j)]] \\
\sigma^2_{\beta_0(s_j)} &= [N + \frac{1}{\sigma^2_{\beta_0}}]^{-1}
\end{aligned}$$

A.1.2.4 Update $\gamma_p(\mathbf{s}_j)$

$$\begin{aligned}
\ln(q^*(\gamma_p(s_j))) &\propto \mathbb{E}_{q(z,\beta,\beta_0,\theta,\Sigma)}[\ln\{p(y_i(s_j), z_i(s_j), \boldsymbol{\beta}(s_j), \beta_0(s_j), \boldsymbol{\gamma}(s_j), \boldsymbol{\theta}(s_j), \boldsymbol{\Sigma}|\boldsymbol{x}_i)\}] \\
&\propto \mathbb{E}_{q(\theta)}[\ln p(\gamma_p(s_j)|\theta_p(s_j))] + \mathbb{E}_{q(\beta)}[\ln p(\beta_p(s_j)|\gamma_p(s_j))] \\
&\propto \mathbb{E}_{q(\beta)}[\ln \mathcal{N}(\beta_p(s_j); 0, \nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j))] \\
&+ \mathbb{E}_{q(\theta)}[\ln Bernoulli(\gamma_p(s_j); \sigma(\theta_p(s_j)))] \\
&\geq \mathbb{E}_{q(\beta)}[-\frac{1}{2}\ln(\nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j)) - \frac{1}{2}\frac{\beta_p(s_j)^2}{\nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j)}] \\
&+ \mathbb{E}_{q(\theta)}[\gamma_p(s_j)[\ln\sigma(\xi_p(s_j)) + \frac{(\theta_p(s_j) - \xi_p(s_j))}{2} - \lambda(\xi_p(s_j))[\theta_p(s_j)^2 - \xi_p(s_j)^2]]] \\
&+ \mathbb{E}_{q(\theta)}[(1-\gamma_p(s_j))[\ln\sigma(\xi_p(s_j)) - \frac{(\theta_p(s_j) + \xi_p(s_j))}{2} \\
&- \lambda(\xi_p(s_j))[\theta_p(s_j)^2 - \xi_p(s_j)^2]]] \\
&= \mathbb{E}_{q(\beta)}[-\frac{1}{2}\ln(\nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j)) - \frac{1}{2}\frac{\beta_p(s_j)^2}{\nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j)}] \\
&+ \mathbb{E}_{q(\theta)}[\gamma_p(s_j)\theta_p(s_j) + \ln\sigma(\xi_p(s_j)) - \frac{(\theta_p(s_j) + \xi_p(s_j))}{2} \\
&- \lambda(\xi_p(s_j))[\theta_p(s_j)^2 - \xi_p(s_j)^2] \\
&\propto \mathbb{E}_{q(\beta)}[-\frac{1}{2}\ln(\nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j)) - \frac{1}{2}\frac{\beta_p(s_j)^2}{\nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j)}] \\
&+ \mathbb{E}_{q(\theta)}[\gamma_p(s_j)\theta_p(s_j)]
\end{aligned}$$

$$\begin{aligned}
q^*(\gamma_p(s_j)) &= Bernoulli(\gamma_p(s_j); \mu_{\gamma_p(s_j)}) \\
\mu_{\gamma_p(s_j)} &= \frac{q(\gamma_p(s_j)=1)}{\sum_{k=\{0,1\}} q(\gamma_p(s_j)=k)} \\
q(\gamma_p(s_j)=1) &= \exp\left\{-\frac{1}{2}\ln\nu_1 - \frac{\mathbb{E}_{q(\beta)}[\beta_p(s_j)^2]}{2\nu_1} + \mathbb{E}_{q(\theta)}[\theta_p(s_j)]\right\} \\
q(\gamma_p(s_j)=0) &= \exp\left\{-\frac{1}{2}\ln\nu_0 - \frac{\mathbb{E}_{q(\beta)}[\beta_p(s_j)^2]}{2\nu_0}\right\}
\end{aligned}$$

A.1.2.5 Update $\theta_p(\mathbf{s}_j)$

$$\begin{aligned}
\ln(q(\boldsymbol{\theta})) &\propto \mathbb{E}_{q(z,\beta,\beta_0,\gamma,\Sigma)}[\ln\{p(y_i(s_j), z_i(s_j), \boldsymbol{\beta}(s_j), \beta_0(s_j), \boldsymbol{\gamma}(s_j), \boldsymbol{\theta}(s_j), \boldsymbol{\Sigma}|\boldsymbol{x}_i)\}] \\
&\propto \mathbb{E}_{q(\gamma)}[\ln p(\boldsymbol{\gamma}(s_j)|\boldsymbol{\theta}(s_j))] + \mathbb{E}_{q(\Sigma^{-1})}[\ln p(\boldsymbol{\theta}|\boldsymbol{\Sigma}^{-1})] \\
&\geq \mathbb{E}_{q(\gamma)}[\sum_{p=1}^{P}\{\theta_p(s_j)\gamma_p(s_j) + \ln\sigma(\xi_p(s_j)) - \frac{(\theta_p(s_j)+\xi_p(s_j))}{2} \\
&- \lambda(\xi_p(s_j))[\theta_p(s_j)^2 - \xi_p(s_j)^2]\}] \\
&- \mathbb{E}_{q(\Sigma^{-1})}[\frac{1}{2}\sum s_i \sim s_j[\boldsymbol{\theta}(s_i)-\boldsymbol{\theta}(s_j)]^T\Sigma^{-1}[\boldsymbol{\theta}(s_i)-\boldsymbol{\theta}(s_j)]] \\
&\propto \sum_{p=1}^{P}\left\{\theta_p(s_j)\mathbb{E}_{q(\gamma)}[\gamma_p(s_j)] - \frac{1}{2}\theta_p(s_j) - \lambda(\xi_p(s_j))\theta_p(s_j)^2\right\} \\
&- \frac{1}{2}tr(\mathbb{E}_{q(\Sigma^{-1})}[\Sigma^{-1}]\sum_{s_i\sim s_j}[\boldsymbol{\theta}(s_i)-\boldsymbol{\theta}(s_j)][\boldsymbol{\theta}(s_i)-\boldsymbol{\theta}(s_j)]^T) \\
&\propto \left[\mathbb{E}_{q(\gamma)}[\boldsymbol{\gamma}(s_j)] - \frac{1}{2} + \mathbb{E}_{q(\Sigma^{-1})}[\Sigma^{-1}]\sum_{s_i\sim s_j}\boldsymbol{\theta}(s_i)\right]^T \boldsymbol{\theta}(s_j) \\
&- \frac{1}{2}\boldsymbol{\theta}(s_j)^T\left[n(s_j)\mathbb{E}_{q(\Sigma^{-1})}[\Sigma^{-1}] + 2diag\{\lambda(\xi_p(s_j)\}_{p=1}^{P}\right]\boldsymbol{\theta}(s_j)
\end{aligned}$$

$$\begin{aligned}
q^*(\boldsymbol{\theta}(s_j)) &= \mathcal{N}(\boldsymbol{\theta}(s_j); \mu_{\theta(s_j)}, \Sigma_{\theta(s_j)}) \\
\mu_{\theta(s_j)} &= \Sigma_{\theta(s_j)}\left[\mathbb{E}_{q(\Sigma^{-1})}[\Sigma^{-1}]\sum_{s_i\sim s_j}\boldsymbol{\theta}(s_i) + \mathbb{E}_{q(\gamma)}[\boldsymbol{\gamma}(s_j)] - \frac{1}{2}\right] \\
\Sigma_{\theta(s_j)} &= \left[n(s_j)\mathbb{E}_{q(\Sigma^{-1})}[\Sigma^{-1}] + 2\text{diag}\{\lambda(\xi_p(s_j))\}_{p=1}^{P}\right]^{-1}
\end{aligned}$$

A.1.2.6 Update Σ^{-1}

$$\begin{aligned}
\ln(q(\boldsymbol{\Sigma}^{-1})) &\propto \mathbb{E}_{q(z,\beta,\beta_0,\gamma,\theta)}[\ln\{p(y_i(s_j), z_i(s_j), \boldsymbol{\beta}(s_j), \beta_0(s_j), \boldsymbol{\gamma}(s_j), \boldsymbol{\theta}(s_j), \boldsymbol{\Sigma}|\boldsymbol{x}_i)\}] \\
&\propto \mathbb{E}_{q(\theta}[\ln p(\boldsymbol{\theta}|\boldsymbol{\Sigma}^{-1})] + \ln p(\boldsymbol{\Sigma}^{-1}) \\
&\propto \frac{M}{2}\ln|\boldsymbol{\Sigma}^{-1}| - \frac{1}{2}\mathrm{tr}(\mathbb{E}_{q(\theta)}[\sum_{s_i\sim s_j}[\boldsymbol{\theta}(s_i)-\boldsymbol{\theta}(s_j)][\boldsymbol{\theta}(s_i)-\boldsymbol{\theta}(s_j)]^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}]) \\
&+ \frac{\nu-P-1}{2}\ln|\boldsymbol{\Sigma}^{-1}| - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1})
\end{aligned}$$

$$\begin{aligned}
q^*(\boldsymbol{\Sigma}^{-1}) &= Wishart(\boldsymbol{\Sigma}^{-1}; \nu_\Sigma, I_\Sigma) \\
\nu_\Sigma &= M + \nu \\
I_\Sigma &= \left[\mathbb{E}_{q(\theta)}[\sum_{s_i\sim s_j}[\boldsymbol{\theta}(s_i)-\boldsymbol{\theta}(s_j)][\boldsymbol{\theta}(s_i)-\boldsymbol{\theta}(s_j)]^{\mathrm{T}}] + \boldsymbol{I}\right]^{-1}
\end{aligned}$$

A.1.2.7 Update $\xi_p(s_j)$

$$\begin{aligned}
p(\gamma_p(s_j)|\theta_p(s_j)) &= \sigma(\theta_p(s_j))^{\gamma_p(s_j)}(1-\sigma(\theta_p(s_j)))^{1-\gamma_p(s_j)} = \exp(\theta_p(s_j)\gamma_p(s_j))\sigma(-\theta_p(s_j)) \\
h(\theta_p(s_j), \gamma_p(s_j), \xi_p(s_j)) &= \sigma(\theta_p(s_j))\exp\left(\theta_p(s_j)\gamma_p(s_j) - \frac{\xi_p(s_j)}{2} - \lambda(\xi_p(s_j))(\theta_p(s_j)^2 - \xi_p(s_j)^2)\right) \\
Q &\propto \sum_{p=1}^{P}\sum_{j=1}^{M}\mathbb{E}_{q(\theta)}[\log h(\theta_p(s_j), \gamma_p(s_j), \xi_p(s_j))] \\
&\propto \sum_{p=1}^{P}\sum_{j=1}^{M}\left[\log\sigma(\theta_p(s_j)) - \frac{\xi_p(s_j)}{2} - \lambda(\xi_p(s_j))(\theta_p(s_j)^2 - \xi_p(s_j)^2)\right] \\
\frac{\partial Q}{\partial \xi_p(s_j)} &= \frac{1}{\sigma(\theta_p(s_j))}\sigma(\theta_p(s_j))(1-\sigma(\theta_p(s_j))) - \frac{1}{2} \\
&- \lambda'(\xi_p(s_j))[\mathbb{E}_{q(\theta)}[\theta_p(s_j)^2 - \xi_p(s_j)^2] + 2\lambda(\xi_p(s_j))\xi_p(s_j) \overset{!}{=} 0 \\
\xi_p(s_j)^{2(t+1)} &= \mathbb{E}_{q(\theta)}[\theta_p(s_j)^2] \\
&= Var(\theta_p(s_j)) + \mathbb{E}_{q(\theta)}[\theta_p(s_j)]^2
\end{aligned}$$

A.1.3 Evidence Lower Bound (ELBO)

$$\begin{aligned}
\mathcal{L}(q) &= \ln\left\{\int\int\int\int\int\int q(z,\beta,\beta_0,\gamma,\theta,\Sigma^{-1})\frac{p(y,z,\beta,\beta_0,\gamma,\theta,\Sigma^{-1}|X)}{q(z,\beta,\beta_0,\gamma,\theta,\Sigma^{-1})}dzd\beta d\beta_0 d\gamma d\theta d\Sigma^{-1}\right\}\\
&= \ln\left\{\mathbb{E}_{q(z,\beta,\beta_0,\gamma,\theta,\Sigma^{-1})}\left[\frac{p(y,z,\beta,\beta_0,\gamma,\theta,\Sigma^{-1}|X)}{q(z,\beta,\beta_0,\gamma,\theta,\Sigma^{-1})}\right]\right\}\\
&\geq \mathbb{E}_{q(z,\beta,\beta_0,\gamma,\theta,\Sigma^{-1})}\left[\ln\left\{p(y,z,\beta,\beta_0,\gamma,\theta,\Sigma^{-1}|X)\right\}\right] -\\
&\quad \mathbb{E}_{q(z,\beta,\beta_0,\gamma,\theta,\Sigma^{-1})}\left[\ln\left\{q(z,\beta,\beta_0,\gamma,\theta,\Sigma^{-1})\right\}\right]\\
&= \mathbb{E}_{q(z)}\left[\ln\left\{p(y|z)\right\}\right] + \mathbb{E}_{q(z,\beta,\beta_0)}\left[\ln\left\{p(z|X,\beta,\beta_0)\right\}\right] + \mathbb{E}_{q(\beta,\gamma)}\left[\ln\left\{p(\beta|\gamma)\right\}\right] +\\
&\quad \mathbb{E}_{q(\gamma|\theta)}\left[\ln\left\{p(\gamma|\theta)\right\}\right] + \mathbb{E}_{q(\theta,\Sigma^{-1})}\left[\ln\left\{p(\theta|\Sigma^{-1})\right\}\right] + \mathbb{E}_{q(\Sigma^{-1})}\left[\ln\left\{p(\Sigma^{-1})\right\}\right] -\\
&\quad \mathbb{E}_{q(z)}\left[\ln\left\{q(z)\right\}\right] - \mathbb{E}_{q(\beta)}\left[\ln\left\{q(\beta)\right\}\right] - \mathbb{E}_{q(\beta_0)}\left[\ln\left\{q(\beta_0)\right\}\right] - \mathbb{E}_{q(\gamma)}\left[\ln\left\{q(\gamma)\right\}\right] -\\
&\quad \mathbb{E}_{q(\theta)}\left[\ln\left\{q(\theta)\right\}\right] - \mathbb{E}_{q(\Sigma^{-1})}\left[\ln\left\{q(\Sigma^{-1})\right\}\right]
\end{aligned}$$

A.1.3.1 $\mathbb{E}_{q(z)}\left[\ln\left\{p(y(s_j)|z(s_j))\right\}\right]$

$$\begin{aligned}
&\mathbb{E}_{q(z)}\left[\ln\left\{p(\boldsymbol{y}(s_j)|\boldsymbol{z}(s_j))\right\}\right]\\
&= \sum_{i=1}^{N}\mathbb{E}_{q(z_i(s_j))}\left[y_i(s_j)\ln\left\{\mathbb{1}(z_i(s_j)>0)\right\} + (1-y_i(s_j))\ln\left\{\mathbb{1}(z_i(s_j)\leq 0)\right\}\right]\\
&= \sum_{i=1}^{N}\int_{-\infty}^{\infty} q(z_i(s_j))\left\{y_i(s_j)\ln\left\{\mathbb{1}(z_i(s_j)>0)\right\} + (1-y_i(s_j))\ln\left\{\mathbb{1}(z_i(s_j)\leq 0)\right\}\right\}dz_i(s_j)\\
&= \begin{cases}\int_{-\infty}^{\infty} q(z_i(s_j))\ln\left\{\mathbb{1}(z_i(s_j)>0)\right\}dz_i(s_j), & \text{if } y_i(s_j)=1,\\ \int_{-\infty}^{\infty} q(z_i(s_j))\ln\left\{\mathbb{1}(z_i(s_j)\leq 0)\right\}dz_i(s_j), & \text{if } y_i(s_j)=0.\end{cases}\\
&= \begin{cases}\int_{-\infty}^{\infty} 0dz_i(s_j), & \text{if } y_i(s_j)=1,\\ \int_{-\infty}^{\infty} 0dz_i(s_j), & \text{if } y_i(s_j)=0.\end{cases}\\
&= 0
\end{aligned}$$

A.1.3.2 $\mathbb{E}_{q(z,\beta,\beta_0)}[\ln\{p(z|X,\beta,\beta_0)\}]$

$$\mathbb{E}_{q(z,\beta,\beta_0)}\left[\ln\left\{p(\boldsymbol{z}(s_j)|\boldsymbol{X},\boldsymbol{\beta}(s_j),\beta_0(s_j))\right\}\right]$$

$$= \mathbb{E}_{q(z,\beta,\beta_0)}[\ln\{\mathcal{N}(\boldsymbol{z}(s_j);\ \boldsymbol{X}\boldsymbol{\beta}(s_j)+\beta_0(s_j),\mathrm{I})\}]$$

$$= -\frac{N}{2}\ln\{2\pi\} - \frac{1}{2}\mathbb{E}_{q(z,\beta,\beta_0)}[[\boldsymbol{z}(s_j)-\boldsymbol{X}\boldsymbol{\beta}(s_j)-\beta_0(s_j)\mathbb{1}]^{\mathrm{T}}[\boldsymbol{z}(s_j)-\boldsymbol{X}\boldsymbol{\beta}(s_j)-\beta_0(s_j)\mathbb{1}]]$$

$$= -\frac{N}{2}\ln\{2\pi\} - \frac{1}{2}\mathbb{E}_{q(z)}[\boldsymbol{z}(s_j)^{\mathrm{T}}\boldsymbol{z}(s_j)] + \mathbb{E}_{q(z)}[\mathbb{E}_{q(\beta)}[\boldsymbol{\beta}(s_j)^{\mathrm{T}}]\boldsymbol{X}^{\mathrm{T}}\boldsymbol{z}(s_j)]-$$

$$\frac{1}{2}\mathbb{E}_{q(\beta)}[\boldsymbol{\beta}(s_j)^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}(s_j)] - \frac{1}{2}\mathbb{E}_{q(\beta_0)}[(\beta_0(s_j)\mathbb{1})^{\mathrm{T}}(\beta_0(s_j)\mathbb{1})]+$$

$$\mathbb{E}_{q(z)}[\boldsymbol{z}(s_j)^{\mathrm{T}}\mathbb{E}_{q(\beta_0)}[\beta_0(s_j)\mathbb{1}]] - \mathbb{E}_{q(\beta)}[\boldsymbol{\beta}(s_j)^{\mathrm{T}}]\boldsymbol{X}^{\mathrm{T}}\mathbb{E}_{q(\beta_0)}[\beta_0(s_j)\mathbb{1}]$$

$$= -\frac{N}{2}\ln\{2\pi\} - \frac{1}{2}\mathbb{E}_{q(z)}[\boldsymbol{z}(s_j)^{\mathrm{T}}\boldsymbol{z}(s_j)] + \mathbb{E}_{q(z)}[\boldsymbol{\eta}(s_j)^{\mathrm{T}}\boldsymbol{z}(s_j)]-$$

$$\frac{1}{2}\mathrm{tr}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\mathbb{E}_{q(\beta)}[\boldsymbol{\beta}(s_j)\boldsymbol{\beta}(s_j)^{\mathrm{T}}]) - \frac{1}{2}\mathbb{E}_{q(\beta_0)}[(\beta_0(s_j)\mathbb{1})^{\mathrm{T}}(\beta_0(s_j)\mathbb{1})]-$$

$$\mathbb{E}_{q(\beta)}[\boldsymbol{\beta}(s_j)^{\mathrm{T}}]\boldsymbol{X}^{\mathrm{T}}\mathbb{E}_{q(\beta_0)}[\beta_0(s_j)\mathbb{1}]$$

$$= -\frac{N}{2}\ln\{2\pi\} - \frac{1}{2}\mathbb{E}_{q(z)}[\boldsymbol{z}(s_j)^{\mathrm{T}}\boldsymbol{z}(s_j)] + \mathbb{E}_{q(z)}[\boldsymbol{\eta}(s_j)^{\mathrm{T}}\boldsymbol{z}(s_j)]-$$

$$\frac{1}{2}\mathrm{tr}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}[\boldsymbol{\Sigma}_{\beta(s_j)}+\boldsymbol{\mu}_{\beta(s_j)}\boldsymbol{\mu}_{\beta(s_j)}^{\mathrm{T}}]) - \frac{1}{2}\mathbb{E}_{q(\beta_0)}[(\beta_0(s_j)\mathbb{1})^{\mathrm{T}}(\beta_0(s_j)\mathbb{1})]-$$

$$\mathbb{E}_{q(\beta)}[\boldsymbol{\beta}(s_j)^{\mathrm{T}}]\boldsymbol{X}^{\mathrm{T}}\mathbb{E}_{q(\beta_0)}[\beta_0(s_j)\mathbb{1}]$$

A.1.3.3 $\mathbb{E}_{q(\beta,\gamma)}[\ln\{p(\beta(s_j)|\gamma(s_j))\}]$

$$= \mathbb{E}_{q(\beta,\gamma)}[\ln\{p(\boldsymbol{\beta}(s_j)|\boldsymbol{\gamma}(s_j))\}]$$

$$= \mathbb{E}_{q(\beta,\gamma)}[\ln\{\mathcal{N}(\boldsymbol{\beta}(s_j);\ \mathbf{0},\ \mathrm{diag}\{\nu_0(1-\gamma_p(s_j))+\nu_1\gamma_p(s_j)\}_{p=1}^{P})\}]$$

$$= -\frac{P}{2}\ln\{2\pi\} - \frac{1}{2}\mathbb{E}_{q(\gamma)}[\ln\{|\mathrm{diag}\{\nu_0(1-\gamma_p(s_j))+\nu_1\gamma_p(s_j)\}_{p=1}^{P})|\}]-$$

$$\frac{1}{2}\mathbb{E}_{q(\beta)}[\boldsymbol{\beta}(s_j)^{\mathrm{T}}\mathbb{E}_{q(\gamma)}[\mathrm{diag}\{\nu_0(1-\gamma_p(s_j))+\nu_1\gamma_p(s_j)\}_{p=1}^{P})]^{-1}\boldsymbol{\beta}(s_j)]$$

$$= -\frac{P}{2}\ln\{2\pi\} - \frac{1}{2}\mathbb{E}_{q(\gamma)}[\ln\{|\mathrm{diag}\{\nu_0(1-\gamma_p(s_j))+\nu_1\gamma_p(s_j)\}_{p=1}^{P})|\}]-$$

$$\frac{1}{2}\mathrm{tr}(\mathbb{E}_{q(\beta)}[\boldsymbol{\beta}(s_j)\boldsymbol{\beta}(s_j)^{\mathrm{T}}]\mathbb{E}_{q(\gamma)}[\mathrm{diag}\{\nu_0(1-\gamma_p(s_j))+\nu_1\gamma_p(s_j)\}_{p=1}^{P}]^{-1})$$

$$= -\frac{P}{2}\ln\{2\pi\} - \frac{1}{2}\mathbb{E}_{q(\gamma)}[\ln\{|\mathrm{diag}\{\nu_0(1-\gamma_p(s_j))+\nu_1\gamma_p(s_j)\}_{p=1}^{P})|\}]-$$

$$\frac{1}{2}\mathrm{tr}([\boldsymbol{\Sigma}_{\beta(s_j)}+\boldsymbol{\mu}_{\beta(s_j)}\boldsymbol{\mu}_{\beta(s_j)}^{\mathrm{T}}]\mathbb{E}_{q(\gamma)}[\mathrm{diag}\{\nu_0(1-\gamma_p(s_j))+\nu_1\gamma_p(s_j)\}_{p=1}^{P}]^{-1})$$

A.1.3.4 $\mathbb{E}_{q(\beta_0)}[\ln\{p(\beta_0(s_j))\}]$

$$
\begin{aligned}
&\mathbb{E}_{q(\beta_0)}\left[\ln\{p(\beta_0(s_j))\}\right]\\
&= \mathbb{E}_{q(\beta_0)}\left[\ln\left\{\frac{1}{\sqrt{2\pi\sigma^2_{\beta_0}}}\exp\left\{-\frac{1}{2\sigma^2_{\beta_0}}\beta_0(s_j)^2\right\}\right\}\right]\\
&= -\frac{1}{2}\ln\{2\pi\sigma^2_{\beta_0}\} - \frac{1}{2\sigma^2_{\beta_0}}\mathbb{E}_{q(\beta_0)}[\beta_0(s_j)^2]\\
&= -\frac{1}{2}\ln\{2\pi\sigma^2_{\beta_0}\} - \frac{1}{2\sigma^2_{\beta_0}}\left[\sigma^2_{\beta_0(s_j)} + \mu^2_{\beta_0(s_j)}\right]
\end{aligned}
$$

A.1.3.5 $\mathbb{E}_{q(\gamma,\theta)}[\ln\{p(\gamma(s_j)|\theta(s_j))\}]$

$$
\begin{aligned}
&\mathbb{E}_{q(\gamma,\theta)}[\ln\{p(\boldsymbol{\gamma}(s_j)|\boldsymbol{\theta}(s_j))\}]\\
&= \mathbb{E}_{q(\gamma,\theta)}\left[\ln\left\{\prod_{p=1}^{P}\sigma(\theta_p(s_j))^{\gamma_p(s_j)}(1-\sigma(\theta_p(s_j)))^{(1-\gamma_p(s_j))}\right\}\right]\\
&\geq \sum_{p=1}^{P}\mathbb{E}_{q(\gamma,\theta)}\left[\theta_p(s_j)\gamma_p(s_j) + \ln\{\sigma(\xi_p(s_j))\} - \frac{1}{2}(\theta_p(s_j)+\xi_p(s_j)) - \right.\\
&\quad\left.\lambda(\xi_p(s_j))[\theta_p(s_j)^2 - \xi_p(s_j)^2]\right]\\
&= \sum_{p=1}^{P}\left\{\mathbb{E}_{q(\gamma)}[\gamma_p(s_j)]\mathbb{E}_{q(\theta)}[\theta_p(s_j)] + \ln\{\sigma(\xi_p(s_j))\} - \frac{1}{2}\mathbb{E}_{q(\theta)}[\theta_p(s_j)] - \frac{1}{2}\xi_p(s_j) - \right.\\
&\quad\left.\lambda(\xi_p(s_j))\mathbb{E}_{q(\theta)}[\theta_p(s_j)^2] + \lambda(\xi_p(s_j))\xi_p(s_j)^2\right\}\\
&= \sum_{p=1}^{P}\left\{\mathbb{E}_{q(\gamma)}[\gamma_p(s_j)]\mu_{\theta_p(s_j)} + \ln\{\sigma(\xi_p(s_j))\} - \frac{1}{2}\mu_{\theta_p(s_j)} - \frac{1}{2}\xi_p(s_j) - \right.\\
&\quad\left.\lambda(\xi_p(s_j))[\Sigma_{\theta_p(s_j)} + \mu_{\theta_p(s_j)}\mu^{\mathrm{T}}_{\theta_p(s_j)}] + \lambda(\xi_p(s_j))\xi_p(s_j)^2\right\}
\end{aligned}
$$

A.1.3.6 $\mathbb{E}_{q(\theta,\Sigma^{-1})}[\ln\{p(\theta|\Sigma^{-1})\}]$

$$
\begin{aligned}
&\mathbb{E}_{q(\theta,\Sigma^{-1})}[\ln\{p(\boldsymbol{\theta}|\boldsymbol{\Sigma}^{-1})\}]\\
&= -\frac{MP}{2}\ln\{2\pi\} + \frac{M}{2}\mathbb{E}_{q(\Sigma^{-1})}[\ln\{|\boldsymbol{\Sigma}^{-1}|\}] - \\
&\quad\frac{1}{2}\sum_{s_i\sim s_j}\mathbb{E}_{q(\theta)}[[\boldsymbol{\theta}(s_i)-\boldsymbol{\theta}(s_j)]^{\mathrm{T}}\mathbb{E}_{q(\Sigma^{-1})}[\boldsymbol{\Sigma}^{-1}][\boldsymbol{\theta}(s_i)-\boldsymbol{\theta}(s_j)]]\\
&= -\frac{MP}{2}\ln\{2\pi\} + \frac{M}{2}\mathbb{E}_{q(\Sigma^{-1})}[\ln\{|\boldsymbol{\Sigma}^{-1}|\}] - \\
&\quad\frac{1}{2}\mathrm{tr}\left(\mathbb{E}_{q(\Sigma^{-1})}\left[\boldsymbol{\Sigma}^{-1}\right]\mathbb{E}_{q(\theta)}\left[\sum_{s_i\sim s_j}[\boldsymbol{\theta}(s_i)-\boldsymbol{\theta}(s_j)][\boldsymbol{\theta}(s_i)-\boldsymbol{\theta}(s_j)]^{\mathrm{T}}\right]\right)
\end{aligned}
$$

A.1.3.7 $\mathbb{E}_{q(\Sigma^{-1})}[\ln\{p(\Sigma^{-1})\}]$

$$\begin{aligned}
&\mathbb{E}_{q(\Sigma^{-1})}[\ln\{p(\boldsymbol{\Sigma}^{-1})\}]\\
&= \mathbb{E}_{q(\Sigma^{-1})}\left[\frac{(\nu-P-1)}{2}\ln\{|\boldsymbol{\Sigma}^{-1}|\} - \frac{1}{2}\mathrm{tr}(\boldsymbol{I}\boldsymbol{\Sigma}^{-1}) - \frac{\nu P}{2}\ln\{2\} - \frac{P(P-1)}{4}\ln(\pi) - \right.\\
&\quad \left.\sum_{p=1}^{P}\ln\left\{\Gamma\left(\frac{\nu+1-p}{2}\right)\right\} - \frac{\nu}{2}\ln(|\boldsymbol{I}|)\right]\\
&= \frac{(\nu-P-1)}{2}\mathbb{E}_{q(\Sigma^{-1})}[\ln\{|\boldsymbol{\Sigma}^{-1}|\}] - \frac{1}{2}\mathrm{tr}(\boldsymbol{I}\mathbb{E}_{q(\Sigma^{-1})}[\boldsymbol{\Sigma}^{-1}]) - \frac{\nu P}{2}\ln\{2\} - \frac{P(P-1)}{4}\ln(\pi) - \\
&\quad \sum_{p=1}^{P}\ln\left\{\Gamma\left(\frac{\nu+1-p}{2}\right)\right\} - \frac{\nu}{2}\ln(P)
\end{aligned}$$

A.1.3.8 $\mathbb{E}_{q(z)}[\ln\{q(z(s_j))\}]$

$$\begin{aligned}
&\mathbb{E}_{q(z)}[\ln\{q(\boldsymbol{z}(s_j))\}]\\
&= \sum_{i=1}^{N}\mathbb{E}_{q(z)}\left[\ln\left\{\mathcal{TN}_{+}(z_i(s_j);\ \eta_i(s_j),\ 1)^{y_i(s_j)}\mathcal{TN}_{-}(z_i(s_j);\ \eta_i(s_j),\ 1)^{(1-y_i(s_j))}\right\}\right]\\
&= \sum_{i=1}^{N}\mathbb{E}_{q(z)}\left[\ln\left\{\mathcal{N}(z_i(s_j);\ \eta_i(s_j),\ 1)\left[\frac{1}{1-\Phi(-\eta_i(s_j))}\right]^{y_i(s_j)}\left[\frac{1}{\Phi(-\eta_i(s_j))}\right]^{(1-y_i(s_j))}\right\}\right]\\
&= \mathbb{E}_{q(z)}\left[\ln\left\{\mathcal{N}(\boldsymbol{z}(s_j);\ \boldsymbol{\eta}(s_j),\ \boldsymbol{I})\right\}\right] - \\
&\quad \sum_{i=1}^{N}\left\{y_i(s_j)\ln\left\{1-\Phi(-\eta_i(s_j))\right\} + (1-y_i(s_j))\ln\left\{\Phi(-\eta_i(s_j))\right\}\right\}\\
&= -\frac{N}{2}\ln\{2\pi\} - \frac{1}{2}\mathbb{E}_{q(z)}\left[[\boldsymbol{z}(s_j)-\boldsymbol{\eta}(s_j)]^{\mathrm{T}}[\boldsymbol{z}(s_j)-\boldsymbol{\eta}(s_j)]\right] - \\
&\quad \sum_{i=1}^{N}\left\{y_i(s_j)\ln\left\{1-\Phi(-\eta_i(s_j))\right\} + (1-y_i(s_j))\ln\left\{\Phi(-\eta_i(s_j))\right\}\right\}\\
&= -\frac{N}{2}\ln\{2\pi\} - \frac{1}{2}\mathbb{E}_{q(z)}[\boldsymbol{z}(s_j)^{\mathrm{T}}\boldsymbol{z}(s_j)] + \mathbb{E}_{q(z)}[\boldsymbol{\eta}(s_j)^{\mathrm{T}}\boldsymbol{z}(s_j)] - \frac{1}{2}\boldsymbol{\eta}(s_j)^{\mathrm{T}}\boldsymbol{\eta}(s_j) - \\
&\quad \sum_{i=1}^{N}\left\{y_i(s_j)\ln\left\{1-\Phi(-\eta_i(s_j))\right\} + (1-y_i(s_j))\ln\left\{\Phi(-\eta_i(s_j))\right\}\right\}
\end{aligned}$$

A.1.3.9 $\mathbb{E}_{q(\beta)}[\ln\{q(\beta(s_j))\}]$

$$
\begin{aligned}
&\mathbb{E}_{q(\beta)}[\ln\{q(\boldsymbol{\beta}(s_j))\}] \\
&= \mathbb{E}_{q(\beta)}[\ln\{\mathcal{N}(\boldsymbol{\beta}(s_j);\ \boldsymbol{\mu}_{\beta(s_j)},\ \boldsymbol{\Sigma}_{\beta(s_j)})\}] \\
&= -\frac{P}{2}\ln\{2\pi\} - \frac{1}{2}\ln\{|\boldsymbol{\Sigma}_{\beta(s_j)}|\} - \frac{1}{2}\mathbb{E}_{q(\beta)}\left[[\boldsymbol{\beta}(s_j) - \boldsymbol{\mu}_{\beta(s_j)}]^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}_{\beta(s_j)}[\boldsymbol{\beta}(s_j) - \boldsymbol{\mu}_{\beta(s_j)}]\right] \\
&= -\frac{P}{2}\ln\{2\pi\} - \frac{1}{2}\ln\{|\boldsymbol{\Sigma}_{\beta(s_j)}|\} - \frac{1}{2}\mathbb{E}_{q(\beta)}\left[\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}_{\beta(s_j)}[\boldsymbol{\beta}(s_j) - \boldsymbol{\mu}_{\beta(s_j)}][\boldsymbol{\beta}(s_j) - \boldsymbol{\mu}_{\beta(s_j)}]^{\mathrm{T}}\right)\right] \\
&= -\frac{P}{2}\ln\{2\pi\} - \frac{1}{2}\ln\{|\boldsymbol{\Sigma}_{\beta(s_j)}|\} - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}_{\beta(s_j)}\boldsymbol{\Sigma}_{\beta(s_j)}) \\
&= -\frac{P}{2}\ln\{2\pi\} - \frac{1}{2}\ln\{|\boldsymbol{\Sigma}_{\beta(s_j)}|\} - \frac{1}{2}\mathrm{tr}(\boldsymbol{I}) \\
&= -\frac{P}{2}\ln\{2\pi\} - \frac{1}{2}\ln\{|\boldsymbol{\Sigma}_{\beta(s_j)}|\} - \frac{P}{2}
\end{aligned}
$$

A.1.3.10 $\mathbb{E}_{q(\beta_0)}[\ln\{q(\beta_0(s_j))\}]$

$$
\begin{aligned}
&\mathbb{E}_{q(\beta_0)}[\ln\{q(\beta_0(s_j))\}] \\
&= \mathbb{E}_{q(\beta_0)}\left[-\frac{1}{2}\ln\left\{2\pi\sigma^2_{\beta_0(s_j)}\right\} - \frac{1}{2\sigma^2_{\beta_0(s_j)}}\left(\beta_0(s_j) - \mu_{\beta_0(s_j)}\right)^2\right] \\
&= -\frac{1}{2}\ln\left\{2\pi\sigma^2_{\beta_0(s_j)}\right\} - \frac{1}{2\sigma^2_{\beta_0(s_j)}}\sigma^2_{\beta_0(s_j)} \\
&= -\frac{1}{2}\ln\left\{2\pi\sigma^2_{\beta_0(s_j)}\right\} - \frac{1}{2}
\end{aligned}
$$

A.1.3.11 $\mathbb{E}_{q(\gamma)}[\ln\{q(\gamma(s_j))\}]$

$$
\begin{aligned}
&\mathbb{E}_{q(\gamma)}[\ln\{q(\boldsymbol{\gamma}(s_j))\}] \\
&= \sum_{p=1}^{P}\mathbb{E}_{q(\gamma)}[\gamma_p(s_j)\ln\{\mu_{\gamma_p(s_j)}\} + (1 - \gamma_p(s_j))\ln\{1 - \mu_{\gamma_p(s_j)}\}] \\
&= \sum_{p=1}^{P}\left\{\mu_{\gamma_p(s_j)}\ln\{\mu_{\gamma_p(s_j)}\} + (1 - \mu_{\gamma_p(s_j)})\ln\{1 - \mu_{\gamma_p(s_j)}\}\right\}
\end{aligned}
$$

A.1.3.12 $\mathbb{E}_{q(\theta)}[\ln\{q(\theta(s_j))\}]$

$$
\begin{aligned}
&\mathbb{E}_{q(\theta)}[\ln\{q(\boldsymbol{\theta}(s_j))\}]\\
&= \mathbb{E}_{q(\theta)}\left[\ln\left\{\mathcal{N}\left(\boldsymbol{\theta}(s_j);\ \boldsymbol{\mu}_{\theta(s_j)},\ \boldsymbol{\Sigma}_{\theta(s_j)}\right)\right\}\right]\\
&= -\frac{P}{2}\ln\{2\pi\} - \frac{1}{2}\ln\left\{|\boldsymbol{\Sigma}_{\theta(s_j)}|\right\} - \frac{1}{2}\mathbb{E}_{q(\theta)}\left[[\boldsymbol{\theta}(s_j) - \boldsymbol{\mu}_{\theta(s_j)}]^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}_{\theta(s_j)}[\boldsymbol{\theta}(s_j) - \boldsymbol{\mu}_{\theta(s_j)}]\right]\\
&= -\frac{P}{2}\ln\{2\pi\} - \frac{1}{2}\ln\left\{|\boldsymbol{\Sigma}_{\theta(s_j)}|\right\} - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}_{\theta(s_j)}\mathbb{E}_{q(\theta)}\left[[\boldsymbol{\theta}(s_j) - \boldsymbol{\mu}_{\theta(s_j)}][\boldsymbol{\theta}(s_j) - \boldsymbol{\mu}_{\theta(s_j)}]^{\mathrm{T}}\right]\right)\\
&= -\frac{P}{2}\ln\{2\pi\} - \frac{1}{2}\ln\left\{|\boldsymbol{\Sigma}_{\theta(s_j)}|\right\} - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}_{\theta(s_j)}\boldsymbol{\Sigma}_{\theta(s_j)}\right)\\
&= -\frac{P}{2}\ln\{2\pi\} - \frac{1}{2}\ln\left\{|\boldsymbol{\Sigma}_{\theta(s_j)}|\right\} - \frac{1}{2}\mathrm{tr}(\boldsymbol{I})\\
&= -\frac{P}{2}\ln\{2\pi\} - \frac{1}{2}\ln\left\{|\boldsymbol{\Sigma}_{\theta(s_j)}|\right\} - \frac{P}{2}
\end{aligned}
$$

A.1.3.13 $\mathbb{E}_{q(\Sigma^{-1})}[\ln\{q(\Sigma^{-1})\}]$

The posterior quantities of $\boldsymbol{\Sigma}^{-1}$ are denoted by the following: posterior matrix $\boldsymbol{I}_{\Sigma^{-1}}$ and the posterior degrees of freedom $\nu_{\Sigma^{-1}}$. For the expressions of those quantities see the derivations above.

$$
\begin{aligned}
&\mathbb{E}_{q(\Sigma^{-1})}[\ln\{q(\boldsymbol{\Sigma}^{-1})\}]\\
&= \mathbb{E}_{q(\Sigma^{-1})}\left[\frac{(\nu_{\Sigma^{-1}} - P - 1)}{2}\ln\left\{|\boldsymbol{\Sigma}^{-1}|\right\} - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{I}^{-1}_{\Sigma^{-1}}\right) - \frac{\nu_{\Sigma^{-1}}P}{2}\ln\{2\}\right.\\
&\left.-\frac{\nu_{\Sigma^{-1}}}{2}\ln\left\{|\boldsymbol{I}_{\Sigma^{-1}}|\right\} - \frac{P(P-1)}{4}\ln\{\pi\} - \sum_{p=1}^{P}\ln\left\{\Gamma\left(\frac{\nu_{\Sigma^{-1}} + 1 - p}{2}\right)\right\}\right]\\
&= \frac{(\nu_{\Sigma^{-1}} - P - 1)}{2}\mathbb{E}_{q(\Sigma^{-1})}\ln\{|\boldsymbol{\Sigma}^{-1}|\}] - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{I}^{-1}_{\Sigma^{-1}}\mathbb{E}_{q(\Sigma^{-1})}[\boldsymbol{\Sigma}^{-1}]\right) - \frac{\nu_{\Sigma^{-1}}P}{2}\ln\{2\}\\
&- \frac{\nu_{\Sigma^{-1}}}{2}\ln\left\{|\boldsymbol{I}_{\Sigma^{-1}}|\right\} - \frac{P(P-1)}{4}\ln\{\pi\} - \sum_{p=1}^{P}\ln\left\{\Gamma\left(\frac{\nu_{\Sigma^{-1}} + 1 - p}{2}\right)\right\}\\
&= \frac{(M + \nu - P - 1)}{2}\mathbb{E}_{q(\Sigma^{-1})}\ln\{|\boldsymbol{\Sigma}^{-1}|\}] - \frac{(M+\nu)P}{2} - \frac{(M+\nu)P}{2}\ln\{2\}\\
&- \frac{(M+\nu)}{2}\ln\left\{\left|\left[\boldsymbol{I} + \mathbb{E}_{q(\theta)}\left[\sum_{s_i \sim s_j}[\boldsymbol{\theta}(s_i) - \boldsymbol{\theta}(s_j)][\boldsymbol{\theta}(s_i) - \boldsymbol{\theta}(s_j)]^{\mathrm{T}}\right]\right]^{-1}\right|\right\}\\
&- \frac{P(P-1)}{4}\ln\{\pi\} - \sum_{p=1}^{P}\ln\left\{\Gamma\left(\frac{\nu_{\Sigma^{-1}} + 1 - p}{2}\right)\right\}
\end{aligned}
$$

A.1.3.14 Total ELBO

The ELBO is an alternative optimization objective which consists of the negative KL-divergence and an added constant, the logarithm of the evidence $p(\boldsymbol{y})$. Therefore, minimizing the KL-divergence is equivalent to maximizing the ELBO (Blei et al., 2017).

The evidence lower bound for BLESS is defined via Jensen's inequality (Jordan et al., 1999) by

$$\begin{aligned}\mathcal{L}(q) \geq \mathbb{E}_{q(\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\beta_0},\boldsymbol{\gamma},\boldsymbol{\theta},\boldsymbol{\Sigma}^{-1})}\left[\ln\left\{p(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\beta_0},\boldsymbol{\gamma},\boldsymbol{\theta},\boldsymbol{\Sigma}^{-1})\right\}\right] - \quad &(\text{A.1})\\ \mathbb{E}_{q(\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\beta_0},\boldsymbol{\gamma},\boldsymbol{\theta},\boldsymbol{\Sigma}^{-1})}\left[\ln\left\{q(\boldsymbol{Z},\boldsymbol{\beta},\boldsymbol{\beta_0},\boldsymbol{\gamma},\boldsymbol{\theta},\boldsymbol{\Sigma}^{-1})\right\}\right]. \quad &(\text{A.2})\end{aligned}$$

$$\begin{aligned}\mathcal{L}(q) = &\sum_{j=1}^{M}\Big\{-\frac{N}{2}\ln\{2\pi\} - \frac{1}{2}\mathbb{E}_{q(z)}[\boldsymbol{z}(s_j)^{\mathrm{T}}\boldsymbol{z}(s_j)] + \mathbb{E}_{q(z)}[\boldsymbol{\eta}(s_j)^{\mathrm{T}}\boldsymbol{z}(s_j)] -\\ &\frac{1}{2}\mathrm{tr}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\left[\boldsymbol{\Sigma}_{\beta(s_j)} + \boldsymbol{\mu}_{\beta(s_j)}\boldsymbol{\mu}_{\beta(s_j)}^{\mathrm{T}}\right]\right) - \frac{1}{2}\mathbb{E}_{q(\beta_0)}\left[[\beta_0(s_j)\mathbb{1}]^{\mathrm{T}}[\beta_0(s_j)\mathbb{1}]\right] -\\ &\mathbb{E}_{q(\beta)}[\boldsymbol{\beta}(s_j)^{\mathrm{T}}]\boldsymbol{X}^{\mathrm{T}}\mathbb{E}_{q(\beta_0)}[\beta_0(s_j)\mathbb{1}]\Big\} +\\ &\sum_{j=1}^{M}\left\{-\frac{P}{2}\ln\{2\pi\} - \frac{1}{2}\sum_{p=1}^{P}\left\{\mathbb{E}_{q(\gamma)}\left[\ln\{\nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j)\}\right]\right\} -\right.\\ &\left.\frac{1}{2}\mathrm{tr}\left(\mathbb{E}_{q(\gamma)}\left[\mathrm{diag}\left\{\nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j)\right\}_{p=1}^{P}\right]^{-1}\left[\boldsymbol{\Sigma}_{\beta(s_j)} + \boldsymbol{\mu}_{\beta(s_j)}\boldsymbol{\mu}_{\beta(s_j)}^{\mathrm{T}}\right]\right)\right\} +\\ &\sum_{j=1}^{M}\left\{-\frac{1}{2}\ln\{2\pi\sigma_{\beta_0}^2\} - \frac{1}{2\sigma_{\beta_0}^2}\left[\sigma_{\beta_0(s_j)}^2 + \mu_{\beta_0(s_j)}^2\right]\right\} +\\ &\sum_{j=1}^{M}\sum_{p=1}^{P}\Big\{\mathbb{E}_{q(\gamma)}[\gamma_p(s_j)]\mu_{\theta_p(s_j)} + \ln\{\sigma(\xi_p(s_j))\} - \frac{1}{2}\mu_{\theta_p(s_j)} - \frac{1}{2}\xi_p(s_j) -\\ &\lambda(\xi_p(s_j))\left[\Sigma_{\theta_p(s_j)} + \mu_{\theta_p(s_j)}\mu_{\theta_p(s_j)}^{\mathrm{T}} - \xi_p(s_j)^2\right]\Big\} - \frac{MP}{2}\ln\{2\pi\} + \frac{M}{2}\mathbb{E}_{q(\Sigma^{-1})}[\ln\{|\Sigma^{-1}|\}] -\\ &\frac{1}{2}\mathrm{tr}\left(\mathbb{E}_{q(\Sigma^{-1})}[\boldsymbol{\Sigma}^{-1}]\sum_{s_i\sim s_j}\mathbb{E}_{q(\theta)}\left[[\boldsymbol{\theta}(s_i)-\boldsymbol{\theta}(s_j)][\boldsymbol{\theta}(s_i)-\boldsymbol{\theta}(s_j)]^{\mathrm{T}}\right]\right) +\\ &\frac{(\nu-P-1)}{2}\mathbb{E}_{q(\Sigma^{-1})}[\ln\{|\Sigma^{-1}|\}] - \frac{1}{2}\mathrm{tr}\left(\mathbb{E}_{q(\Sigma^{-1})}[\boldsymbol{\Sigma}^{-1}]\boldsymbol{I}^{-1}\right) - \frac{\nu P}{2}\ln\{2\} -\\ &\frac{P(P-1)}{4}\ln(\pi) - \sum_{p=1}^{P}\ln\left\{\Gamma\left(\frac{\nu+1-p}{2}\right)\right\} - \frac{\nu}{2}\ln\{P\}\\ &\sum_{j=1}^{M}\Big\{-\frac{N}{2}\ln\{2\pi\} - \frac{1}{2}\mathbb{E}_{q(z)}[\boldsymbol{z}(s_j)^{\mathrm{T}}\boldsymbol{z}(s_j)] + \mathbb{E}_{q(z)}[\boldsymbol{\eta}(s_j)^{\mathrm{T}}\boldsymbol{z}(s_j)] - \frac{1}{2}\boldsymbol{\eta}(s_j)^{\mathrm{T}}\boldsymbol{\eta} -\end{aligned}$$

$$\sum_{i=1}^{N} \{y_i(s_j) \ln \{1 - \Phi(-\eta_i(s_j))\} + (1 - y_i(s_j)) \ln \{\Phi(-\eta_i(s_j))\}\} \Bigg\} -$$

$$\sum_{j=1}^{M} \left\{ -\frac{P}{2} \ln\{2\pi\} - \frac{1}{2} \ln\{|\boldsymbol{\Sigma}_{\beta(s_j)}|\} - \frac{P}{2} \right\} - \sum_{j=1}^{M} \left\{ -\frac{1}{2} \ln \left\{ 2\pi\sigma^2_{\beta_0(s_j)} \right\} - \frac{1}{2} \right\} -$$

$$\sum_{j=1}^{M} \sum_{p=1}^{P} \Big\{ \mu_{\gamma_p(s_j)} \ln\{\mu_{\gamma_p(s_j)}\} + (1 - \mu_{\gamma_p(s_j)}) \ln\{1 - \mu_{\gamma_p(s_j)}\Big\} -$$

$$\sum_{j=1}^{M} \left\{ -\frac{P}{2} \ln\{2\pi\} - \frac{1}{2} \ln \Big\{ |\boldsymbol{\Sigma}_{\theta(s_j)}| \Big\} - \frac{P}{2} \right\} -$$

$$\frac{(M + \nu - P - 1)}{2} \mathbb{E}_{q(\Sigma^{-1})} \ln\{|\boldsymbol{\Sigma}^{-1}|\}\Big] + \frac{(M + \nu)P}{2} + \frac{(M + \nu)P}{2} \ln\{2\} +$$

$$\frac{(M + \nu)}{2} \ln \left\{ \left| \left[\boldsymbol{I} + \mathbb{E}_{q(\theta)} \left[\sum_{s_i \sim s_j} [\boldsymbol{\theta}(s_i) - \boldsymbol{\theta}(s_j)][\boldsymbol{\theta}(s_i) - \boldsymbol{\theta}(s_j)]^{\mathrm{T}} \right] \right]^{-1} \right| \right\} +$$

$$\frac{P(P - 1)}{4} \ln\{\pi\} + \sum_{p=1}^{P} \ln \left\{ \Gamma \left(\frac{\nu_{\Sigma^{-1}} + 1 - p}{2} \right) \right\}$$

A.1.3.15 Simplified ELBO

$$\mathcal{L}(q) = -\frac{1}{2}\sum_{j=1}^{M} \operatorname{tr}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\left[\boldsymbol{\Sigma}_{\beta(s_j)} + \boldsymbol{\mu}_{\beta(s_j)}\boldsymbol{\mu}_{\beta(s_j)}^{\mathrm{T}}\right]\right) - \frac{1}{2}\sum_{j=1}^{M} \mathbb{E}_{q(\beta_0)}\left[[\beta_0(s_j)\mathbb{1}]^{\mathrm{T}}[\beta_0(s_j)\mathbb{1}]\right] -$$

$$\sum_{j=1}^{M} \mathbb{E}_{q(\beta)}[\boldsymbol{\beta}(s_j)^{\mathrm{T}}]\boldsymbol{X}^{\mathrm{T}}\mathbb{E}_{q(\beta_0)}[\beta_0(s_j)\mathbb{1}] + \frac{1}{2}\sum_{j=1}^{M}\boldsymbol{\eta}(s_j)^{\mathrm{T}}\boldsymbol{\eta}(s_j) +$$

$$\sum_{j=1}^{M}\sum_{i=1}^{N}\left\{y_i(s_j)\ln\left\{1-\Phi(-\eta_i(s_j))\right\} + (1-y_i(s_j))\ln\left\{\Phi(-\eta_i(s_j))\right\}\right\} -$$

$$\frac{1}{2}\sum_{j=1}^{M}\sum_{p=1}^{P}\mathbb{E}_{q(\gamma)}[\ln\{\nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j)\}] -$$

$$\frac{1}{2}\sum_{j=1}^{M}\operatorname{tr}\left(\mathbb{E}_{q(\gamma)}\left[\operatorname{diag}\left\{\nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j)\right\}_{p=1}^{P}\right]^{-1}\left[\boldsymbol{\Sigma}_{\beta(s_j)} + \boldsymbol{\mu}_{\beta(s_j)}\boldsymbol{\mu}_{\beta(s_j)}^{\mathrm{T}}\right]\right) +$$

$$\frac{1}{2}\sum_{j=1}^{M}\ln\left\{|\boldsymbol{\Sigma}_{\beta(s_j)}|\right\} + \frac{PM}{2} - \frac{M}{2}\ln\left\{\sigma_{\beta_0}^2\right\} - \frac{1}{2\sigma_{\beta_0}^2}\sum_{j=1}^{M}\left[\sigma_{\beta_0(s_j)}^2 + \mu_{\beta_0(s_j)}^2\right] +$$

$$\frac{1}{2}\sum_{j=1}^{M}\ln\left\{\sigma_{\beta_0(s_j)}^2\right\} + \frac{M}{2} +$$

$$\sum_{j=1}^{M}\sum_{p=1}^{P}\left\{\mu_{\gamma_p(s_j)}\mu_{\theta_p(s_j)} + \ln\{\sigma(\xi_p(s_j))\} - \frac{1}{2}(\mu_{\theta_p(s_j)} + \xi_p(s_j)) -\right.$$

$$\left.\lambda(\xi_p(s_j))\left[\Sigma_{\theta_p(s_j)} + \mu_{\theta_p(s_j)}\mu_{\theta_p(s_j)}^{\mathrm{T}} - \xi_p(s_j)^2\right]\right\} -$$

$$\sum_{j=1}^{M}\sum_{p=1}^{P}\left\{\mu_{\gamma_p(s_j)}\ln\{\mu_{\gamma_p(s_j)}\} + (1-\mu_{\gamma_p(s_j)})\ln\{1-\mu_{\gamma_p(s_j)}\right\} -$$

$$\frac{1}{2}\operatorname{tr}\left(\mathbb{E}_{q(\Sigma^{-1})}\left[\boldsymbol{\Sigma}^{-1}\right]\mathbb{E}_{q(\theta)}\left[\sum_{s_i\sim s_j}[\boldsymbol{\theta}(s_i)-\boldsymbol{\theta}(s_j)][\boldsymbol{\theta}(s_i)-\boldsymbol{\theta}(s_j)]^{\mathrm{T}}\right]\right) +$$

$$\frac{1}{2}\sum_{j=1}^{M}\ln\left\{|\boldsymbol{\Sigma}_{\theta(s_j)}|\right\} + \frac{PM}{2} - \frac{1}{2}\operatorname{tr}(\boldsymbol{I}\mathbb{E}_{q(\Sigma^{-1})}[\boldsymbol{\Sigma}^{-1}]) - \frac{\nu P}{2}\ln\{2\} -$$

$$\sum_{p=1}^{P}\ln\left\{\Gamma\left(\frac{\nu+1-p}{2}\right)\right\} - \frac{\nu}{2}\ln\{P\} +$$

$$\frac{(M+\nu)P}{2} + \frac{(M+\nu)P}{2}\ln\{2\} + \sum_{p=1}^{P}\ln\left\{\Gamma\left(\frac{\nu_{\Sigma^{-1}}+1-p}{2}\right)\right\} +$$

$$\frac{(M+\nu)}{2}\ln\left\{\left|\left[\boldsymbol{I} + \mathbb{E}_{q(\theta)}\left[\sum_{s_i\sim s_j}[\boldsymbol{\theta}(s_i)-\boldsymbol{\theta}(s_j)][\boldsymbol{\theta}(s_i)-\boldsymbol{\theta}(s_j)]^{\mathrm{T}}\right]\right]^{-1}\right|\right\}$$

A.1.4 Log Marginal Posterior of γ

$$\begin{aligned}
\nu_0 &= 0 \\
\boldsymbol{X}_\gamma &= \begin{cases} \boldsymbol{x}_p, & \text{if } \gamma_p = 1, \\ 0, & \text{if } \gamma_p = 0. \end{cases} \\
\boldsymbol{\beta}_\gamma &= \begin{cases} \beta_p, & \text{if } \gamma_p = 1, \\ 0, & \text{if } \gamma_p = 0. \end{cases} \\
\gamma_p &= \begin{cases} 1, & \text{if } p(\gamma_p = 1|\boldsymbol{y}) > 0.5, \\ 0, & \text{if } p(\gamma_p = 1|\boldsymbol{y}) \leq 0.5. \end{cases} \\
\boldsymbol{\eta} &= \boldsymbol{X}_\gamma \mathbb{E}_{q(\beta)}[\boldsymbol{\beta}_\gamma] + \mathbb{E}_{q(\beta_0)}[\beta_0] \\
\boldsymbol{\Sigma}_\beta &= \left[\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma + \operatorname{diag}\left\{\mathbb{E}_{q(\gamma)}[\gamma_p]\frac{1}{\nu_1}\right\}\right] \\
\boldsymbol{\mu}_\beta &= \mathbb{E}_{q(\beta)}[\boldsymbol{\beta}_\gamma] \\
Q &= \operatorname{rank}(\boldsymbol{X}_\gamma) \\
\theta &= \mathbb{E}_{q(\theta)}[\theta] \\
\Sigma^{-1} &= \mathbb{E}_{q(\Sigma^{-1})}[\Sigma^{-1}]
\end{aligned}$$

All rows/columns where $\gamma_p(s_j) = 0$ are set to 0. Therefore, forming a sub-matrices/sub-vectors. For example, $\boldsymbol{X} \in \mathbb{R}^{N \times P}$ is represented by $\boldsymbol{X}_\gamma \in \mathbb{R}^{N \times Q}$.

$$\begin{aligned}
\ln\{p(\boldsymbol{\gamma}|\boldsymbol{y})\} &= \ln\left\{\int\int\int q(z,\beta,\beta_0)\frac{p(y,z,\beta,\beta_0,\gamma,\theta,\Sigma^{-1}|X)}{q(z,\beta,\beta_0)} dz d\beta d\beta_0\right\} \\
&= \ln\left\{\mathbb{E}_{q(z,\beta,\beta_0)}\left[\frac{p(y,z,\beta,\beta_0,\gamma,\theta,\Sigma^{-1}|X)}{q(z,\beta,\beta_0)}\right]\right\} \\
&\geq \mathbb{E}_{q(z,\beta,\beta_0)}\left[\ln\left\{p(y,z,\beta,\beta_0,\gamma,\theta,\Sigma^{-1}|X)\right\}\right] - \mathbb{E}_{q(z,\beta,\beta_0)}\left[\ln\left\{q(z,\beta,\beta_0)\right\}\right] \\
&= \mathbb{E}_{q(z)}[\ln\{p(y|z)\}] + \mathbb{E}_{q(z,\beta,\beta_0)}[\ln\{p(z|X,\beta,\beta_0)\}] + \mathbb{E}_{q(\beta)}[\ln\{p(\beta|\gamma)\}] \\
&+ \mathbb{E}_{q(\beta_0)}[\ln\{p(\beta_0)\}] + \ln\{p(\gamma|\theta)\} + \ln\{p(\theta|\Sigma^{-1})\} + \ln\{p(\Sigma^{-1})\} \\
&- \mathbb{E}_{q(z)}[\ln\{q(z)\}] - \mathbb{E}_{q(\beta)}[\ln\{q(\beta)\}] - \mathbb{E}_{q(\beta_0)}[\ln\{q(\beta_0)\}] \\
&\equiv \ln\{g(\boldsymbol{\gamma})\}
\end{aligned}$$

$$
\begin{aligned}
\ln\{g(\boldsymbol{\gamma})\} = &-\frac{1}{2}\sum_{j=1}^{M} \operatorname{tr}\left(\boldsymbol{X}_{\gamma}^{T}\boldsymbol{X}_{\gamma}\left[\boldsymbol{\Sigma}_{\beta(s_j)} + \boldsymbol{\mu}_{\beta(s_j)}\boldsymbol{\mu}_{\beta(s_j)}^{T}\right]\right) - \frac{N}{2}\sum_{j=1}^{M}\mathbb{E}_{q(\beta_0)}\left[\beta_0(s_j)^2\right] \\
&- \sum_{i=1}^{N}\sum_{j=1}^{M}\mathbb{E}_{q(\beta_0)}[\beta_0(s_j)]\boldsymbol{x}_{\gamma(s_j)}^{T}\boldsymbol{\mu}_{\beta(s_j)} - \sum_{j=1}^{M}\frac{Q(s_j)}{2}\ln\{\nu_1\} \\
&- \frac{1}{2}\sum_{j=1}^{M}\operatorname{tr}\left(\operatorname{diag}\left\{\frac{1}{\nu_1}\right\}_{q=1}^{Q(s_j)}\left[\boldsymbol{\Sigma}_{\beta(s_j)} + \boldsymbol{\mu}_{\beta(s_j)}\boldsymbol{\mu}_{\beta(s_j)}^{T}\right]\right) \\
&- \frac{M}{2}\ln\left\{\sigma_{\beta_0}^{2}\right\} - \frac{1}{2\sigma_{\beta_0}^{2}}\sum_{j=1}^{M}\left[\sigma_{\beta_0(s_j)}^{2} + \mu_{\beta_0(s_j)}^{2}\right] \\
&+ \sum_{j=1}^{M}\sum_{p=1}^{P}\{\gamma_p(s_j)\ln\{\sigma(\theta_p(s_j))\} + (1-\gamma_p(s_j))\ln\{1-\sigma(\theta_p(s_j))\}\} \\
&- \frac{MP}{2}\ln\{2\pi\} + \frac{M}{2}\ln\left\{|\boldsymbol{\Sigma}^{-1}|\right\} - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\left[\sum_{s_i\sim s_j}[\boldsymbol{\theta}(s_i)-\boldsymbol{\theta}(s_j)][\boldsymbol{\theta}(s_i)-\boldsymbol{\theta}(s_j)]^{T}\right]\right) \\
&+ \frac{\nu-P-1}{2}\ln\{|\boldsymbol{\Sigma}^{-1}|\} - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{I}\boldsymbol{\Sigma}^{-1}\right) - \frac{\nu P}{2}\ln\{2\} - \frac{P(P-1)}{4}\ln\{\pi\} \\
&- \sum_{p=1}^{P}\ln\left\{\Gamma\left(\frac{\nu+1-p}{2}\right)\right\} - \frac{\nu}{2}\ln\{P\} + \frac{1}{2}\sum_{j=1}^{M}\boldsymbol{\eta}(s_j)^{T}\boldsymbol{\eta}(s_j) \\
&+ \sum_{i=1}^{N}\sum_{j=1}^{M}\{y_i(s_j)\ln\{1-\Phi(-\eta_i(s_j))\} + (1-y_i(s_j))\ln\{\Phi(-\eta_i(s_j))\}\} \\
&+ \frac{1}{2}\ln\left\{|\boldsymbol{\Sigma}_{\beta(s_j)}|\right\} + \sum_{j=1}^{M}\frac{Q(s_j)}{2} + \frac{1}{2}\sum_{j=1}^{M}\ln\{\sigma_{\beta_0(s_j)}^{2}\} + \frac{M}{2}
\end{aligned}
$$

A.1.5 Relevant Expectations

A.1.5.1 $\mathbb{E}_{q(z)}[z_i(s_j)]$

$$\mathbb{E}_{q(z)}[z_i(s_j)] = \begin{cases} \eta_i(s_j) + \frac{\phi(-\eta_i(s_j))}{1-\Phi(-\eta_i(s_j))}, & \text{if } y_i(s_j) = 1, \\ \eta_i(s_j) - \frac{\phi(-\eta_i(s_j))}{\Phi(-\eta_i(s_j))}, & \text{if } y_i(s_j) = 0. \end{cases}$$

A.1.5.2 $\mathbb{E}_{q(\gamma)}\left[\ln\left\{\nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j)\right\}\right]$

$$\begin{aligned} \mathbb{E}_{q(\gamma)}\left[\ln\left\{\nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j)\right\}\right] &= \sum_{\gamma_k} q(\gamma_k)\ln\{\nu_0(1-\gamma_k) + \nu_1\gamma_k\} \\ &= q(\gamma_k = 1)\ln\{\nu_1\} + q(\gamma_k = 0)\ln\{\nu_0\} \\ &= \mathbb{E}_{q(\gamma)}[\gamma_p(s_j)]\ln\{\nu_1\} + (1 - \mathbb{E}_{q(\gamma)}[\gamma_p(s_j)])\ln\{\nu_0\} \\ &= \mu_{\gamma_p(s_j)}\ln\{\nu_1\} + (1 - \mu_{\gamma_p(s_j)})\ln\{\nu_0\} \end{aligned}$$

A.1.5.3 $\mathbb{E}_{q(\gamma)}\left[\frac{1}{\nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j)}\right]$

$$\begin{aligned} \mathbb{E}_{q(\gamma)}\left[\frac{1}{\nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j)}\right] &= \sum_{\gamma_k} q(\gamma_k)\frac{1}{\nu_0(1-\gamma_k) + \nu_1\gamma_k} \\ &= \frac{q(\gamma_k = 1)}{\nu_1} + \frac{q(\gamma_k = 0)}{\nu_0} \\ &= \frac{\mathbb{E}_{q(\gamma)}[\gamma_p(s_j)]}{\nu_1} + \frac{1 - \mathbb{E}_{q(\gamma)}[\gamma_p(s_j)]}{\nu_0} \\ &= \frac{\mu_{\gamma_p(s_j)}}{\nu_1} + \frac{1 - \mu_{\gamma_p(s_j)}}{\nu_0} \end{aligned}$$

A.1.6 Initialization and Convergence of Variational Approximation

The ELBO continuously increases until the algorithm converges to a local optimum when the difference in ELBO between two consecutive steps reaches a pre-determined convergence threshold ϵ. It should be noted that variational inference is also sensitive to parameter initialization considering the highly non-convex nature of the optimization problem. Hence, we choose to initialize $\boldsymbol{\beta}$ and $\boldsymbol{\beta_0}$ with estimates obtained by the mass-univariate approach Firth regression, $\boldsymbol{\gamma}$ with the fixed value 0.5, $\boldsymbol{\theta}$ with the logit transformed fraction of active voxels for every covariate from Firth regression, and $\boldsymbol{\Sigma}^{-1}$ with an identity matrix in order to ensure convergence to a local optimum.

A.2 Further UK Biobank Results

A.2.1 UKBB Analysis with Sample Size: N=40,000

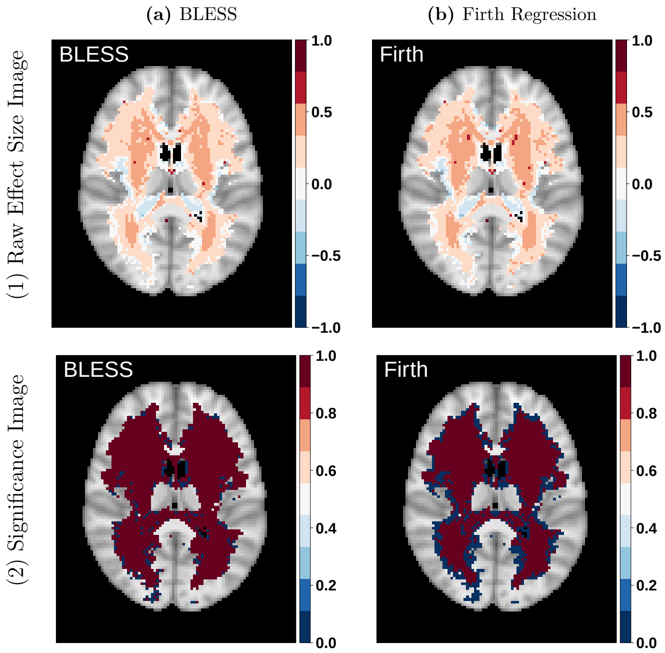

(a) BLESS **(b)** Firth Regression

Figure A.1: Comparison of results between (a) BLESS and (b) Firth Regression. (1) spatially varying coefficient maps of the covariate age. (2) Thresholded significance maps of the covariate age where the threshold for BLESS is determined via the probability of inclusion/exclusion $P(\gamma_p(s_j)|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) \geq 0.5$ and the threshold for Firth regression via the test statistic $t = |\hat{\beta}/\hat{\sigma}_{\hat{\beta}}| \geq 1.96$ (significant voxels are red, p-values have been adjusted via FDR-correction at 5%).

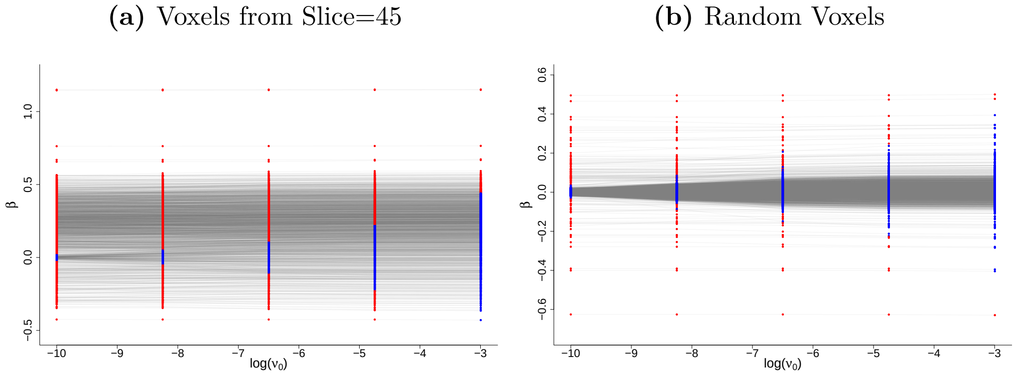

(a) Voxels from Slice=45 **(b)** Random Voxels

Figure A.2: Regularisation plots of estimated parameters from voxels from slice z = 45 and from a set of random voxels across the 3D image for the UKBB analysis of lesion incidence regressed on age, sex, age by sex, and head size with a sample size of 40,000 subjects.

A.3 Simulation Study: Varying Base Rate Intensities and Sample Sizes

	$\boldsymbol{\lambda = 1}$			$\boldsymbol{\lambda = 2}$			$\boldsymbol{\lambda = 3}$		
N=500	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
Truth	-2.0868	0.5882	0.5586	-1.7933	0.6559	0.6237	-1.6089	0.7051	0.6685
BLESS	-2.0564	0.4799	0.3941	-1.7836	0.6202	0.5482	-1.6036	0.6910	0.6231
BSGLMM	-2.1314	0.6027	0.5732	-1.8168	0.6590	0.6306	-1.6233	0.7058	0.6661
Firth	-2.0864	0.5840	0.5599	-1.7925	0.6494	0.6243	-1.6059	0.6993	0.6625
N=1,000	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
Truth	-2.0868	0.5882	0.5586	-1.7933	0.6559	0.6237	-1.6089	0.7051	0.6685
BLESS	-2.0852	0.5725	0.5088	-1.7952	0.6520	0.6113	-1.6083	0.6982	0.6573
BSGLMM	-2.1066	0.5918	0.5547	-1.8056	0.6559	0.6238	-1.6182	0.7013	0.6668
Firth	-2.0836	0.5837	0.5492	-1.7925	0.6518	0.6219	-1.6091	0.6988	0.6664
N=5,000	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
Truth	-2.0868	0.5882	0.5586	-1.7933	0.6559	0.6237	-1.6089	0.7051	0.6685
BLESS	-2.0874	0.5826	0.5555	-1.7933	0.6516	0.6200	-1.6060	0.6982	0.6653
BSGLMM	-2.0913	0.5836	0.5573	-1.7965	0.6525	0.6222	-1.6087	0.6990	0.6672
Firth	-2.0861	0.5826	0.5574	-1.7934	0.6523	0.6228	-1.6065	0.6991	0.6679

Table A.1: Comparison of parameter estimates from the methods, BLESS, BSGLMM and Firth Regression with true coefficient values.

Parameter Estimate: $\hat{\beta}_0$	**Bias**			**Variance**			**MSE**		
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	0.0242	0.0014	-0.0043	0.0020	0.0020	0.0020	0.0026	0.0020	0.0020
BSGLMM	-0.0464	-0.0243	-0.0160	0.0208	0.0131	0.0104	0.0229	0.0136	0.0106
Firth	-0.0032	-0.0001	0.0010	0.0523	0.0310	0.0238	0.0523	0.0310	0.0238
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	-0.0054	-0.0090	-0.0059	0.0010	0.0010	0.0010	0.0010	0.0011	0.0010
BSGLMM	-0.0206	-0.0140	-0.0106	0.0102	0.0067	0.0054	0.0106	0.0069	0.0055
Firth	0.0020	-0.0014	-0.0018	0.0246	0.0152	0.0118	0.0246	0.0152	0.0118
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	-0.0038	-0.0030	-0.0004	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
BSGLMM	-0.0056	-0.0044	-0.0015	0.0023	0.0016	0.0013	0.0023	0.0016	0.0013
Firth	-0.0006	-0.0015	0.0005	0.0048	0.0030	0.0023	0.0048	0.0030	0.0023

Parameter Estimate: $\hat{\beta}_1$	**Bias**			**Variance**			**MSE**		
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	-0.0961	-0.0237	-0.0009	0.0014	0.0019	0.0020	0.0106	0.0024	0.0020
BSGLMM	0.0280	0.0129	0.0130	0.0117	0.0080	0.0067	0.0125	0.0082	0.0068
Firth	0.0068	-0.0024	0.0017	0.0562	0.0348	0.0272	0.0563	0.0348	0.0272
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	-0.0031	0.0082	0.0019	0.0010	0.0010	0.0010	0.0010	0.0011	0.0010
BSGLMM	0.0127	0.0106	0.0066	0.0063	0.0045	0.0039	0.0064	0.0046	0.0039
Firth	-0.0002	0.0026	0.0005	0.0271	0.0171	0.0135	0.0271	0.0171	0.0135
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	0.0032	0.0039	-0.0011	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
BSGLMM	0.0057	0.0054	-0.0006	0.0018	0.0014	0.0012	0.0018	0.0014	0.0012
Firth	0.0022	0.0031	-0.0023	0.0053	0.0034	0.0027	0.0053	0.0034	0.0027

Table A.2: Evaluation of parameter estimates from the methods, BLESS, BSGLMM and Firth Regression via bias, variance and MSE of the spatially varying coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$.

Parameter Estimate: $\hat{\boldsymbol{\beta}}_2$	**Bias**			**Variance**			**MSE**		
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	-0.1661	-0.0584	-0.0284	0.0006	0.0009	0.0009	0.0282	0.0043	0.0018
BSGLMM	0.0168	0.0072	-0.0053	0.0110	0.0076	0.0064	0.0113	0.0077	0.0064
Firth	0.0040	-0.0009	-0.0102	0.0535	0.0335	0.0263	0.0535	0.0335	0.0265
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	-0.0373	0.0058	0.0013	0.0005	0.0005	0.0005	0.0019	0.0005	0.0005
BSGLMM	-0.0034	0.0051	0.0005	0.0059	0.0043	0.0037	0.0059	0.0044	0.0037
Firth	-0.0104	0.0025	-0.0003	0.0258	0.0165	0.0131	0.0259	0.0165	0.0131
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	-0.0002	0.0014	-0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
BSGLMM	-0.0032	-0.0002	-0.0010	0.0017	0.0013	0.0012	0.0017	0.0013	0.0012
Firth	-0.0034	0.0002	-0.0003	0.0051	0.0033	0.0026	0.0051	0.0033	0.0026

Predictive Performance: $\hat{\boldsymbol{y}}$	**Bias**			**Variance**			**MSE**		
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	-0.0078	-0.0052	-0.0032	0.0011	0.0017	0.0018	0.0022	0.0031	0.0034
BSGLMM	-0.0027	-0.0025	-0.0020	0.0002	0.0004	0.0007	0.0002	0.0004	0.0007
Firth	0.0170	0.0140	0.0122	0.0009	0.0016	0.0022	0.0018	0.0032	0.0043
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	-0.0015	0.0007	-0.0013	0.0004	0.0005	0.0007	0.0008	0.0010	0.0012
BSGLMM	-0.0010	-0.0018	-0.0020	0.0001	0.0002	0.0003	0.0001	0.0002	0.0003
Firth	0.0082	0.0082	0.0056	0.0005	0.0008	0.0011	0.0009	0.0016	0.0021
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	-0.0006	-0.0000	0.0008	0.0001	0.0001	0.0001	0.0001	0.0002	0.0003
BSGLMM	-0.0004	-0.0008	-0.0001	0.0000	0.0001	0.0001	0.0000	0.0001	0.0001
Firth	0.0010	0.0015	0.0020	0.0001	0.0002	0.0002	0.0002	0.0003	0.0004

Table A.3: Evaluation of parameter estimates $\hat{\beta}_2$ and the predictive performance $\hat{y}$ from the methods, BLESS, BSGLMM and Firth Regression via bias, variance and MSE.

		TPR			**TDR**	
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	0.6635	0.9080	0.9770	0.9520	0.9808	0.9893
BSGLMM (uncorrected)	0.9983	1.0000	1.0000	0.8498	0.8617	0.8621
BSGLMM (corrected)	0.9972	1.0000	1.0000	0.8829	0.8941	0.8949
Firth (uncorrected)	0.8258	0.9822	0.9984	0.9553	0.9565	0.9550
Firth (corrected)	0.6600	0.9629	0.9963	0.8911	0.9368	0.9489
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	0.9458	0.9983	1.0000	0.9713	0.9882	0.9936
BSGLMM (uncorrected)	1.0000	1.0000	1.0000	0.8578	0.8668	0.8773
BSGLMM (corrected)	1.0000	1.0000	1.0000	0.8906	0.8993	0.9092
Firth (uncorrected)	0.9876	0.9999	1.0000	0.9577	0.9553	0.9554
Firth (corrected)	0.9738	0.9999	1.0000	0.9406	0.9607	0.9644
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	1.0000	1.0000	1.0000	0.9873	0.9973	0.9968
BSGLMM (uncorrected)	1.0000	1.0000	1.0000	0.8776	0.8956	0.8982
BSGLMM (corrected)	1.0000	1.0000	1.0000	0.9084	0.9252	0.9272
Firth (uncorrected)	1.0000	1.0000	1.0000	0.9516	0.9548	0.9515
Firth (corrected)	1.0000	1.0000	1.0000	0.9702	0.9743	0.9716
		FPR			**FDR**	
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	0.0334	0.0178	0.0106	0.0480	0.0192	0.0107
BSGLMM (uncorrected)	0.1783	0.1619	0.1616	0.1502	0.1383	0.1379
BSGLMM (corrected)	0.1340	0.1196	0.1188	0.1171	0.1059	0.1051
Firth (uncorrected)	0.0388	0.0449	0.0472	0.0447	0.0435	0.0450
Firth (corrected)	0.0802	0.0652	0.0540	0.1089	0.0632	0.0511
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	0.0282	0.0120	0.0065	0.0287	0.0118	0.0064
BSGLMM (uncorrected)	0.1674	0.1551	0.1410	0.1422	0.1332	0.1227
BSGLMM (corrected)	0.1243	0.1132	0.1009	0.1094	0.1007	0.0908
Firth (uncorrected)	0.0438	0.0470	0.0469	0.0423	0.0447	0.0446
Firth (corrected)	0.0619	0.0413	0.0372	0.0594	0.0393	0.0356
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	0.0129	0.0027	0.0032	0.0127	0.0027	0.0032
BSGLMM (uncorrected)	0.1407	0.1175	0.1141	0.1224	0.1044	0.1018
BSGLMM (corrected)	0.1019	0.0815	0.0790	0.0916	0.0748	0.0728
Firth (uncorrected)	0.0512	0.0476	0.0511	0.0484	0.0452	0.0485
Firth (corrected)	0.0309	0.0265	0.0293	0.0298	0.0257	0.0284

Table A.4: Evaluation of inference results from the methods, BLESS, BSGLMM and Firth Regression via TPR, TDR, FPR and FDR for parameter estiamte $\hat{\beta}_1$.

	TPR			**TDR**		
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	0.5448	0.8336	0.9344	0.9288	0.9805	0.9930
BSGLMM (uncorrected)	0.9981	1.0000	1.0000	0.6659	0.6777	0.6889
BSGLMM (corrected)	0.9951	1.0000	1.0000	0.7871	0.7969	0.8094
Firth (uncorrected)	0.8260	0.9840	0.9978	0.8791	0.8800	0.8801
Firth (corrected)	0.4798	0.9357	0.9897	0.6342	0.8321	0.8670
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	0.8888	0.9942	0.9998	0.9683	0.9909	0.9957
BSGLMM (uncorrected)	1.0000	1.0000	1.0000	0.6820	0.6883	0.7078
BSGLMM (corrected)	1.0000	1.0000	1.0000	0.8037	0.8078	0.8221
Firth (uncorrected)	0.9848	1.0000	1.0000	0.8862	0.8726	0.8733
Firth (corrected)	0.9388	0.9993	1.0000	0.8417	0.8937	0.9134
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	1.0000	1.0000	1.0000	0.9851	0.9970	0.9984
BSGLMM (uncorrected)	1.0000	1.0000	1.0000	0.7222	0.7394	0.7599
BSGLMM (corrected)	1.0000	1.0000	1.0000	0.8400	0.8594	0.8763
Firth (uncorrected)	1.0000	1.0000	1.0000	0.8723	0.8710	0.8727
Firth (corrected)	1.0000	1.0000	1.0000	0.9499	0.9556	0.9565
	FPR			**FDR**		
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	0.0139	0.0056	0.0022	0.0712	0.0195	0.0070
BSGLMM (uncorrected)	0.1692	0.1610	0.1523	0.3341	0.3223	0.3111
BSGLMM (corrected)	0.0914	0.0870	0.0800	0.2129	0.2031	0.1906
Firth (uncorrected)	0.0382	0.0451	0.0456	0.1209	0.1200	0.1199
Firth (corrected)	0.0908	0.0636	0.0515	0.3658	0.1679	0.1330
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	0.0098	0.0031	0.0014	0.0317	0.0091	0.0043
BSGLMM (uncorrected)	0.1579	0.1527	0.1396	0.3180	0.3117	0.2922
BSGLMM (corrected)	0.0833	0.0807	0.0736	0.1963	0.1922	0.1779
Firth (uncorrected)	0.0425	0.0490	0.0488	0.1138	0.1274	0.1267
Firth (corrected)	0.0600	0.0409	0.0324	0.1583	0.1063	0.0866
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS	0.0051	0.0010	0.0005	0.0149	0.0030	0.0016
BSGLMM (uncorrected)	0.1298	0.1188	0.1066	0.2778	0.2606	0.2401
BSGLMM (corrected)	0.0646	0.0554	0.0477	0.1600	0.1406	0.1237
Firth (uncorrected)	0.0492	0.0497	0.0490	0.1277	0.1290	0.1273
Firth (corrected)	0.0179	0.0157	0.0153	0.0501	0.0444	0.0435

Table A.5: Evaluation of inference results from the methods, BLESS, BSGLMM and Firth Regression via TPR, TDR, FPR and FDR for parameter estiamte $\hat{\beta}_2$.

A.4 Simulation Study: Varying Spatial Priors

	$\boldsymbol{\lambda = 1}$			$\boldsymbol{\lambda = 2}$			$\boldsymbol{\lambda = 3}$		
N=500	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
Truth	-2.0868	0.5882	0.5586	-1.7933	0.6559	0.6237	-1.6089	0.7051	0.6685
BLESS (CAR)	-1.8840	0.1218	0.1155	-1.6392	0.2978	0.2811	-1.5691	0.6123	0.5579
BLESS (PxCAR)	-2.0216	0.3912	0.3712	-1.7546	0.5469	0.5126	-1.5947	0.6693	0.6126
BLESS (MCAR)	-2.0564	0.4799	0.3941	-1.7836	0.6202	0.5482	-1.6036	0.6910	0.6231
N=1,000	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
Truth	-2.0868	0.5882	0.5586	-1.7933	0.6559	0.6237	-1.6089	0.7051	0.6685
BLESS (CAR)	-1.9952	0.3810	0.3368	-1.7900	0.6428	0.6013	-1.6061	0.6948	0.6529
BLESS (PxCAR)	-2.0559	0.5115	0.4555	-1.7916	0.6461	0.6051	-1.6065	0.6954	0.6542
BLESS (MCAR)	-2.0852	0.5725	0.5088	-1.7952	0.6520	0.6113	-1.6083	0.6982	0.6573
N=5,000	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
Truth	-2.0868	0.5882	0.5586	-1.7933	0.6559	0.6237	-1.6089	0.7051	0.6685
BLESS (CAR)	-2.0858	0.5807	0.5514	-1.7928	0.6508	0.6183	-1.6055	0.6977	0.6637
BLESS (PxCAR)	-2.0859	0.5809	0.5520	-1.7928	0.6509	0.6186	-1.6056	0.6977	0.6639
BLESS (MCAR)	-2.0874	0.5826	0.5555	-1.7933	0.6516	0.6200	-1.6060	0.6982	0.6653

Table A.6: Comparison of parameter estimates from the methods, BLESS with three different types of spatial prior, specifically three formulations of a conditional autoregressive prior, BSGLMM and Firth Regression with true coefficient values.

Parameter Estimate: $\hat{\boldsymbol{\beta}}_0$	**Bias**			**Variance**			**MSE**		
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (CAR)	0.2076	0.1523	0.0283	-0.4831	-0.3658	-0.0789	-0.4855	-0.3560	-0.0934
BLESS (PxCAR)	0.0601	0.0294	0.0021	-0.1909	-0.0983	-0.0194	-0.1925	-0.0943	-0.0344
BLESS (MCAR)	0.0242	0.0014	-0.0043	-0.0961	-0.0237	-0.0009	-0.1661	-0.0584	-0.0284
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (CAR)	0.0862	-0.0065	-0.0055	-0.2035	0.0027	0.0013	-0.2238	0.0016	0.0010
BLESS (PxCAR)	0.0217	-0.0078	-0.0056	-0.0641	0.0057	0.0014	-0.0884	0.0045	0.0011
BLESS (MCAR)	-0.0054	-0.0090	-0.0059	-0.0031	0.0082	0.0019	-0.0373	0.0058	0.0013
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (CAR)	-0.0036	-0.0030	-0.0004	0.0030	0.0038	-0.0011	-0.0002	0.0014	-0.0002
BLESS (PxCAR)	-0.0036	-0.0030	-0.0004	0.0030	0.0038	-0.0011	-0.0002	0.0014	-0.0002
BLESS (MCAR)	-0.0038	-0.0030	-0.0004	0.0032	0.0039	-0.0011	-0.0002	0.0014	-0.0002

Parameter Estimate: $\hat{\boldsymbol{\beta}}_1$	**Bias**			**Variance**			**MSE**		
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (CAR)	0.0020	0.0020	0.0020	0.0002	0.0007	0.0017	0.0001	0.0004	0.0008
BLESS (PxCAR)	0.0020	0.0020	0.0020	0.0010	0.0015	0.0019	0.0005	0.0008	0.0009
BLESS (MCAR)	0.0020	0.0020	0.0020	0.0014	0.0019	0.0020	0.0006	0.0009	0.0009
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (CAR)	0.0010	0.0010	0.0010	0.0006	0.0010	0.0010	0.0003	0.0005	0.0005
BLESS (PxCAR)	0.0010	0.0010	0.0010	0.0008	0.0010	0.0010	0.0004	0.0005	0.0005
BLESS (MCAR)	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	0.0005	0.0005	0.0005
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (CAR)	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001
BLESS (PxCAR)	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001
BLESS (MCAR)	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001

Table A.7: Evaluation of parameter estimates from the methods, BLESS, BSGLMM and Firth Regression via bias, variance and MSE of the spatially varying coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$.

Parameter Estimate: $\hat{\boldsymbol{\beta}}_2$	**Bias**			**Variance**			**MSE**		
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (CAR)	0.0451	0.0252	0.0028	0.2336	0.1345	0.0079	0.2358	0.1271	0.0096
BLESS (PxCAR)	0.0056	0.0029	0.0020	0.0375	0.0112	0.0023	0.0376	0.0097	0.0021
BLESS (MCAR)	0.0026	0.0020	0.0020	0.0106	0.0024	0.0020	0.0282	0.0043	0.0018
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (CAR)	0.0084	0.0010	0.0010	0.0420	0.0010	0.0010	0.0504	0.0005	0.0005
BLESS (PxCAR)	0.0015	0.0011	0.0010	0.0049	0.0010	0.0010	0.0082	0.0005	0.0005
BLESS (MCAR)	0.0010	0.0011	0.0010	0.0010	0.0011	0.0010	0.0019	0.0005	0.0005
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (CAR)	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001
BLESS (PxCAR)	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001
BLESS (MCAR)	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001

Predictive Performance: $\hat{\boldsymbol{y}}$	**Bias**			**Variance**			**MSE**		
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (CAR)	-0.0229	-0.0296	-0.0103	0.0010	0.0034	0.0032	0.0037	0.0090	0.0062
BLESS (PxCAR)	0.0297	0.0239	0.0092	0.0010	0.0017	0.0014	0.0012	0.0019	0.0014
BLESS (MCAR)	-0.0078	-0.0052	-0.0032	0.0011	0.0017	0.0018	0.0022	0.0031	0.0034
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (CAR)	-0.0103	0.0003	-0.0014	0.0008	0.0006	0.0007	0.0019	0.0011	0.0012
BLESS (PxCAR)	0.0110	0.0006	0.0002	0.0004	0.0003	0.0004	0.0005	0.0003	0.0004
BLESS (MCAR)	-0.0015	0.0007	-0.0013	0.0004	0.0005	0.0007	0.0008	0.0010	0.0012
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (CAR) -0.0007	-0.0000	0.0007	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	
BLESS (PxCAR)	-0.0001	-0.0004	0.0003	0.0000	0.0000	0.0001	0.0000	0.0000	0.0001
BLESS (MCAR)	-0.0006	-0.0000	0.0008	0.0001	0.0001	0.0001	0.0001	0.0002	0.0003

Table A.8: Evaluation of parameter estimates $\hat{\beta}_2$ and the predictive performance $\hat{y}$ from the methods, BLESS, BSGLMM and Firth Regression via bias, variance and MSE.

		TPR			**TDR**	
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (CAR)	0.1193	0.3630	0.8372	0.9867	0.9997	0.9993
BLESS (PxCAR)	0.5060	0.7654	0.9395	0.9714	0.9953	0.9987
BLESS (MCAR)	0.6635	0.9080	0.9770	0.9520	0.9808	0.9893
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (CAR)	0.5667	0.9850	0.9994	0.9987	0.9997	0.9997
BLESS (PxCAR)	0.8111	0.9923	0.9997	0.9970	0.9997	0.9998
BLESS (MCAR)	0.9458	0.9983	1.0000	0.9713	0.9882	0.9936
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (CAR)	1.0000	1.0000	1.0000	0.9997	0.9999	0.9999
BLESS (PxCAR)	1.0000	1.0000	1.0000	0.9996	0.9999	0.9999
BLESS (MCAR)	1.0000	1.0000	1.0000	0.9873	0.9973	0.9968
		FPR			**FDR**	
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (CAR)	0.0014	0.0001	0.0006	0.0133	0.0003	0.0007
BLESS (PxCAR)	0.0150	0.0036	0.0012	0.0286	0.0047	0.0013
BLESS (MCAR)	0.0334	0.0178	0.0106	0.0480	0.0192	0.0107
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (CAR)	0.0010	0.0003	0.0003	0.0013	0.0003	0.0003
BLESS (PxCAR)	0.0024	0.0003	0.0002	0.0030	0.0003	0.0002
BLESS (MCAR)	0.0282	0.0120	0.0065	0.0287	0.0118	0.0064
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (CAR)	0.0003	0.0001	0.0001	0.0003	0.0001	0.0001
BLESS (PxCAR)	0.0004	0.0001	0.0001	0.0004	0.0001	0.0001
BLESS (MCAR)	0.0129	0.0027	0.0032	0.0127	0.0027	0.0032

Table A.9: Evaluation of inference results from the methods, BLESS, BSGLMM and Firth Regression via TPR, TDR, FPR and FDR for parameter estiamte $\hat{\beta}_1$.

	TPR			**TDR**		
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (CAR)	0.1192	0.3757	0.8221	0.9718	0.9995	0.9983
BLESS (PxCAR)	0.5064	0.7681	0.9227	0.9352	0.9893	0.9974
BLESS (MCAR)	0.5448	0.8336	0.9344	0.9288	0.9805	0.9930
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (CAR)	0.5446	0.9845	0.9995	0.9972	0.9989	0.9994
BLESS (PxCAR)	0.7805	0.9912	0.9998	0.9946	0.9991	0.9995
BLESS (MCAR)	0.8888	0.9942	0.9998	0.9683	0.9909	0.9957
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (CAR)	1.0000	1.0000	1.0000	0.9994	0.9997	0.9999
BLESS (PxCAR)	1.0000	1.0000	1.0000	0.9994	0.9997	0.9999
BLESS (MCAR)	1.0000	1.0000	1.0000	0.9851	0.9970	0.9984
	FPR			**FDR**		
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (CAR)	0.0011	0.0001	0.0005	0.0282	0.0005	0.0017
BLESS (PxCAR)	0.0117	0.0028	0.0008	0.0648	0.0107	0.0026
BLESS (MCAR)	0.0139	0.0056	0.0022	0.0712	0.0195	0.0070
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (CAR)	0.0006	0.0003	0.0002	0.0028	0.0011	0.0006
BLESS (PxCAR)	0.0014	0.0003	0.0002	0.0054	0.0009	0.0005
BLESS (MCAR)	0.0098	0.0031	0.0014	0.0317	0.0091	0.0043
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (CAR)	0.0002	0.0001	0.0000	0.0006	0.0003	0.0001
BLESS (PxCAR)	0.0002	0.0001	0.0000	0.0006	0.0003	0.0001
BLESS (MCAR)	0.0051	0.0010	0.0005	0.0149	0.0030	0.0016

Table A.10: Evaluation of inference results from the methods, BLESS, BSGLMM and Firth Regression via TPR, TDR, FPR and FDR for parameter estiamte $\hat{\beta}_2$.

A.5 Simulation Study: Varying Slab Variances

	$\boldsymbol{\lambda = 1}$			$\boldsymbol{\lambda = 2}$			$\boldsymbol{\lambda = 3}$		
N=500	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
Truth	-2.0868	0.5882	0.5586	-1.7933	0.6559	0.6237	-1.6089	0.7051	0.6685
BLESS ($\nu_1 = 1$)	-2.0737	0.5197	0.4479	-1.7905	0.6287	0.5780	-1.6036	0.6852	0.6340
BLESS ($\nu_1 = 10$)	-2.0564	0.4799	0.3941	-1.7836	0.6202	0.5482	-1.6036	0.6910	0.6231
BLESS ($\nu_1 = 100$)	-1.9736	0.3059	0.2429	-1.7480	0.5535	0.4576	-1.5882	0.6662	0.5755
N=1,000	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
Truth	-2.0868	0.5882	0.5586	-1.7933	0.6559	0.6237	-1.6089	0.7051	0.6685
BLESS ($\nu_1 = 1$)	-2.0894	0.5752	0.5271	-1.7932	0.6451	0.6099	-1.6067	0.6920	0.6548
BLESS ($\nu_1 = 10$)	-2.0852	0.5725	0.5088	-1.7952	0.6520	0.6113	-1.6083	0.6982	0.6573
BLESS ($\nu_1 = 100$)	-2.0859	0.5799	0.5501	-1.7911	0.6470	0.5993	-1.6061	0.6963	0.6511
N=5,000	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
Truth	-2.0868	0.5882	0.5586	-1.7933	0.6559	0.6237	-1.6089	0.7051	0.6685
BLESS ($\nu_1 = 1$)	-2.0866	0.5806	0.5539	-1.7930	0.6504	0.6197	-1.6057	0.6972	0.6651
BLESS ($\nu_1 = 10$)	-2.0874	0.5826	0.5555	-1.7933	0.6516	0.6200	-1.6060	0.6982	0.6653
BLESS ($\nu_1 = 100$)	-2.0870	0.5822	0.5546	-1.7929	0.6512	0.6191	-1.6056	0.6978	0.6639

Table A.11: Comparison of parameter estimates from the methods, BLESS with three different types of spatial prior, specifically three formulations of a conditional autoregressive prior, BSGLMM and Firth Regression with true coefficient values.

Parameter Estimate: $\hat{\boldsymbol{\beta}}_0$	**Bias**			**Variance**			**MSE**		
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS ($\nu_1 = 1$)	0.0077	-0.0030	-0.0014	-0.0552	-0.0174	-0.0104	-0.1074	-0.0324	-0.0236
BLESS ($\nu_1 = 10$)	0.0242	0.0014	-0.0043	-0.0961	-0.0237	-0.0009	-0.1661	-0.0584	-0.0284
BLESS ($\nu_1 = 100$)	0.1111	0.0371	0.0095	-0.2824	-0.0932	-0.0247	-0.3391	-0.1561	-0.0758
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS ($\nu_1 = 1$)	-0.0070	-0.0047	-0.0022	-0.0029	-0.0014	-0.0057	-0.0229	-0.0025	-0.0068
BLESS ($\nu_1 = 10$)	-0.0054	-0.0090	-0.0059	-0.0031	0.0082	0.0019	-0.0373	0.0058	0.0013
BLESS ($\nu_1 = 100$)	0.0088	-0.0072	-0.0059	-0.0287	0.0057	0.0022	-0.0783	-0.0012	0.0011
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS ($\nu_1 = 1$)	-0.0024	-0.0020	0.0004	0.0006	0.0020	-0.0028	-0.0030	-0.0007	-0.0019
BLESS ($\nu_1 = 10$)	-0.0038	-0.0030	-0.0004	0.0032	0.0039	-0.0011	-0.0002	0.0014	-0.0002
BLESS ($\nu_1 = 100$)	-0.0039	-0.0031	-0.0005	0.0034	0.0040	-0.0009	0.0001	0.0016	-0.0000

Parameter Estimate: $\hat{\boldsymbol{\beta}}_1$	**Bias**			**Variance**			**MSE**		
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS ($\nu_1 = 1$)	0.0020	0.0020	0.0020	0.0018	0.0021	0.0021	0.0008	0.0010	0.0010
BLESS ($\nu_1 = 10$)	0.0020	0.0020	0.0020	0.0014	0.0019	0.0020	0.0006	0.0009	0.0009
BLESS ($\nu_1 = 100$)	0.0020	0.0020	0.0020	0.0007	0.0016	0.0019	0.0003	0.0007	0.0008
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS ($\nu_1 = 1$)	0.0010	0.0010	0.0010	0.0011	0.0011	0.0010	0.0005	0.0005	0.0005
BLESS ($\nu_1 = 10$)	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	0.0005	0.0005	0.0005
BLESS ($\nu_1 = 100$)	0.0010	0.0010	0.0010	0.0009	0.0010	0.0010	0.0004	0.0005	0.0005
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS ($\nu_1 = 1$)	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001
BLESS ($\nu_1 = 10$)	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001
BLESS ($\nu_1 = 100$)	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001

Table A.12: Evaluation of parameter estimates from the methods, BLESS, BSGLMM and Firth Regression via bias, variance and MSE of the spatially varying coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$.

Parameter Estimate: $\hat{\boldsymbol{\beta}}_2$	**Bias**			**Variance**			**MSE**		
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS ($\nu_1 = 1$)	0.0021	0.0020	0.0020	0.0048	0.0024	0.0022	0.0124	0.0020	0.0016
BLESS ($\nu_1 = 10$)	0.0026	0.0020	0.0020	0.0106	0.0024	0.0020	0.0282	0.0043	0.0018
BLESS ($\nu_1 = 100$)	0.0144	0.0034	0.0021	0.0805	0.0102	0.0025	0.1153	0.0250	0.0066
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS ($\nu_1 = 1$)	0.0010	0.0010	0.0010	0.0011	0.0011	0.0011	0.0011	0.0005	0.0006
BLESS ($\nu_1 = 10$)	0.0010	0.0011	0.0010	0.0010	0.0011	0.0010	0.0019	0.0005	0.0005
BLESS ($\nu_1 = 100$)	0.0011	0.0011	0.0010	0.0017	0.0010	0.0010	0.0065	0.0005	0.0005
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS ($\nu_1 = 1$)	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001
BLESS ($\nu_1 = 10$)	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001
BLESS ($\nu_1 = 100$)	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001

Predictive Performance: $\hat{\boldsymbol{y}}$	**Bias**			**Variance**			**MSE**		
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS ($\nu_1 = 1$)	-0.0054	-0.0042	-0.0039	0.0009	0.0012	0.0015	0.0017	0.0024	0.0029
BLESS ($\nu_1 = 10$)	0.0206	0.0107	0.0053	0.0008	0.0011	0.0011	0.0009	0.0012	0.0011
BLESS ($\nu_1 = 100$)	-0.0160	-0.0117	-0.0069	0.0012	0.0024	0.0025	0.0030	0.0048	0.0048
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS ($\nu_1 = 1$)	-0.0014	-0.0002	-0.0024	0.0003	0.0005	0.0007	0.0006	0.0010	0.0013
BLESS ($\nu_1 = 10$)	0.0037	0.0002	0.0001	0.0002	0.0003	0.0004	0.0002	0.0003	0.0004
BLESS ($\nu_1 = 100$)	-0.0031	0.0003	-0.0013	0.0005	0.0006	0.0007	0.0010	0.0011	0.0012
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS ($\nu_1 = 1$)	-0.0008	-0.0002	0.0005	0.0001	0.0001	0.0001	0.0001	0.0002	0.0003
BLESS ($\nu_1 = 10$)	-0.0001	-0.0004	0.0003	0.0000	0.0000	0.0001	0.0000	0.0000	0.0001
BLESS ($\nu_1 = 100$)	-0.0006	-0.0000	0.0008	0.0001	0.0001	0.0001	0.0001	0.0002	0.0003

Table A.13: Evaluation of parameter estimates $\hat{\beta}_2$ and the predictive performance $\hat{y}$ from the methods, BLESS, BSGLMM and Firth Regression via bias, variance and MSE.

	TPR			**TDR**		
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS ($\nu_1 = 1$)	0.8149	0.9709	0.9957	0.9181	0.9475	0.9615
BLESS ($\nu_1 = 10$)	0.6635	0.9080	0.9770	0.9520	0.9808	0.9893
BLESS ($\nu_1 = 100$)	0.3584	0.7728	0.9259	0.9750	0.9931	0.9969
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS ($\nu_1 = 1$)	0.9866	0.9998	1.0000	0.9228	0.9548	0.9676
BLESS ($\nu_1 = 10$)	0.9458	0.9983	1.0000	0.9713	0.9882	0.9936
BLESS ($\nu_1 = 100$)	0.8815	0.9893	0.9994	0.9861	0.9971	0.9991
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS ($\nu_1 = 1$)	1.0000	1.0000	1.0000	0.9744	0.9823	0.9852
BLESS ($\nu_1 = 10$)	1.0000	1.0000	1.0000	0.9873	0.9973	0.9968
BLESS ($\nu_1 = 100$)	1.0000	1.0000	1.0000	0.9936	0.9993	0.9996
	FPR			**FDR**		
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS ($\nu_1 = 1$)	0.0729	0.0541	0.0401	0.0819	0.0525	0.0385
BLESS ($\nu_1 = 10$)	0.0334	0.0178	0.0106	0.0480	0.0192	0.0107
BLESS ($\nu_1 = 100$)	0.0090	0.0054	0.0029	0.0250	0.0069	0.0031
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS ($\nu_1 = 1$)	0.0831	0.0476	0.0337	0.0772	0.0452	0.0324
BLESS ($\nu_1 = 10$)	0.0282	0.0120	0.0065	0.0287	0.0118	0.0064
BLESS ($\nu_1 = 100$)	0.0125	0.0029	0.0009	0.0139	0.0029	0.0009
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS ($\nu_1 = 1$)	0.0264	0.0181	0.0151	0.0256	0.0177	0.0148
BLESS ($\nu_1 = 10$)	0.0129	0.0027	0.0032	0.0127	0.0027	0.0032
BLESS ($\nu_1 = 100$)	0.0065	0.0007	0.0004	0.0064	0.0007	0.0004

Table A.14: Evaluation of inference results from the methods, BLESS, BSGLMM and Firth Regression via TPR, TDR, FPR and FDR for parameter estiamte $\hat{\beta}_1$.

		TPR			**TDR**	
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS ($\nu_1 = 1$)	0.7094	0.9280	0.9798	0.8852	0.9518	0.9741
BLESS ($\nu_1 = 10$)	0.5448	0.8336	0.9344	0.9288	0.9805	0.9930
BLESS ($\nu_1 = 100$)	0.2839	0.6634	0.8451	0.9528	0.9925	0.9978
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS ($\nu_1 = 1$)	0.9608	0.9990	1.0000	0.9175	0.9589	0.9757
BLESS ($\nu_1 = 10$)	0.8888	0.9942	0.9998	0.9683	0.9909	0.9957
BLESS ($\nu_1 = 100$)	0.7980	0.9759	0.9976	0.9847	0.9974	0.9993
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS ($\nu_1 = 1$)	1.0000	1.0000	1.0000	0.9702	0.9844	0.9913
BLESS ($\nu_1 = 10$)	1.0000	1.0000	1.0000	0.9851	0.9970	0.9984
BLESS ($\nu_1 = 100$)	1.0000	1.0000	1.0000	0.9920	0.9987	0.9998
		FPR			**FDR**	
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS ($\nu_1 = 1$)	0.0309	0.0158	0.0087	0.1148	0.0482	0.0259
BLESS ($\nu_1 = 10$)	0.0139	0.0056	0.0022	0.0712	0.0195	0.0070
BLESS ($\nu_1 = 100$)	0.0045	0.0017	0.0006	0.0472	0.0075	0.0022
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS ($\nu_1 = 1$)	0.0290	0.0144	0.0084	0.0825	0.0411	0.0243
BLESS ($\nu_1 = 10$)	0.0098	0.0031	0.0014	0.0317	0.0091	0.0043
BLESS ($\nu_1 = 100$)	0.0042	0.0009	0.0002	0.0153	0.0026	0.0007
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS ($\nu_1 = 1$)	0.0103	0.0053	0.0030	0.0298	0.0156	0.0087
BLESS ($\nu_1 = 10$)	0.0051	0.0010	0.0005	0.0149	0.0030	0.0016
BLESS ($\nu_1 = 100$)	0.0027	0.0004	0.0001	0.0080	0.0013	0.0002

Table A.15: Evaluation of inference results from the methods, BLESS, BSGLMM and Firth Regression via TPR, TDR, FPR and FDR for parameter estiamte $\hat{\beta}_2$.

A.6 Simulation Study: Sensitivity Analysis of Neighbourhood Structure

In order to ensure the reproducibility of our method and to strengthen the robustness of BLESS we perform an additional sensitivity analysis on the size of the neighbourhood structure of the MCAR prior. To illustrate the difference in neighbourhood structures we provide the below figure where the left depicts the previous neighbourhood structure where the sum of the neighbours of $\theta(s_5)$ on a 2D lattice is calculated by the blue coloured set of four neighbours $\partial\theta(s_5) = \{\theta(s_2), \theta(s_4), \theta(s_6), \theta(s_8)\}$. In the sensitivity analysis, we now extend the neighbourhood structure to a set of eight neighbours where the set is given by $\partial\theta(s_5) = \{\theta(s_1), \theta(s_2), \theta(s_3), \theta(s_4), \theta(s_6), \theta(s_7), \theta(s_8), \theta(s_9)\}$.

$\theta(s_1)$	$\boldsymbol{\theta(s_2)}$	$\theta(s_3)$
$\boldsymbol{\theta(s_4)}$	$\theta(s_5)$	$\boldsymbol{\theta(s_6)}$
$\theta(s_7)$	$\boldsymbol{\theta(s_8)}$	$\theta(s_9)$

$\boldsymbol{\theta(s_1)}$	$\boldsymbol{\theta(s_2)}$	$\boldsymbol{\theta(s_3)}$
$\boldsymbol{\theta(s_4)}$	$\theta(s_5)$	$\boldsymbol{\theta(s_6)}$
$\boldsymbol{\theta(s_7)}$	$\boldsymbol{\theta(s_8)}$	$\boldsymbol{\theta(s_9)}$

We show that BB-BLESS is robust to changing the tuning parameter which influences the spatial dependence structure of our method. Both the performance of parameter estimates and predictive performance as well as inference results yield very similar results for the extended neighborhood structure, consisting of 8 neighbors, versus the original choice of tuning parameter of 4 neighbors, see Table A.16 and A.17.

Parameter Estimate: $\hat{\boldsymbol{\beta}}_1$	**Bias**			**Variance**			**MSE**		
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (4 neighbours)	-0.0961	-0.0237	-0.0009	0.0014	0.0019	0.0020	0.0106	0.0024	0.0020
BLESS (8 neighbours)	-0.0945	-0.0205	-0.0001	0.0014	0.0019	0.0020	0.0103	0.0023	0.0020
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (4 neighbours)	-0.0031	0.0082	0.0019	0.0010	0.0010	0.0010	0.0010	0.0011	0.0010
BLESS (8 neighbours)	0.0104	-0.0047	0.0088	0.0010	0.0010	0.0010	0.0011	0.0010	0.0011
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (4 neighbours)	0.0032	0.0039	-0.0011	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
BLESS (8 neighbours)	0.0032	0.0039	-0.0011	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002

Predictive Performance: $\hat{\boldsymbol{y}}$	**Bias**			**Variance**			**MSE**		
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (4 neighbours)	-0.0078	-0.0052	-0.0032	0.0011	0.0017	0.0018	0.0022	0.0031	0.0034
BLESS (8 neighbours)	-0.0077	-0.0049	-0.0031	0.0011	0.0016	0.0018	0.0022	0.0031	0.0033
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (4 neighbours)	-0.0015	0.0007	-0.0013	0.0004	0.0005	0.0007	0.0008	0.0010	0.0012
BLESS (8 neighbours)	-0.0072	-0.0015	-0.0007	0.0004	0.0005	0.0007	0.0008	0.0009	0.0013
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (4 neighbours)	-0.0006	0.0000	0.0008	0.0001	0.0001	0.0001	0.0001	0.0002	0.0003
BLESS (8 neighbours)	-0.0006	0.0000	0.0008	0.0001	0.0001	0.0001	0.0001	0.0002	0.0003

Table A.16: Sensitivity analysis evaluating parameter estimates of BLESS with a spatial prior of four and eight neighbours via bias, variance and MSE of the spatially varying coefficients $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$, and the predictive performance $\hat{\boldsymbol{y}}$. Variations of the neighbourhood in the spatial prior of BLESS leave the results relatively unchanged. Accounting for a larger spatial neighbourhood potentially reduces bias in the parameter estimates and prediction results for lower sample sizes, such as $N = 500$.

	TPR			**TDR**		
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (4 neighbours)	0.6635	0.9080	0.9770	0.9520	0.9808	0.9893
BLESS (8 neighbours)	0.6679	0.9140	0.9788	0.9512	0.9795	0.9881
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (4 neighbours)	0.9458	0.9983	1.0000	0.9713	0.9882	0.9936
BLESS (8 neighbours)	0.9499	0.9989	1.0000	0.9665	0.9768	0.9915
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (4 neighbours)	1.0000	1.0000	1.0000	0.9873	0.9973	0.9968
BLESS (8 neighbours)	1.0000	1.0000	1.0000	0.9867	0.9964	0.9960

	FPR			**FDR**		
N=500	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (4 neighbours)	0.0334	0.0178	0.0106	0.0480	0.0192	0.0107
BLESS (8 neighbours)	0.0342	0.0191	0.0119	0.0488	0.0205	0.0119
N=1,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (4 neighbours)	0.0282	0.0120	0.0065	0.0287	0.0118	0.0064
BLESS (8 neighbours)	0.0331	0.0238	0.0086	0.0335	0.0232	0.0085
N=5,000	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
BLESS (4 neighbours)	0.0129	0.0027	0.0032	0.0127	0.0027	0.0032
BLESS (8 neighbours)	0.0135	0.0037	0.0040	0.0133	0.0036	0.0040

Table A.17: Sensitivity analysis evaluating inference results of BLESS with a spatial prior of four and eight neighbours via TPR, TDR, FPR and FDR for parameter estimate $\hat{\beta}_1$. Extending the neighbourhood in the spatial prior of BLESS to eight voxels instead of the original four voxels leaves the inference results fairly unchanged. Sensitivity is slightly increased and specificity is slightly descreased by a larger spatial neighbourhood.

B

Appendix: Scalable Uncertainty Quantification and Cluster Size-Based Imaging Statistics for Large-scale Lesion Mapping Applications

Contents

B.1 Derivation of Approximate Posterior Sampling Method for BB-BLESS

B.1.1 Joint Distribution

Re-weighting of the likelihood with Dirichlet weights $\boldsymbol{w}^{(b)} \sim N \times Dir(1,\dots,1)$ and jitter the spike-and-slab prior by applying the mean shift $\mu_p(s_j) \sim \mathcal{N}(0,\nu_0)$, for all $p = 1, \dots, P$ and $j = 1, \dots, M$.

$$
\begin{aligned}
Q = &\prod_{i=1}^{N}\prod_{j=1}^{M} \Pr(y_i(s_j)|z_i(s_j))^{w_i^{(b)}} \\
&\times \prod_{i=1}^{N}\prod_{j=1}^{M} \mathcal{N}(z_i(s_j); \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(s_j) + \beta_0(s_j), 1)^{w_i^{(b)}} \\
&\times \prod_{j=1}^{M} \mathcal{N}(\beta_0(s_j); \mu_{\beta_0}, \sigma_{\beta_0}^2) \\
&\times \prod_{j=1}^{M} \mathcal{N}(\boldsymbol{\beta}(s_j); \boldsymbol{\mu}(s_j), \mathrm{diag}\{\nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j)\}_{p=1}^{P}) \\
&\times \prod_{p=1}^{P}\prod_{j=1}^{M} Bernoulli(\gamma_p(s_j); \sigma(\theta_p(s_j))) \\
&\times \mathcal{N}(\boldsymbol{\theta}; \mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{D}-\boldsymbol{W})^{-1}) \\
&\times Wishart(\boldsymbol{\Sigma}^{-1}; P, \boldsymbol{I})
\end{aligned}
$$

$$
\begin{aligned}
Q^* \geq \mathbb{E}_{q(z,\beta,\beta_0,\gamma,\theta,\Sigma^{-1})} &\Bigg[-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M} w_i^{(b)}(z_i(s_j) - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}(s_j) - \beta_0(s_j))^2 - \frac{1}{2\sigma_{\beta_0}^2}\sum_{j=1}^{M}(\beta_0(s_j) - \mu_{\beta_0})^2 \\
&- \frac{1}{2}\sum_{j=1}^{M}\log(|\mathrm{diag}\{\nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j)\}_{p=1}^{P}|) \\
&- \frac{1}{2}\sum_{j=1}^{M}[\boldsymbol{\beta}(s_j) - \boldsymbol{\mu}(s_j)]^{\mathrm{T}}\,\mathrm{diag}\{\nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j)\}_{p=1}^{P}\,[\boldsymbol{\beta}(s_j) - \boldsymbol{\mu}(s_j)] \\
&+ \sum_{p=1}^{P}\sum_{j=1}^{J}\left[\log(\sigma(\xi_p(s_j))) + \theta_p(s_j)\gamma_p(s_j) - \frac{(\theta_p(s_j) + \xi_p(s_j))}{2} - \lambda(\xi_p(s_j))(\theta_p(s_j)^2 - \xi_p(s_j)^2)\right] \\
&+ \frac{1}{2}\sum_{j=1}^{M}\log\left\{|\boldsymbol{\Sigma}^{-1}|\right\} - \frac{1}{2}\sum_{s_i\sim s_j}[\boldsymbol{\theta}(s_i) - \boldsymbol{\theta}(s_j)]^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}[\boldsymbol{\theta}(s_i) - \boldsymbol{\theta}(s_j)] \\
&+ \frac{1}{2}\left\{(\nu - P - 1)\log\{|\boldsymbol{\Sigma}^{-1}|\} - \mathrm{tr}(\boldsymbol{\Sigma}^{-1})\right\}\Bigg].
\end{aligned}
$$

B.1.2 Variational Approximations

B.1.2.1 Update $z_i(s_j)$

Assume $y_i(s_j) = 1$ and $\boldsymbol{\eta}(s_j) = \boldsymbol{x}_i^T \mathbb{E}_{q(\beta)}[\boldsymbol{\beta}(s_j)] + \mathbb{E}_{q(\beta_0)}[\beta_0(s_j)]$:

$$
\begin{aligned}
\ln(q^*(z_i(s_j))) &\propto w_i^{(b)} \ln\{p(y_i(s_j)|z_i(s_j))\} + w_i^{(b)} \mathbb{E}_{q(\beta,\beta_0)}\left[\ln\{p(z_i(s_j)|\boldsymbol{x}_i, \boldsymbol{\beta}(s_j), \beta_0(s_j))\}\right] \\
&\propto w_i^{(b)} y_i(s_j) \ln\left\{\mathbb{1}(z_i(s_j) > 0)\right\} + w_i^{(b)} (1 - y_i(s_j)) \ln\left\{\mathbb{1}(z_i(s_j) \leq 0)\right\} \\
&- \frac{1}{2} w_i^{(b)} \mathbb{E}_{q(\beta,\beta_0)}[(z_i(s_j) - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}(s_j) - \beta_0(s_j))^2] \\
&\propto w_i^{(b)} \ln\left\{\mathbb{1}(z_i(s_j) > 0)\right\} - \frac{1}{2} w_i^{(b)} z_i(s_j)^2 + w_i^{(b)} z_i(s_j)[\mathbb{E}_{q(\beta)}[\boldsymbol{\beta}(s_j)]^{\mathrm{T}} \boldsymbol{x}_i \\
&+ \mathbb{E}_{q(\beta_0)}[\beta_0(s_j)]] \\
&\propto w_i^{(b)} \ln\left\{\mathbb{1}(z_i(s_j) > 0)\right\} - \frac{1}{2} w_i^{(b)} z_i(s_j)^2 + w_i^{(b)} z_i(s_j) \eta_i(s_j)
\end{aligned}
$$

$$
q^*(z_i(s_j)) = \begin{cases} \mathcal{TN}_+(z_i(s_j); \eta_i(s_j), 1), & \text{if } y_i(s_j) = 1, \\ \mathcal{TN}_-(z_i(s_j); \eta_i(s_j), 1), & \text{if } y_i(s_j) = 0. \end{cases}
$$

B.1.2.2 Update $\beta(s_j)$

$$
\begin{aligned}
\ln(q^*(\boldsymbol{\beta}(s_j))) &\propto \mathbb{E}_{q(\gamma)}[\ln\{p(\boldsymbol{\beta}(s_j)|\boldsymbol{\gamma}(s_j))\}] + \sum_i^N w_i^{(b)} \mathbb{E}_{q(z,\beta_0)}[\ln\{p(z_i(s_j)|\boldsymbol{x}_i, \boldsymbol{\beta}(s_j), \beta_0(s_j))\}] \\
&\propto -\frac{1}{2} \left[\boldsymbol{\beta}(s_j) - \boldsymbol{\mu}(s_j)\right]^{\mathrm{T}} \mathbb{E}_{q(\gamma)}[\mathrm{diag}\{\nu_0(1 - \gamma_p(s_j)) + \nu_1 \gamma_p(s_j)\}_{p=1}^P]^{-1} \left[\boldsymbol{\beta}(s_j) - \boldsymbol{\mu}(s_j)\right] \\
&- \frac{1}{2} \mathbb{E}_{q(z,\beta_0)} \left[[\boldsymbol{z}(s_j) - X\boldsymbol{\beta}(s_j) - \beta_0(s_j)]^{\mathrm{T}} \mathrm{diag}\left\{w_i^{(b)}\right\}_i^N [\boldsymbol{z}(s_j) - X\boldsymbol{\beta}(s_j) - \beta_0(s_j)] \right] \\
&\propto \mathbb{E}_{q(z,\beta_0)} \left[\boldsymbol{\beta}(s_j)^T X^T \mathrm{diag}\left\{w_i^{(b)}\right\}_i^N [\boldsymbol{z}(s_j) - \beta_0(s_j)] \right] \\
&- \frac{1}{2} \boldsymbol{\beta}(s_j)^T X^T \mathrm{diag}\left\{w_i^{(b)}\right\}_i^N X\boldsymbol{\beta}(s_j) \\
&+ \boldsymbol{\beta}(s_j)^{\mathrm{T}} \mathbb{E}_{q(\gamma)}[\mathrm{diag}\{\nu_0(1 - \gamma_p(s_j)) + \nu_1 \gamma_p(s_j)\}_{p=1}^P]^{-1} \boldsymbol{\mu}(s_j) \\
&- \frac{1}{2} \mathrm{tr}(\mathbb{E}_{q(\gamma)}[\mathrm{diag}\{\nu_0(1 - \gamma_p(s_j)) + \nu_1 \gamma_p(s_j)\}_{p=1}^P]^{-1} \boldsymbol{\beta}(s_j) \boldsymbol{\beta}(s_j)^T) \\
&\propto \boldsymbol{\beta}(s_j)^T \Big[X^T \mathrm{diag}\left\{w_i^{(b)}\right\}_i^N [\mathbb{E}_{q(z}[\boldsymbol{z}(s_j)] - \mathbb{E}_{q(\beta_0)}[\beta_0(s_j)] \\
&+ \mathbb{E}_{q(\gamma)}[\mathrm{diag}\{\nu_0(1 - \gamma_p(s_j)) + \nu_1 \gamma_p(s_j)\}_{p=1}^P]^{-1} \boldsymbol{\mu}(s_j) \Big] \\
&- \frac{1}{2} \mathrm{tr}([X^T \mathrm{diag}\left\{w_i^{(b)}\right\}_i^N X \\
&+ \mathbb{E}_{q(\gamma)}[\mathrm{diag}\{\nu_0(1 - \gamma_p(s_j)) + \nu_1 \gamma_p(s_j)\}_{p=1}^P]^{-1}] \boldsymbol{\beta}(s_j) \boldsymbol{\beta}(s_j)^T)
\end{aligned}
$$

$$\begin{aligned}
q^*(\boldsymbol{\beta}(s_j)) &= \mathcal{N}(\boldsymbol{\beta}(s_j); \mu_{\beta(s_j)}, \Sigma_{\beta(s_j)}) \\
\mu_{\beta(s_j)} &= \Sigma_{\beta(s_j)} \Big[X^{\mathrm{T}} \mathrm{diag} \left\{ w_i^{(b)} \right\}_i^N \left[\mathbb{E}_{q(z)}[\boldsymbol{z}(s_j)] - \mathbb{E}_{q(\beta_0)}[\beta_0(s_j)] \right] \\
&\quad + \mathbb{E}_{q(\gamma)}[\mathrm{diag}\{\nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j)\}_{p=1}^P]^{-1} \boldsymbol{\mu}(s_j) \Big] \\
\Sigma_{\beta(s_j)} &= \Big[X^{\mathrm{T}} \mathrm{diag} \left\{ w_i^{(b)} \right\}_i^N X + \mathbb{E}_{q(\gamma)}[\mathrm{diag}\{\nu_0(1-\gamma_p(s_j)) + \nu_1\gamma_p(s_j)\}_{p=1}^P]^{-1} \Big]^{-1}
\end{aligned}$$

B.1.2.3 Update $\beta_0(s_j)$

$$\begin{aligned}
\ln(q^*(\beta_0(s_j))) &\propto \mathbb{E}_{q(z,\beta)}[\sum_i^N w_i^{(b)} \ln p(z_i(s_j) | \boldsymbol{x}_i, \beta_0(s_j), \boldsymbol{\beta}(s_j))] + \ln p(\beta_0(s_j)) \\
&\propto \sum_i^N \mathbb{E}_{q(z,\beta)}[w_i^{(b)} \ln \mathcal{N}(z_i(s_j); \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}(s_j) + \beta_0(s_j), 1)] + \ln \mathcal{N}(\beta_0(s_j); 0, \sigma^2_{\beta_0}) \\
&\propto \sum_{i=1}^N \beta_0(s_j) w_i^{(b)} [\mathbb{E}_{q(z)}[z_i(s_j)] - \boldsymbol{x}_i^T \mathbb{E}_{q(\beta)}[\boldsymbol{\beta}(s_j)]] - \frac{1}{2} \sum_i^N w_i^{(b)} \beta_0(s_j)^2 - \frac{1}{2\sigma^2_{\beta_0}} \beta_0(s_j)^2
\end{aligned}$$

$$\begin{aligned}
q^*(\beta_0(s_j)) &= \mathcal{N}(\beta_0(s_j); \mu_{\beta_0(s_j)}, \sigma^2_{\beta_0(s_j)}) \\
\mu_{\beta_0(s_j)} &= [\sum_i^N w_i^{(b)} + \frac{1}{\sigma^2_{\beta_0}}]^{-1} [\sum_{i=1}^N w_i^{(b)} \left[\mathbb{E}_{q(z)}[z_i(s_j)] - \boldsymbol{x}_i^T \mathbb{E}_{q(\beta)}[\boldsymbol{\beta}(s_j)] \right]] \\
\sigma^2_{\beta_0(s_j)} &= [\sum_i^n w_i^{(b)} + \frac{1}{\sigma^2_{\beta_0}}]^{-1}
\end{aligned}$$

B.1.3 Other variational updates

The variational updates for the parameters $\boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\Sigma^{-1}}$ for BB-BLESS are identical to the derived variational distributions of the standard BLESS-VI approach described in the previous section as the updates for those parameters are unaffected by the re-weighting of the likelihood and the perturbation of the prior mean in the spike-and-slab prior.

B.2 Computational Complexity of BLESS

Quantity	Dimensions	Space Complexity
$q(\boldsymbol{Z})$	$N \times M$	$\mathcal{O}([N \times (P+1) \times M]^2)$
$q(\boldsymbol{\beta})$	$P \times M$	$\mathcal{O}(N \times P^2 \times M)$
$q(\boldsymbol{\beta_0})$	$M \times 1$	$\mathcal{O}(N \times P \times M + M^2)$
$q(\boldsymbol{\gamma})$	$P \times M$	$\mathcal{O}([P \times M]^2)$
$q(\boldsymbol{\theta})$	$P \times M$	$\mathcal{O}(P^2 \times M)$
$q(\boldsymbol{\xi})$	$P \times M$	$\mathcal{O}(P^2 \times M)$
$q(\boldsymbol{\Sigma^{-1}})$	$P \times P$	$\mathcal{O}(P^2)$
ELBO	1×1	$\mathcal{O}([N \times M]^2)$

Table B.1: Computational complexity of BLESS-VI algorithm for each variational update and the calculation of the ELBO.

B.3 Further Simulation Study Results

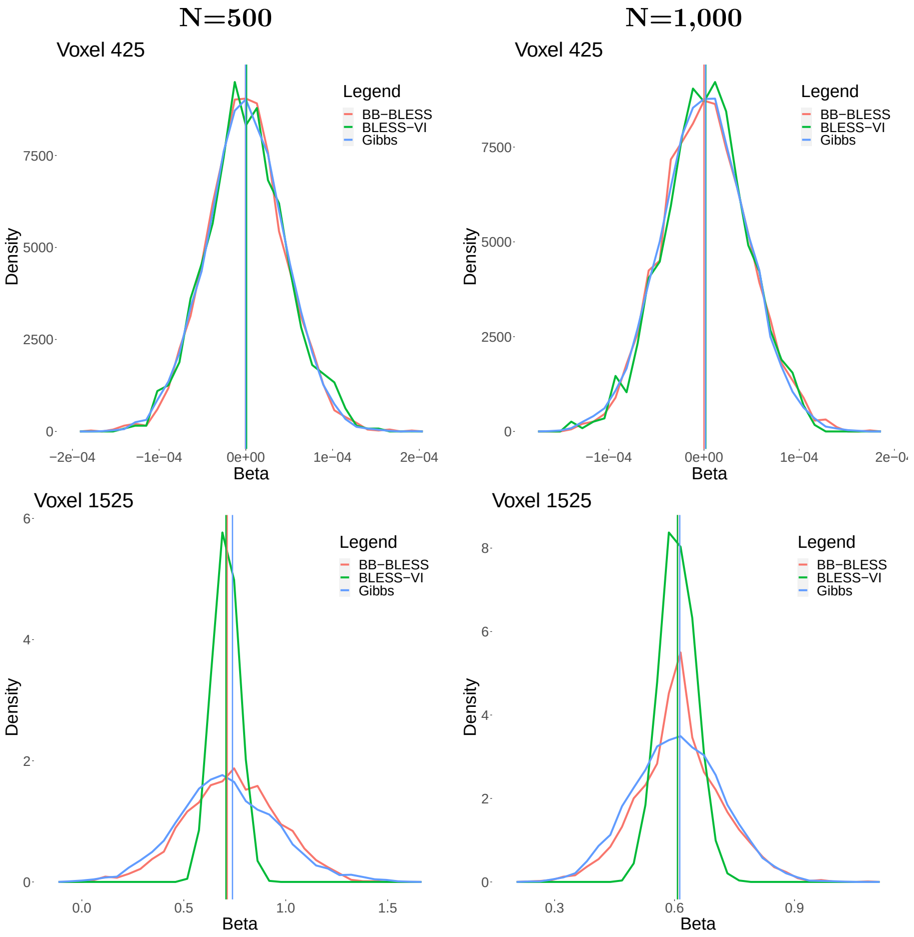


Figure B.1: Comparison of marginal posterior distributions between BB-BLESS, BLESS-Gibbs and BLESS-VI where the posterior mean is indicated via a vertical line.

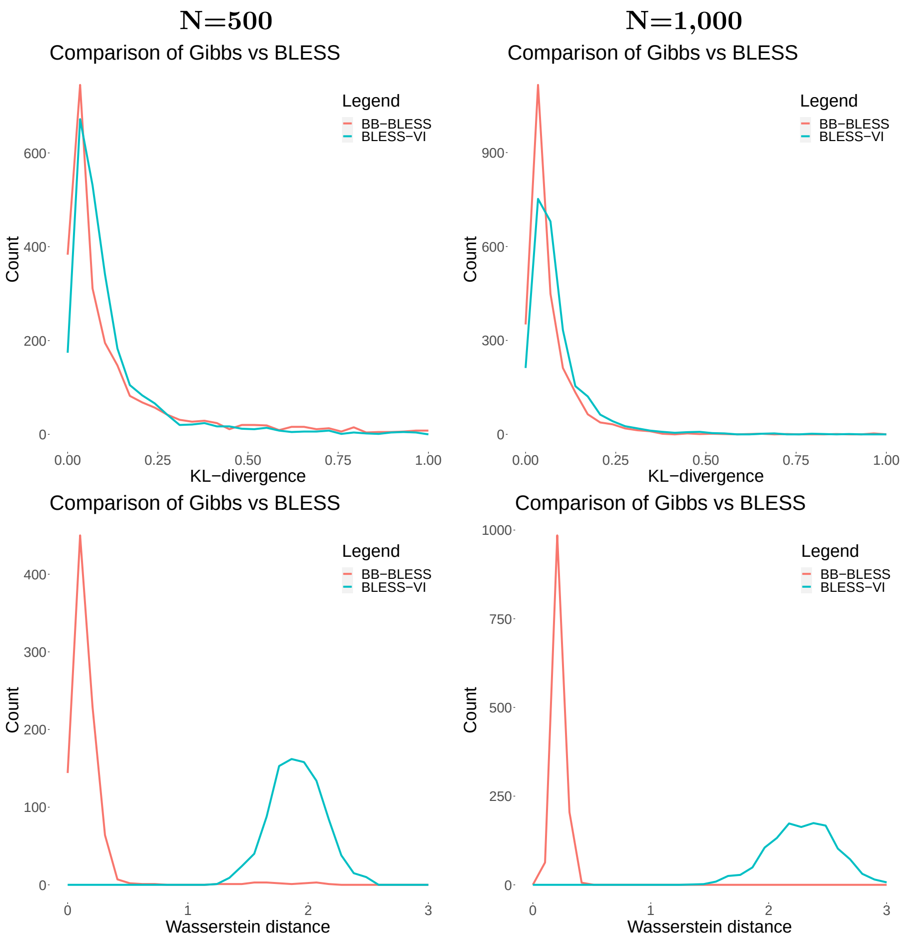
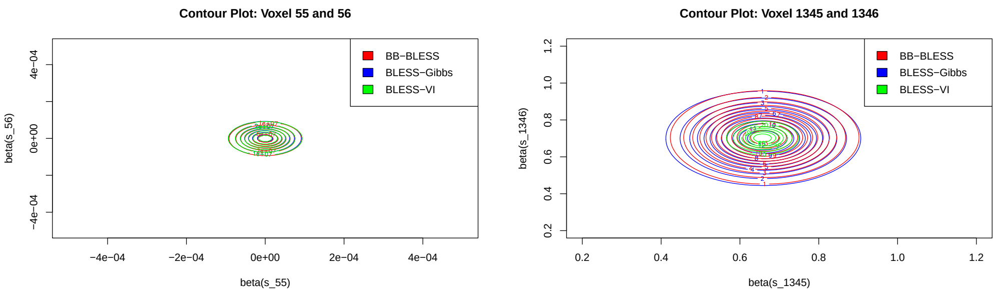


Figure B.2: Evaluation of marginal posterior distribution between Gibbs and BB-BLESS and BLESS-VI via KL-divergence and Wasserstein distance.



Inactive voxels **Active voxels**

Figure B.3: Contour plots between two sets of randomly chosen inactive and active neighbouring voxels estimated via BB-Gibbs (in blue), BB-BLESS (in red), and BLESS-VI (in green) showcasing that BLESS-VI does underestimate the tails of the distribution whereas BB-BLESS is able to reflect the posterior distribution more accurately compared to BLESS-Gibbs for the contour plot comparing two neighbouring active voxels.

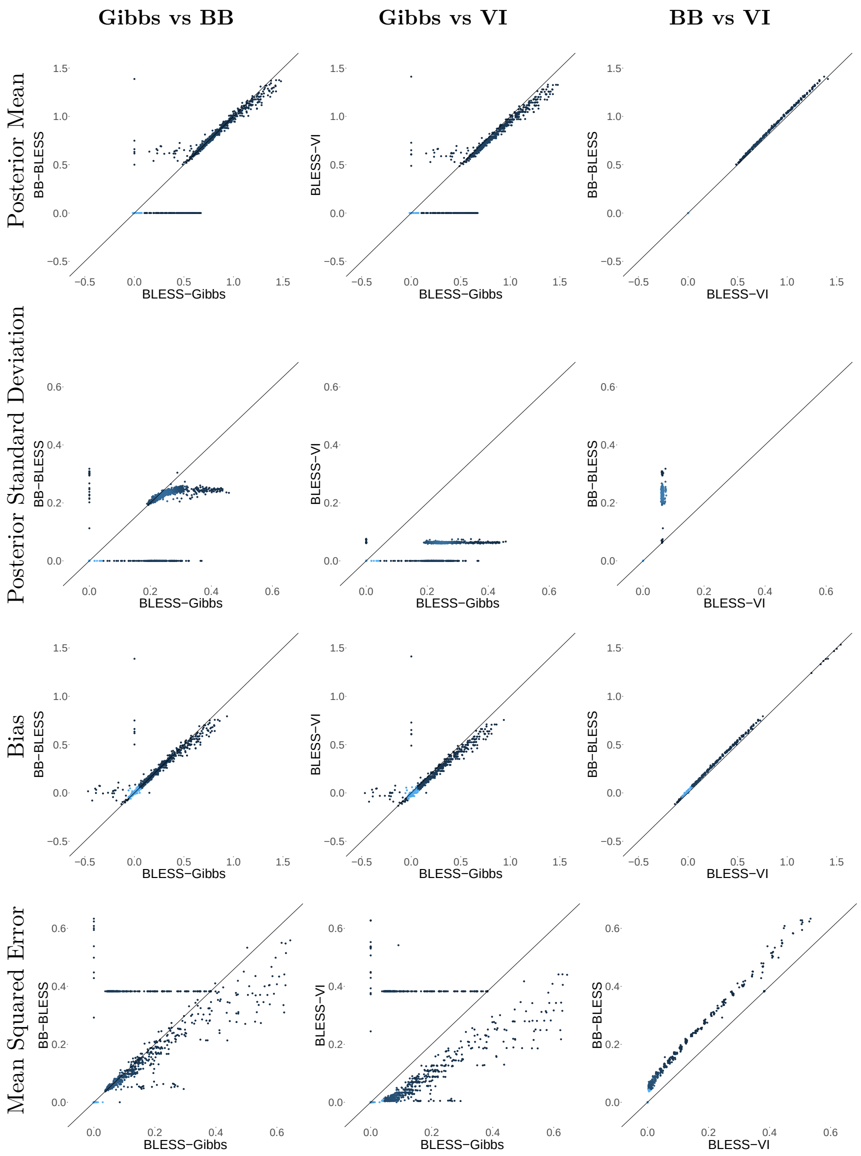


Figure B.4: Comparison of posterior quantities, such as posterior mean and standard deviation, bias and mean squared error of the parameter estimates, between BB-BLESS, BLESS-Gibbs and BLESS-VI for N=500.

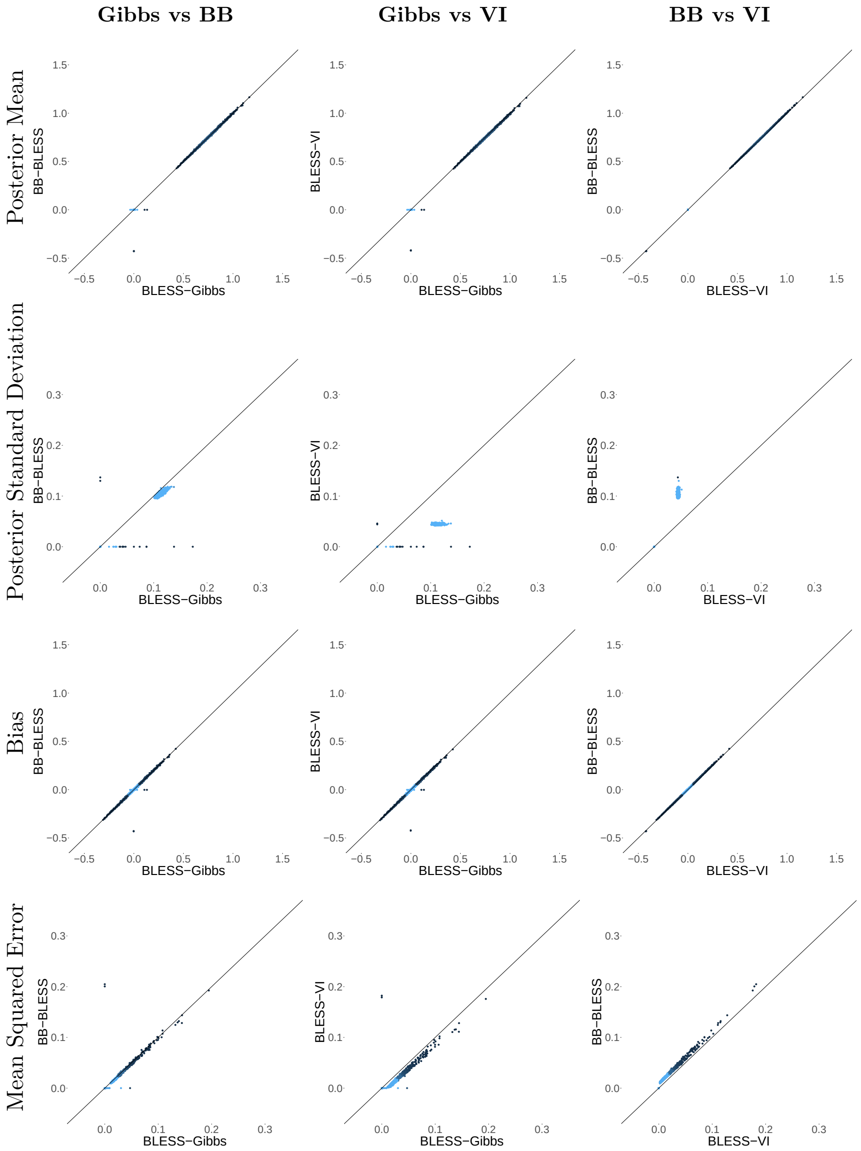


Figure B.5: Comparison of posterior quantities, such as posterior mean and standard deviation, bias and mean squared error of the parameter estimates, between BB-BLESS, BLESS-Gibbs and BLESS-VI for N=1, 000.

	$\boldsymbol{N} = \mathbf{500}$ and $\boldsymbol{\lambda} = 1$			**$\boldsymbol{N} = \mathbf{1,000}$ and $\boldsymbol{\lambda} = 3$**		
	β_0	β_1	β_2	β_0	β_1	β_2
Truth	-2.0868	0.5882	0.5586	-1.6089	0.7051	0.6685
BLESS-MCMC	-2.1461	0.5994	0.5802	-1.6102	0.7003	0.6517
BB-BLESS	-2.1037	0.4911	0.4848	-1.6148	0.7028	0.6622
BLESS-VI	-2.0777	0.4741	0.4685	-1.6089	0.6981	0.6576
BSGLMM	-2.1507	0.6156	0.5976	-1.6197	0.7030	0.6647
Firth	-2.1090	0.5952	0.5835	-1.6106	0.7006	0.6644

Table B.2: Evaluation of parameter estimates $\hat{\boldsymbol{\beta}}_0$, $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ for BLESS (Gibbs, BB, VI), BSGLMM and Firth regression compared to the truth for $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ for scenario 1 ($N = 500$ and $\lambda = 1$) and scenario 2 ($N = 1,000$ and $\lambda = 3$).

	$\boldsymbol{N} = \mathbf{500}$ and $\boldsymbol{\lambda} = \mathbf{1}$			**$\boldsymbol{N} = \mathbf{1,000}$ and $\boldsymbol{\lambda} = \mathbf{3}$**		
$\hat{\boldsymbol{\beta}}_\mathbf{0}$	**Bias**	**Variance**	**MSE**	**Bias**	**Variance**	**MSE**
BLESS-MCMC	-0.0701	0.0576	0.0625	-0.0094	0.0076	0.0077
BB-BLESS	-0.0210	0.0267	0.0271	-0.0122	0.0075	0.0077
BLESS-VI	0.0055	0.0020	0.0020	-0.0061	0.0010	0.0010
BSGLMM	-0.0658	0.0213	0.0256	-0.0115	0.0054	0.0055
Firth	-0.0182	0.0539	0.0542	-0.0027	0.0118	0.0118
$\hat{\boldsymbol{\beta}}_\mathbf{1}$	**Bias**	**Variance**	**MSE**	**Bias**	**Variance**	**MSE**
BLESS-MCMC	0.0288	0.0406	0.0414	0.0060	0.0065	0.0065
BB-BLESS	-0.0857	0.0171	0.0245	0.0078	0.0063	0.0064
BLESS-VI	-0.1037	0.0013	0.0121	0.0027	0.0010	0.0010
BSGLMM	0.0518	0.0120	0.0147	0.0126	0.0039	0.0040
Firth	-0.0182	0.0539	0.0542	-0.0027	0.0118	0.0118
$\hat{\boldsymbol{\beta}}_\mathbf{2}$	**Bias**	**Variance**	**MSE**	**Bias**	**Variance**	**MSE**
BLESS-MCMC	0.0441	0.0234	0.0253	0.0012	0.0032	0.0032
BB-BLESS	-0.0746	0.0092	0.0148	0.0038	0.0032	0.0032
BLESS-VI	-0.0929	0.0007	0.0093	-0.0013	0.0005	0.0005
BSGLMM	0.0364	0.0114	0.0127	-0.0038	0.0038	0.0038
Firth	-0.0182	0.0539	0.0542	-0.0027	0.0118	0.0118
$\hat{\boldsymbol{y}}$	**Bias**	**Variance**	**MSE**	**Bias**	**Variance**	**MSE**
BLESS-MCMC	-0.0117	0.0007	0.0014	-0.0037	0.0007	0.0012
BB-BLESS	-0.0174	0.0011	0.0021	-0.0051	0.0007	0.0013
BLESS-VI	-0.0010	0.0011	0.0021	-0.0014	0.0007	0.0012
BSGLMM	-0.0058	0.0002	0.0002	-0.0025	0.0003	0.0003
Firth	0.0125	0.0009	0.0018	0.0051	0.0011	0.0022

Table B.3: Evaluation of parameter estimates $\hat{\boldsymbol{\beta}}_0$, $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ and the predictive perfomance of the estimated lesion probabilities $\hat{\boldsymbol{y}}$ for BLESS (MCMC, BB, VI), BSGLMM and Firth regression via bias, variance and MSE.

	$N = 500$ and $\lambda = 1$				**$N = 1,000$ and $\lambda = 3$**			
$\boldsymbol{t_{\hat{\beta}_1}}$	**TPR**	**TDR**	**FPR**	**FDR**	**TPR**	**TDR**	**FPR**	**FDR**
BLESS-MCMC	0.7319	0.9998	0.0001	0.0002	0.9999	0.9999	0.0001	0.0001
BB-BLESS	0.6263	0.9606	0.0257	0.0394	0.9999	0.9970	0.0031	0.0030
BLESS-VI	0.6263	0.9606	0.0257	0.0394	1.0000	0.9970	0.0031	0.0031
BSGLMM	0.9991	0.9004	0.1128	0.0996	1.0000	0.9027	0.1088	0.0973
Firth	0.6566	0.8953	0.0768	0.1047	1.0000	0.9637	0.0379	0.0363
$\boldsymbol{t_{\hat{\beta}_2}}$	**TPR**	**TDR**	**FPR**	**FDR**	**TPR**	**TDR**	**FPR**	**FDR**
BLESS-MCMC	0.7376	0.9997	0.0001	0.0003	1.0000	0.9998	0.0001	0.0002
BB-BLESS	0.6478	0.9135	0.0204	0.0865	0.9998	0.9900	0.0034	0.0100
BLESS-VI	0.6479	0.9136	0.0204	0.0864	0.9998	0.9900	0.0034	0.0100
BSGLMM	0.9955	0.7664	0.1033	0.2336	1.0000	0.8093	0.0803	0.1907
Firth	0.4853	0.6625	0.0813	0.3375	1.0000	0.9181	0.0303	0.0819

Table B.4: Evaluation of inference results for BLESS (MCMC, BB, VI), BSGLMM and Firth regression for two simulation study scenarios ($N = 500$, $\lambda = 1$ and $N = 1,000$, $\lambda = 3$) for the two sizes of effect wihtin the image.

B.4 Simulation Study with Realistic Lesion Masks

The results of this section are based on a simulation study framework developed by Kindalova et al. (2021). In their work, they simulate artificial lesion maps with realistic lesion patterns by generating a voxel-wise distribution of lesions across age which uses the age effect on lesion probability estimated from a subset of the UK Biobank population. The simulation study framework yields realistic lesion masks by matching brain lesion summaries, such as total lesion volume, average lesion size, and lesion count, across the reference dataset from the UK Biobank and the simulated datasets. We simulate a dataset of 1,000 subjects within the simulation framework provided by Kindalova et al. (2021) on their Github page (`https://github.com/petyakindalova/LesionMaskSimulation`) for the evaluation of our methods (BLESS-Gibbs, BB-BLESS, and BLESS-VI) compared to other approaches (BSGLMM and Firth Regression). The ground truth spatially varying regression coefficients were obtained from the reference dataset (13,000 subjects) estimated regression coefficients for the age effect where the effect map was thresholded to have both null and real effects. We do this by running the command `'cluster'` in the neuroimaging software `FSL` with a cluster-defining threshold of 0.05.

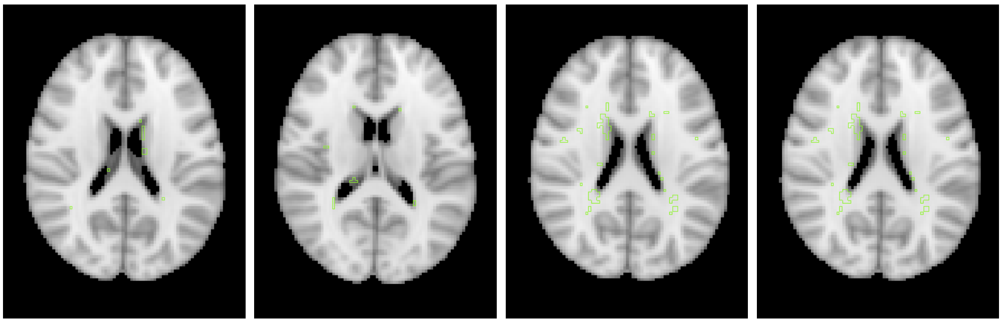


Figure B.6: Lesion masks generated via simulation framework based on UK Biobank summary statistics. Two most left-hand lesion masks stem from younger subjects with fewer lesions opposed to the two right-hand lesion masks which were generated for older subjects. Lesion masks are overlayed on the MNI template and lesion borders are depicted by green outlines.

	Parameter Estimate: $\hat{\boldsymbol{\beta}}_1$			**Predictive Performance: $\hat{\boldsymbol{y}}$**		
	Bias	Variance	MSE	Bias	Variance	MSE
BLESS-Gibbs	0.00073	0.00024	0.00035	-0.00020	0.00544	2.19×10^{-5}
BB-BLESS	0.00095	0.00011	0.00046	-0.00023	0.00544	2.25×10^{-5}
BLESS-VI	0.00083	0.00005	0.00045	-0.00022	0.00543	2.21×10^{-5}
BSGLMM	-0.00227	0.00027	0.00069	-0.00050	0.00544	2.30×10^{-5}
Firth	-0.00209	0.00041	0.00050	0.00082	0.00542	6.90×10^{-7}

Table B.5: Evaluation of parameter estimates for the effect of age from various methods where truth is determined by a simulation framework which utilises UK Biobank summary statistics by Kindalova et al. (2021).

Inference Results	TPR	TDR	FPR	FDR
BLESS-Gibbs	0.9278	1.0000	0.0000	0.0000
BB-BLESS	0.8910	1.0000	0.0000	0.0000
BLESS-VI	0.9049	1.0000	0.0000	0.0000
BSGLMM	0.8990	0.9932	0.0031	0.0068
Firth	0.7393	0.9959	0.0015	0.0041

Table B.6: Evaluation of inference results from various methods where truth is determined by a simulation framework which utilises UK Biobank summary statistics by Kindalova et al. (2021).

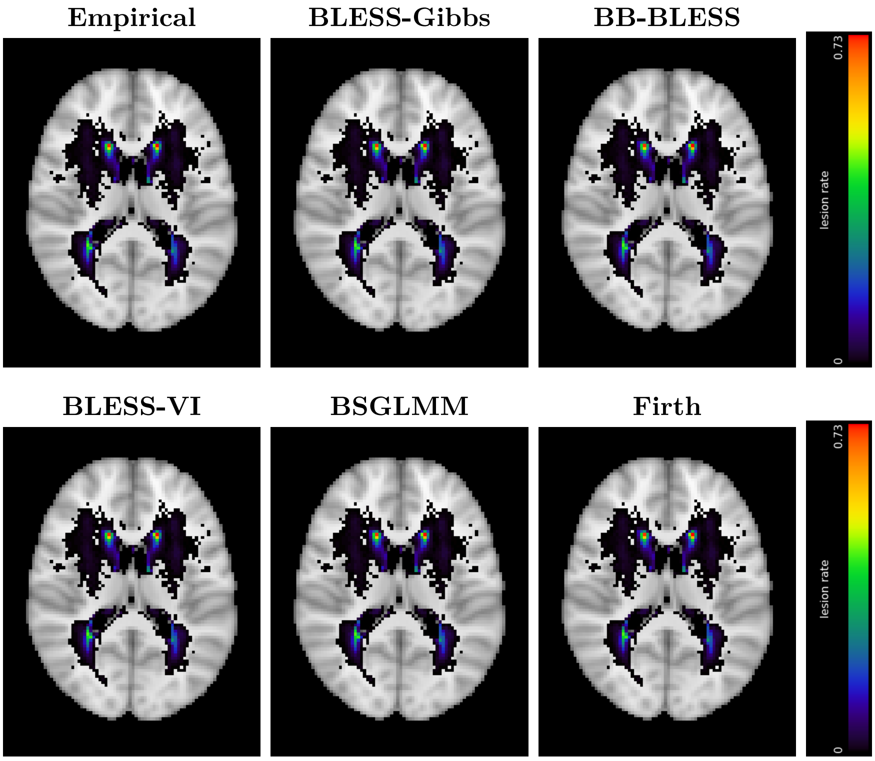


Figure B.7: True empirical lesion rate map and predicted lesion rate maps (BLESS-Gibbs, BB-BLESS, BLESS-VI, BSGLMM, and Firth Regression).

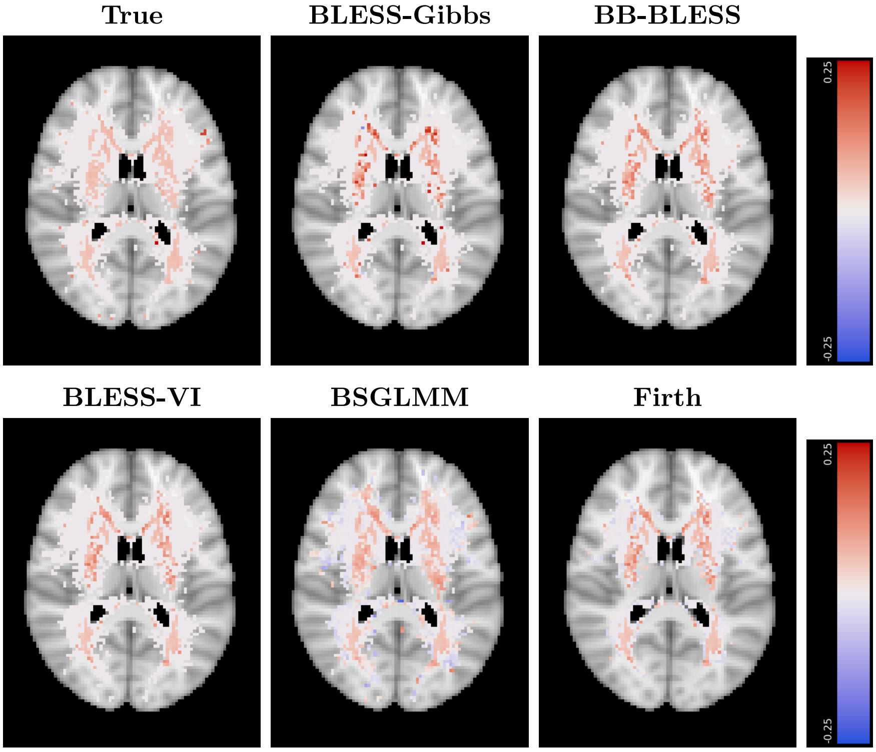


Figure B.8: True parameter map and estimated parameter maps for the covariate age via realistic lesion mask simulation framework for the methods, BLESS-Gibbs, BB-BLESS, BLESS-VI, BSGLMM, and Firth Regression.

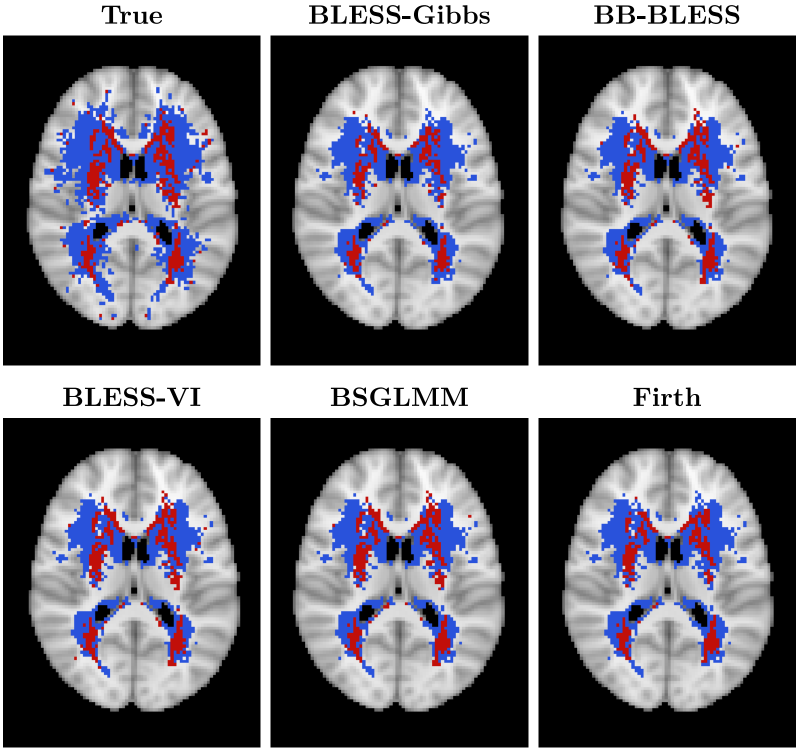


Figure B.9: True binary significance map and estimated binary significance maps for the covariate age via realistic lesion mask simulation framework for the methods, BLESS-Gibbs, BB-BLESS, BLESS-VI, BSGLMM, and Firth Regression.

B.5 Posterior Predictive Checks

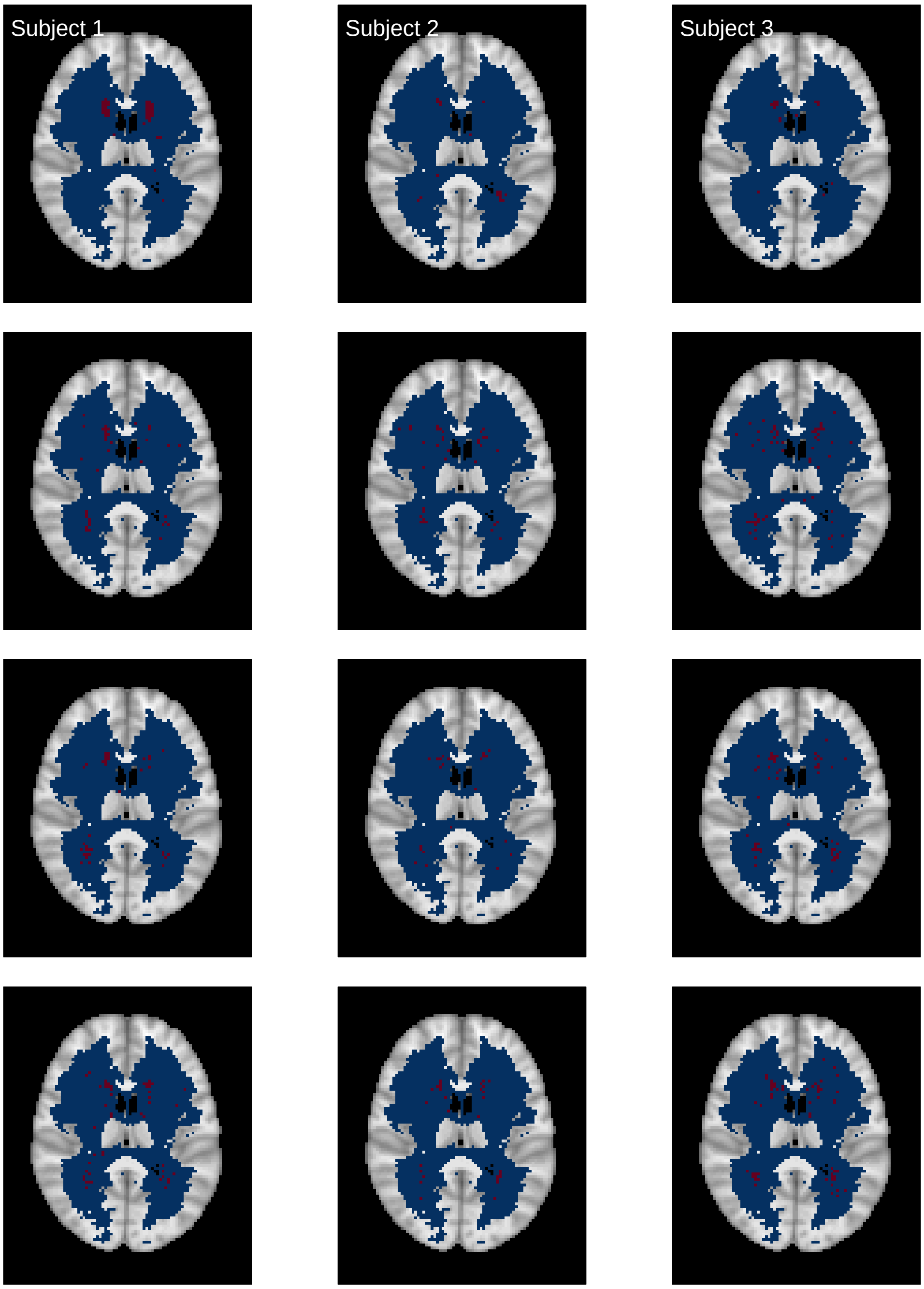
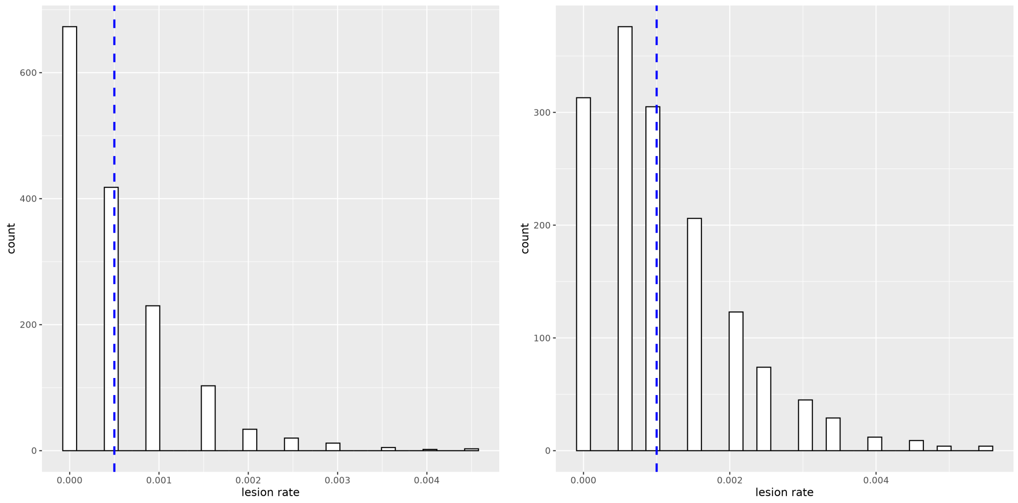


Figure B.10: First row contains true lesion images and second to fourth row contains lesion masks generated via posterior predictive distribution $p(\hat{y}|y) = \int p(\hat{y}|\beta, \beta_0)p(\beta, \beta_0|y)d\beta d\beta_0$, where $\hat{y}(s_j) \sim Bernoulli(\Phi^{-1}\{\beta_0(s_j)^{(b)} + \boldsymbol{x}_i^T\boldsymbol{\beta}(s_j)^{(b)}\})$.

(a) Minimum **(b)** Median



(c) Mean **(d)** Maximum

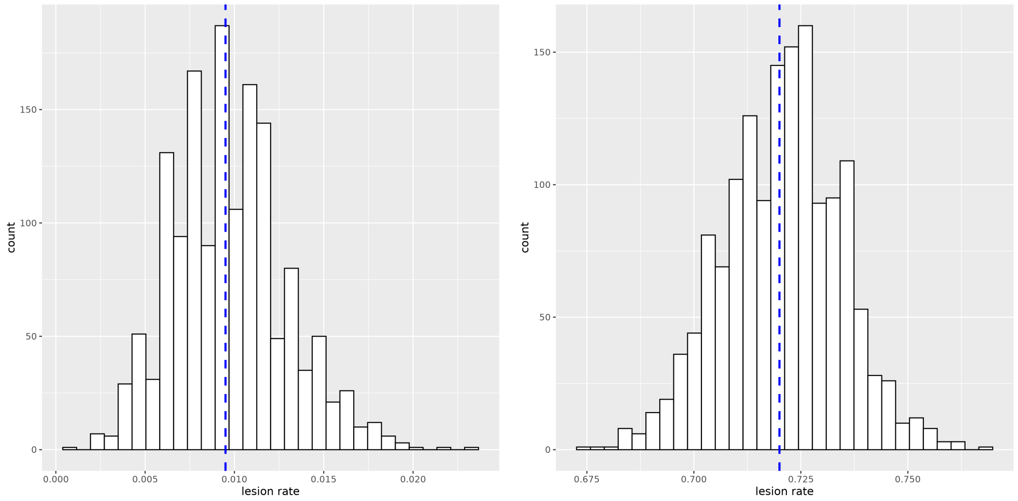


Figure B.11: Lesion rate comparisons of a voxel location with the (a) minimum, (b) median, (c) mean and (d) maximum empirical lesion rate across all voxels. Blue line indicates true lesion rate at particular voxel based on dataset consisting of 2,000 subjects. Histogram is based on lesion rates based on datasets generated via posterior predictive distribution $\hat{y}(s_j) \sim Bernoulli(\Phi^{-1}\{\beta_0(s_j)^{(b)} + \boldsymbol{x}_i^T\boldsymbol{\beta}(s_j)^{(b)}\})$. The p-values $p = Pr(T(y_{rep}, \theta) \geq T(y, \theta)|y)$ are (a) 0.55133, (b) 0.54067, (c) 0.54267, and (d) 0.55, where values close to 0 or 1 indicate a poor fit. All voxel locations pass the posterior predictive check.

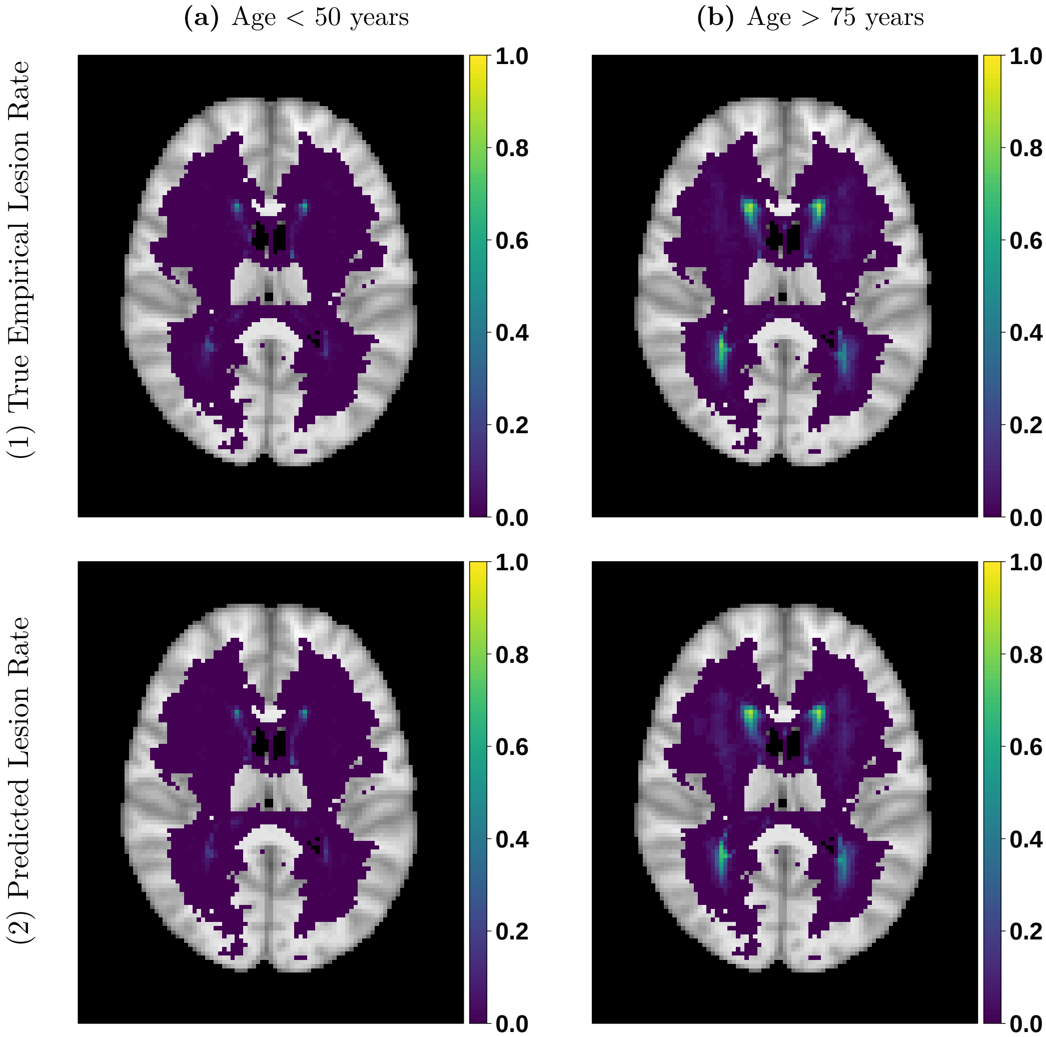


Figure B.12: Comparison of (1) true empirical lesion rates to (2) predicted lesion rates (estimated via posterior predictive) of 100 held-out test subjects with an age (1) of lower than 50 years and (b) of higher than 75 years. We show that our model is able to reflect differences in lesion prevalence among different age groups on held-out test data.

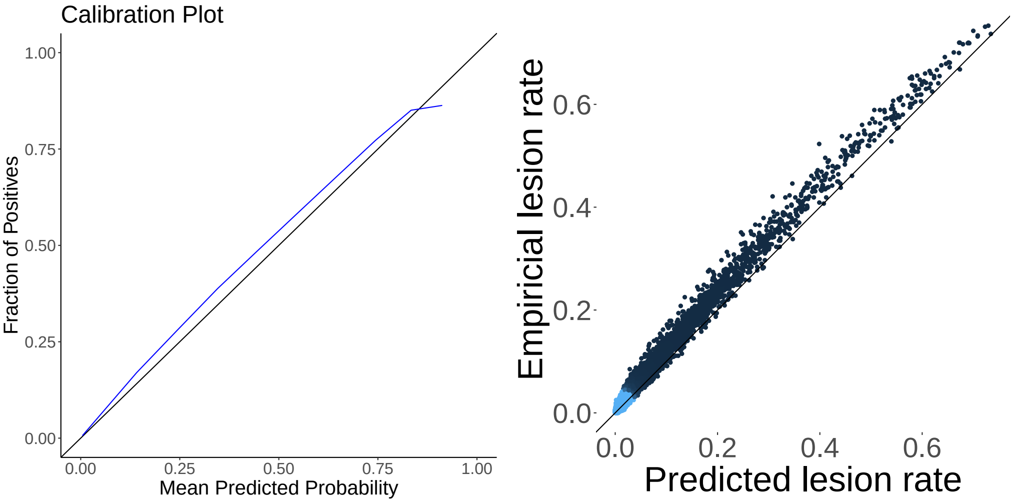


Figure B.13: Calibration plot and histogram of predicted probabilities from BB-BLESS for all voxels within lesion mask for 1000 out of sample subjects. The calibration curve showcases how well the average predicted probability for binned predictions is aligned with true fraction of lesions within that bin. The scatterplot on the right of the predicted lesion probabilities by BB-BLESS against the true empirical lesion rates at each voxel highlights the same trend as the calibration plots by identifying a good alignment between predicted and empirical lesion rates with a slight overestimation of predicted lesion rates compared to the empirical lesion rates.

In this section, we evaluate the posterior predictive distribution and model calibration using out-of-sample data. We firstly assess the model fit of BLESS with a calibration plot derived from 1,000 out-of-sample subjects across all voxels within the lesion mask from the UK Biobank. Figure B.13 shows that the mean predicted probability within each bin (calculated in steps of 0.1) over all voxels from BB-BLESS is well aligned with the fraction of subjects who truly exhibit a lesion over all out-of-sample subjects within a respective bin.

The scatterplot, on the right in Figure B.13, of the predicted lesion probabilities by BB-BLESS against the true empirical lesion rates at each voxel for the out-of-sample dataset highlights the same trend as the calibration plot to the left. The predicted and empirical lesion rates are aligned well along the 45 degree line with a slight overestimation of predicted lesion rates compared to the empirical lesion rates. We therefore show that our model is robust to violations in model assumptions by emphasising that BLESS is well calibrated in its uncertainty for the finite out-of-sample predictive performance. Moreover, by ensuring that our Bayesian model is well calibrated, we show that our model has desirable frequentist properties with respect to the estimation of uncertainty for finite out-of-sample sample sizes.

C

Appendix: Scalable Scalar-on-Image Cortical Surface Regression with a Relaxed-Thresholded Gaussian Process Prior

Contents

C.1 Derivation of Gibbs Sampling Algorithm

C.1.1 Joint Distribution

$$
\begin{aligned}
& \ln\left\{p\left(\boldsymbol{y}, \boldsymbol{X}, \tilde{\boldsymbol{\beta}}(\boldsymbol{s}), \boldsymbol{\alpha}(\boldsymbol{s}), \beta_0, \sigma_\epsilon^2, \sigma_\beta^2, \sigma_\alpha^2, a_\epsilon, a_\beta, a_\alpha\right)\right\} \\
= \; & \ln\{p\left(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\alpha}(\boldsymbol{s}), \boldsymbol{\theta}, \boldsymbol{\Psi}, \boldsymbol{\lambda}, \beta_0, \sigma_\epsilon^2\right) p(\beta_0|\sigma_{\beta_0}) p\left(\boldsymbol{\theta}|\sigma_\beta^2\right) p\left(\boldsymbol{\alpha}(\boldsymbol{s})|\boldsymbol{\theta}, \boldsymbol{\Psi}, \boldsymbol{\lambda}, \sigma_\alpha^2\right) \\
& p(\sigma_\epsilon|a_\epsilon)\; p(\sigma_\beta|a_\beta)\; p(\sigma_\alpha|a_\alpha)\; p(a_\epsilon|s_\epsilon^2)\; p(a_\beta|s_\beta^2)\; p(a_\alpha|s_\alpha^2)\} \\
= \; & -\frac{N}{2}\ln\{\sigma_\epsilon^2\} - \frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^N \left\{y_i - \beta_0 - \sum_{j=1}^M \sum_{l=1}^L \theta_l \sqrt{\lambda_l}\psi_l(s_j) I(|\alpha(s_j)| > \delta) x_{i,j}\right\}^2 \\
& - \frac{1}{2\sigma_{\beta_0}^2}\beta_0^2 - \frac{L}{2}\ln\left\{\sigma_\beta^2\right\} - \frac{1}{2\sigma_\beta^2}\sum_{l=1}^L \theta_l^2 \\
& - \frac{M}{2}\ln\left\{\sigma_\alpha^2\right\} - \frac{1}{2\sigma_\alpha^2}\sum_{j=1}^M \left\{\alpha(s_j) - \sum_{l=1}^L \theta_l \sqrt{\lambda_l}\psi_l(s_j)\right\}^2 \\
& + \frac{1}{2}\ln\{1/a_\beta\} - (1/2+1)\ln\left\{\sigma_\beta^2\right\} - \frac{1/a_\beta}{\sigma_\beta^2} \\
& + \frac{1}{2}\ln\{1/a_\epsilon\} - (1/2+1)\ln\left\{\sigma_\epsilon^2\right\} - \frac{1/a_\epsilon}{\sigma_\epsilon^2} \\
& + \frac{1}{2}\ln\{1/a_\alpha\} - (1/2+1)\ln\left\{\sigma_\alpha^2\right\} - \frac{1/a_\alpha}{\sigma_\alpha^2} \\
& + \frac{1}{2}\ln\{1/s_\beta^2\} - (1/2+1)\ln\left\{a_\beta\right\} - \frac{1/s_\beta^2}{a_\beta} \\
& + \frac{1}{2}\ln\{1/s_\epsilon^2\} - (1/2+1)\ln\left\{a_\epsilon\right\} - \frac{1/s_\epsilon^2}{a_\epsilon} \\
& + \frac{1}{2}\ln\{1/s_\alpha^2\} - (1/2+1)\ln\left\{a_\alpha\right\} - \frac{1/s_\alpha^2}{a_\alpha}
\end{aligned}
$$

C.1.2 Update β_0

$$
\begin{aligned}
& [\beta_0|\text{rest}] \sim \mathcal{N}\left(\mu_{\beta_0}, \Sigma_{\beta_0}\right) \\
& \quad \mu_{\beta_0} = \Sigma_{\beta_0}\sigma_\epsilon^{-2}\left[\sum_{i=1}^N \left(y_i - \sum_{j=1}^M \sum_{l=1}^L \theta_l \sqrt{\lambda_l}\psi_l(s_j) I(|\alpha(s_j)| > \delta) x_{i,j}\right)\right] \\
& \quad \Sigma_{\beta_0} = \left[N\sigma_\epsilon^{-2} + \sigma_{\beta_0}^{-2}\right]^{-1}
\end{aligned}
$$

C.1.3 Update θ

$$
\begin{aligned}
&[\boldsymbol{\theta}|\text{rest}] \sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)\\
&\quad \boldsymbol{\mu}_{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\left[\sigma_\epsilon^{-2}\left[\boldsymbol{y}-\beta_0\mathbb{1}_N\right]^T\left[\boldsymbol{X}\text{diag}\left\{I(|\alpha(s_j)|>\delta)\right\}_{j=1}^M\boldsymbol{\Psi}^T\text{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^L\right]+\right.\\
&\qquad\qquad \left.\sigma_\alpha^{-2}\boldsymbol{\alpha}(\boldsymbol{s})^T\boldsymbol{\Psi}^T\text{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^L\right]^T\\
&\quad \boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \left[\sigma_\epsilon^{-2}\left[\boldsymbol{X}\text{diag}\left\{I(|\alpha(s_j)|>\delta)\right\}_{j=1}^M\boldsymbol{\Psi}^T\text{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^L\right]^T\right.\\
&\qquad\qquad \left[\boldsymbol{X}\text{diag}\left\{I(|\alpha(s_j)|>\delta)\right\}_{j=1}^M\boldsymbol{\Psi}^T\text{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^L\right]+\sigma_\beta^{-2}\boldsymbol{I}+\\
&\qquad\qquad \left.\sigma_\alpha^{-2}\left[\boldsymbol{\Psi}^T\text{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^L\right]^T\left[\boldsymbol{\Psi}^T\text{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^L\right]\right]^{-1}
\end{aligned}
$$

C.1.4 Update α

$$
\begin{aligned}
[\alpha(s_j)|\text{rest}] \sim\ & w_{-1,j}\ \text{Truncated-Normal}_{(-\infty,-\delta)}\left(\sum_{l=1}^L\theta_l\sqrt{\lambda_l}\psi_l(s_j),\sigma_\alpha^2\right)+\\
& w_{0,j}\ \text{Truncated-Normal}_{(-\delta,\delta)}\left(\sum_{l=1}^L\theta_l\sqrt{\lambda_l}\psi_l(s_j),\sigma_\alpha^2\right)+\\
& w_{1,j}\ \text{Truncated-Normal}_{(\delta,\infty)}\left(\sum_{l=1}^L\theta_l\sqrt{\lambda_l}\psi_l(s_j),\sigma_\alpha^2\right)
\end{aligned}
$$

Constraint of $\sum_{k\in\{-1,0,1\}} w_{k,j} = 1$.

$$
\begin{aligned}
w_{-1,j} &\propto \Phi\left(-\frac{\left[\delta+\sum_{l=1}^L\theta_l\sqrt{\lambda_l}\psi_l(s_j)\right]}{\sigma_\alpha}\right)\ c_j\\
w_{0,j} &\propto \left[\Phi\left(\frac{\delta-\sum_{l=1}^L\theta_l\sqrt{\lambda_l}\psi_l(s_j)}{\sigma_\alpha}\right)-\Phi\left(-\frac{\left[\delta+\sum_{l=1}^L\theta_l\sqrt{\lambda_l}\psi_l(s_j)\right]}{\sigma_\alpha}\right)\ \right]\\
w_{1,j} &\propto \left[1-\Phi\left(\frac{\delta-\sum_{l=1}^L\theta_l\sqrt{\lambda_l}\psi_l(s_j)}{\sigma_\alpha}\right)\right]\ c_j
\end{aligned}
$$

$$
\begin{aligned}
c_j = \exp&\left\{-\frac{1}{2\sigma_\epsilon^2}\left[\sum_{l=1}^L\theta_l\sqrt{\lambda_l}\psi_l(s_j)\right]\left[\sum_{l=1}^L\theta_l\sqrt{\lambda_l}\psi_l(s_j)\sum_{i=1}^N x_{i,j}^2-\right.\right.\\
&\left.\left.2\sum_{i=1}^N\left(y_i-\sum_{j'\neq j}\sum_{l=1}^L\theta_l\sqrt{\lambda_l}\psi_l(s_{j'})I(|\alpha_{j'}|>\delta)x_{i,j'}-\beta_0\right)x_{i,j}\left(\sum_{l=1}\theta_l\sqrt{\lambda_l}\psi_l(s_j)\right)\right]\right\}
\end{aligned}
$$

C.1.5 Update σ_ϵ^2

$$\left[\sigma_\epsilon^2|\text{rest}\right] \sim \text{Inverse-Gamma}\Bigg\{1/2+\frac{N}{2}, 1/a_\epsilon + \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \left[\boldsymbol{X}\text{diag}\left\{I(|\alpha(s_j)|>\delta)\right\}_{j=1}^{M}\boldsymbol{\Psi}^T\text{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^{L}\right]^T\boldsymbol{\theta}-\beta_0\right)^2\Bigg\}$$

C.1.6 Update σ_β^2

$$\left[\sigma_\beta^2|\text{rest}\right] \sim \text{Inverse-Gamma}\left\{1/2+\frac{L}{2}, 1/a_\beta+\frac{1}{2}\sum_{l=1}^{L}\theta_l^2\right\}$$

C.1.7 Update σ_α^2

$$\left[\sigma_\alpha^2|\text{rest}\right] \sim \text{Inverse-Gamma}\left\{1/2+\frac{M}{2}, 1/a_\alpha+\frac{1}{2}\sum_{j=1}^{M}\left[\alpha(s_j)-\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\psi_l(s_j)\right]^2\right\}$$

C.1.8 Update a_ϵ

$$[a_\epsilon|\text{rest}] \sim \text{Inverse-Gamma}(1,\ \sigma_\epsilon^{-2}+s_\epsilon^{-2})$$

C.1.9 Update a_β

$$[a_\beta|\text{rest}] \sim \text{Inverse-Gamma}(1,\ \sigma_\beta^{-2}+s_\beta^{-2})$$

C.1.10 Update a_α

$$[a_\alpha|\text{rest}] \sim \text{Inverse-Gamma}(1,\ \sigma_\alpha^{-2}+s_\alpha^{-2})$$

C.1.11 Update δ (with discrete uniform prior)

Suppose δ follows a discrete uniform distribution on $\{\delta_1,\dots,\delta_{n_\delta}\}$, i.e.

$$\Pr(\delta=\delta_t)=\frac{1}{n_\delta}$$

$$
\begin{aligned}
\ln\{\pi(\delta|\cdot)\} &\propto \ln\{p(\boldsymbol{y}|\beta_0, \tilde{\boldsymbol{\beta}}(\boldsymbol{s}), \boldsymbol{\alpha}(\boldsymbol{s}), \delta, \sigma_\epsilon^2)p(\delta)\} \\
&\propto -\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{N}\left\{y_i - \beta_0 - \sum_{j=1}^{M}\tilde{\beta}(s_j)I(|\alpha(s_j)| > \delta)x_{i,j}\right\}^2 \\
&\quad + \ln\left\{\frac{1}{n_\delta}\right\}
\end{aligned}
$$

where n_δ is the number of prior options for threshold δ.

The conditional probability of $\delta = \delta_t$ is

$$
\Pr(\delta = \delta_t \mid \cdot) \propto \frac{1}{n_\delta}\exp\left(-\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{N}\left\{y_i - \beta_0 - \sum_{j=1}^{M}\tilde{\beta}(s_j)I(|\alpha(s_j)| > \delta_t)x_{i,j}\right\}^2\right)
$$

C.1.12 Update δ (with continuous uniform prior)

$$
\begin{aligned}
\ln\{\pi(\delta|\cdot)\} &\propto \ln\{p(\boldsymbol{y}|\beta_0, \tilde{\boldsymbol{\beta}}(\boldsymbol{s}), \boldsymbol{\alpha}(\boldsymbol{s}), \delta, \sigma_\epsilon^2)p(\delta)\} \\
&\propto -\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{N}\left\{y_i - \beta_0 - \sum_{j=1}^{M}\tilde{\beta}(s_j)I(|\alpha(s_j)| > \delta)x_{i,j}\right\}^2 \\
&\quad + \ln\left\{\frac{1}{t_{max} - t_{min}}\right\}
\end{aligned}
$$

where t_{min} and t_{max} are the lower and upper bound of the continuous uniform prior on the threshold δ.

Let $\tau_1, \ldots, \tau_M$ be a permutation for the indices $\{1, \ldots, M\}$, such that

$$
0 = |\alpha_{\tau_0}| < |\alpha_{\tau_1}| < |\alpha_{\tau_2}| < \cdots < |\alpha_{\tau_M}| < |\alpha_{\tau_{M+1}}| = \infty
$$

Then we can rewrite

$$
\begin{aligned}
\ln\{\pi(\delta|\cdot)\} &= C - \frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{N}\left\{y_i - \beta_0 - \sum_{j=1}^{M}\beta(s_{\tau_j})I(|\alpha_{\tau_j}| > \delta)x_{i,\tau_j}\right\}^2 \\
&= C - \left(\sum_{k=0}^{M} I\{\delta \in (|\alpha_{\tau_k}|, |\alpha_{\tau_{k+1}}|)\}\right)\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{N}\left\{y_i - \beta_0 - \sum_{j=1}^{M}\beta(s_{\tau_j})I(|\alpha_{\tau_j}| > \delta)x_{i,\tau_j}\right\}^2 \\
&= C - \sum_{k=0}^{M} I\{\delta \in (|\alpha_{\tau_k}|, |\alpha_{\tau_{k+1}}|)\}\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{N}\left\{y_i - \beta_0 - \sum_{j=k+1}^{M}\beta(s_{\tau_j})x_{i,\tau_j}\right\}^2 \\
&= C - \sum_{k=0}^{M} I\{\delta \in (|\alpha_{\tau_k}|, |\alpha_{\tau_{k+1}}|)Q_k
\end{aligned}
$$

where $\sum_{j=M+1}^{M}\beta(s_{\tau_j})x_{i,\tau_j}$ is defined as zero and

$$Q_k = \frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{N}\left\{y_i - \beta_0 - \sum_{j=k+1}^{M}\beta(s_{\tau_j})x_{i,\tau_j}\right\}^2$$

This further implies that

$$\pi(\delta|\cdot) \propto \sum_{k=0}^{M}\exp(-Q_k)I\{\delta\in(|\alpha_{\tau_k}|,|\alpha_{\tau_{k+1}}|)\} \tag{C.1}$$

$$\propto \sum_{k=0}^{M}\exp(-Q_k)(|\alpha_{k+1}|-|\alpha_k|)\frac{I\{\delta\in(|\alpha_{\tau_k}|,|\alpha_{\tau_{k+1}}|)\}}{(|\alpha_{t_{k+1}}|-|\alpha_{\tau_k}|)} \tag{C.2}$$

$$\begin{aligned}
\pi(\delta|\cdot) &= \sum_{k=0}^{M}\pi(|\alpha_{\tau_k}|\leq\delta\leq|\alpha_{\tau_{k+1}}|)\prod_{i=1}^{N}\pi(y_i||\alpha_{\tau_k}|\leq\delta\leq|\alpha_{\tau_{k+1}}|,\dots)\\
[\delta|\cdot] &\sim \sum_{k=0}^{M}w_k\text{Uniform}(|\alpha_{\tau_k}|,|\alpha_{\tau_{k+1}}|)\\
w_k &= \frac{exp(-Q_k)(|\alpha_{\tau_{k+1}}|-|\alpha_{\tau_k}|)}{\sum_{k=0}^{M}exp(-Q_k)(|\alpha_{\tau_{k+1}}|-|\alpha_{\tau_k}|)}
\end{aligned}$$

C.2 Derivation of Variational Inference Algorithm

Note that we add an additional set of parameters to this derivation to model the confounding variables within our real data analysis in Section 4.3.2 where we denote the input confounding variables with $\boldsymbol{z_i}\in\mathbb{R}^K$ for every subject $i=1,\dots,N$ and the attached parameter vector with $\boldsymbol{\beta^c}\in\mathbb{R}^K$. To fully specify our hierarchical Bayesian spatial model we place non-informative Normal priors on the coefficients modelling the confounding variables $\beta_k^c\sim\mathcal{N}(0,\sigma_{\beta^c}^2)$ for every $k=1,\dots,K$.

C.2.1 Joint Distribution

$$
\begin{aligned}
&\ln\left\{p\left(\boldsymbol{y},\boldsymbol{X},\tilde{\boldsymbol{\beta}}(\boldsymbol{s}),\boldsymbol{\alpha}(\boldsymbol{s}),\beta_0,\boldsymbol{\beta^c},\sigma^2_\epsilon,\sigma^2_\beta,\sigma^2_\alpha,a_\epsilon,a_\beta,a_\alpha\right)\right\}\\
=\;&\ln\Big\{p\left(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\alpha}(\boldsymbol{s}),\boldsymbol{\theta},\boldsymbol{\Psi},\boldsymbol{\lambda},\beta_0,\boldsymbol{\beta^c},\sigma^2_\epsilon\right)p(\beta_0|\sigma_{\beta_0})p\left(\boldsymbol{\theta}|\sigma^2_\beta\right)p\left(\boldsymbol{\alpha}(\boldsymbol{s})|\boldsymbol{\theta},\boldsymbol{\Psi},\boldsymbol{\lambda},\sigma^2_\alpha\right)\;p(\boldsymbol{\beta^c}|\sigma^2_{\beta^c})\\
&\quad p(\delta)\;p(\sigma_\epsilon|a_\epsilon)\;p(\sigma_\beta|a_\beta)\;p(\sigma_\alpha|a_\alpha)\;p(a_\epsilon|s^2_\epsilon)\;p(a_\beta|s^2_\beta)\;p(a_\alpha|s^2_\alpha)\Big\}\\
=\;&-\frac{N}{2}\ln\{\sigma^2_\epsilon\}-\frac{1}{2\sigma^2_\epsilon}\sum_{i=1}^{N}\left\{y_i-\beta_0-\sum_{j=1}^{M}\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\psi_l(s_j)I(|\alpha(s_j)|>\delta)x_{i,j}-\sum_{k=1}^{K}z_k\beta^c_k\right\}^2\\
&-\frac{1}{2\sigma^2_{\beta_0}}\beta_0^2-\frac{1}{2\sigma^2_{\beta^c}}\sum_{k=1}^{K}\beta_k^{2,k}-\frac{L}{2}\ln\left\{\sigma^2_\beta\right\}-\frac{1}{2\sigma^2_\beta}\sum_{l=1}^{L}\theta_l^2\\
&-\frac{M}{2}\ln\left\{\sigma^2_\alpha\right\}-\frac{1}{2\sigma^2_\alpha}\sum_{j=1}^{M}\left\{\alpha(s_j)-\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\psi_l(s_j)\right\}^2\\
&+\frac{1}{2}\ln\{1/a_\beta\}-(1/2+1)\ln\left\{\sigma^2_\beta\right\}-\frac{1/a_\beta}{\sigma^2_\beta}\\
&+\frac{1}{2}\ln\{1/a_\epsilon\}-(1/2+1)\ln\left\{\sigma^2_\epsilon\right\}-\frac{1/a_\epsilon}{\sigma^2_\epsilon}\\
&+\frac{1}{2}\ln\{1/a_\alpha\}-(1/2+1)\ln\left\{\sigma^2_\alpha\right\}-\frac{1/a_\alpha}{\sigma^2_\alpha}\\
&+\frac{1}{2}\ln\{1/s^2_\beta\}-(1/2+1)\ln\left\{a_\beta\right\}-\frac{1/s^2_\beta}{a_\beta}\\
&+\frac{1}{2}\ln\{1/s^2_\epsilon\}-(1/2+1)\ln\left\{a_\epsilon\right\}-\frac{1/s^2_\epsilon}{a_\epsilon}\\
&+\frac{1}{2}\ln\{1/s^2_\alpha\}-(1/2+1)\ln\left\{a_\alpha\right\}-\frac{1/s^2_\alpha}{a_\alpha}
\end{aligned}
$$

C.2.2 Update β_0

$$
\begin{aligned}
\ln\{q^*(\beta_0)\}&\propto\mathbb{E}_{q(\boldsymbol{\theta},\boldsymbol{\alpha}(\boldsymbol{s}),\boldsymbol{\beta^c},\delta,\sigma^2_\epsilon,\sigma^2_\beta,\sigma^2_\alpha,a_\epsilon,a_\beta,a_\alpha)}[\ln\{p(\boldsymbol{y},\beta_0,\boldsymbol{\beta^c},\boldsymbol{\theta},\boldsymbol{\alpha}(\boldsymbol{s}),\delta,\sigma^2_\epsilon,\sigma^2_\beta,\sigma^2_\alpha,a_\epsilon,a_\beta,a_\alpha)\}]\\
&\propto\mathbb{E}_{q(\boldsymbol{\theta},\boldsymbol{\alpha}(\boldsymbol{s}),\boldsymbol{\beta^c},\delta,\sigma^2_\epsilon)}[\ln\{p(\boldsymbol{y}|\beta_0,\boldsymbol{\theta},\boldsymbol{\alpha}(\boldsymbol{s}),\boldsymbol{\beta^c},\delta,\sigma^2_\epsilon)p(\beta_0)\}]\\
&\propto\mathbb{E}_{q(\boldsymbol{\theta},\boldsymbol{\alpha}(\boldsymbol{s}),\delta,\boldsymbol{\beta^c},\sigma^2_\epsilon)}[\sum_{i=1}^{N}\ln\{\mathcal{N}(y_i|\beta_0+\sum_{k=1}^{K}z_k\beta^c_k+\sum_{j=1}^{M}\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\Psi_l(s_j)I(|\alpha(s_j)|>\delta),\sigma^2_\epsilon)\}]\\
&+\ln\{\mathcal{N}(\beta_0|0,\sigma^2_{\beta_0}\}\\
&\propto-\frac{1}{2}\mathbb{E}_{q(\sigma^2_\epsilon)}\left[\frac{1}{\sigma^2_\epsilon}\right]\left[\beta_0\sum_{i=1}^{N}\left[y_i-\sum_{k=1}^{K}z_k\mathbb{E}_{q(\beta^c_k)}[\beta^c_k]-\right.\right.\\
&\left.\left.\sum_{j=1}^{M}\sum_{l=1}^{L}\mathbb{E}_{q(\theta_l)}[\theta_l]\sqrt{\lambda_l}\Psi_l(s_j)\mathbb{E}_{q(\alpha(s_j))}[I(|\alpha(s_j)|>\delta)]x_{i,j}\right]+N\beta_0^2\right]-\frac{1}{2\sigma^2_{\beta_0}}\beta_0^2
\end{aligned}
$$

$$
\begin{aligned}
q^*(\beta_0) &= \mathcal{N}(\beta_0; \mu_{\beta_0}, \Sigma_{\beta_0}) \\
\mu_{\beta_0} &= \Sigma_{\beta_0} \left[\mathbb{E}_{q(\sigma_\epsilon^2)} \left[\frac{1}{\sigma_\epsilon^2} \right] \sum_{i=1}^{N} \left\{ y_i - \sum_{k=1}^{K} z_k \mathbb{E}_{q(\beta_k^c)}[\beta_k^c] \right. \right. \\
&\quad \left. \left. - \sum_{j=1}^{M} \sum_{l=1}^{L} \mathbb{E}_{q(\theta_l)}[\theta_l] \sqrt{\lambda_l} \Psi_l(s_j) \mathbb{E}_{q(\alpha(s_j))}[I(|\alpha(s_j)| > \delta)] x_{i,j} \right\} \right] \\
&= \Sigma_{\beta_0} \left[\mathbb{E}_{q(\sigma_\epsilon^2)} \left[\frac{1}{\sigma_\epsilon^2} \right] \sum_{i=1}^{N} \left\{ y_i - \sum_{k=1}^{K} z_k \mathbb{E}_{q(\beta_k^c)}[\beta_k^c] \right. \right. \\
&\quad \left. \left. - \sum_{j=1}^{M} \sum_{l=1}^{L} \mathbb{E}_{q(\theta_l)}[\theta_l] \sqrt{\lambda_l} \Psi_l(s_j) [w_{1,j} + w_{-1,j}] x_{i,j} \right\} \right] \\
\Sigma_{\beta_0} &= [N \mathbb{E}_{q(\sigma_\epsilon^2)} \left[\frac{1}{\sigma_\epsilon^2} \right] + \frac{1}{\sigma_{\beta_0}^2}]^{-1}
\end{aligned}
$$

C.2.3 Update θ

$$
\begin{aligned}
\ln\{q^*(\boldsymbol{\theta})\} &\propto \mathbb{E}_{q(\beta_0, \boldsymbol{\beta^c}, \boldsymbol{\alpha(s)}, \delta, \sigma_\epsilon^2, \sigma_\beta^2, \sigma_\alpha^2, a_\epsilon, a_\beta, a_\alpha)}[\ln\{p(\boldsymbol{y}, \beta_0, \boldsymbol{\beta^c}, \boldsymbol{\theta}, \boldsymbol{\alpha(s)}, \delta, \sigma_\epsilon^2, \sigma_\beta^2, \sigma_\alpha^2, a_\epsilon, a_\beta, a_\alpha)\}] \\
&\propto \mathbb{E}_{q(\beta_0, \boldsymbol{\beta^c}, \boldsymbol{\alpha(s)}, \delta, \sigma_\epsilon^2, \sigma_\beta^2, \sigma_\alpha^2)}[\ln\{p(\boldsymbol{y}|\beta_0, \boldsymbol{\beta^c}, \boldsymbol{\theta}, \boldsymbol{\alpha(s)}, \delta, \sigma_\epsilon^2) p(\boldsymbol{\theta}|\sigma_\beta^2) p(\boldsymbol{\alpha(s)}|\boldsymbol{\theta}, \sigma_\alpha^2)\}] \\
&\propto \mathbb{E}_{q(\beta_0, \boldsymbol{\beta^c}, \boldsymbol{\alpha(s)}, \delta, \sigma_\epsilon^2)} \left[\sum_{i=1}^{N} \ln\{\mathcal{N}(y_i | \beta_0 + \sum_{k=1}^{K} z_k \beta_k^c \right. \\
&\quad \left. + \sum_{j=1}^{M} \sum_{l=1}^{L} \theta_l \sqrt{\lambda_l} \Psi_l(s_j) I(|\alpha(s_j)| > \delta) x_{i,j}, \sigma_\epsilon^2)\} \right] \\
&\quad + \mathbb{E}_{q(\sigma_\beta^2)} \left[\sum_{l=1}^{L} \ln\{\mathcal{N}(\theta_l | 0, \sigma_\beta^2)\} \right] + \mathbb{E}_{q(\sigma_\alpha^2)} \left[\sum_{j=1}^{M} \mathcal{N}(\alpha(s_j) | \sum_{l=1}^{L} \theta_l \sqrt{\lambda_l} \Psi_l(s_j), \sigma_\alpha^2) \right] \\
&\propto \mathbb{E}_{q(\beta_0, \boldsymbol{\beta^c}, \boldsymbol{\alpha(s)}, \delta, \sigma_\epsilon^2)} \left[-\frac{1}{2\sigma_\epsilon^2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{k=1}^{K} z_k \beta_k^c \right. \right. \\
&\quad \left. \left. - \sum_{j=1}^{M} \sum_{l=1}^{L} \theta_l \sqrt{\lambda_l} \Psi_l(s_j) I(|\alpha(s_j)| > \delta) x_{i,j} \right)^2 \right] \\
&\quad - \mathbb{E}_{q(\sigma_\beta^2)} \left[\frac{1}{2\sigma_\beta^2} \right] \sum_{l=1}^{L} \theta_l^2 + \mathbb{E}_{q(\boldsymbol{\theta}, \sigma_\alpha^2)} \left[-\frac{1}{2\sigma_\alpha^2} \sum_{j=1}^{M} (\alpha(s_j) - \sum_{l=1}^{L} \theta_l \sqrt{\lambda_l} \Psi_l(s_j))^2 \right] \\
&\propto -\frac{1}{2} \mathbb{E}_{q(\sigma_\epsilon^2)} \left[\frac{1}{\sigma_\epsilon^2} \right] \sum_{i=1}^{N} \left[2 \left(y_i - \mathbb{E}_{q(\beta_0)}[\beta_0] - \sum_{k=1}^{K} z_k \mathbb{E}_{q(\beta_k^c)}[\beta_k^c] \right) \right. \\
&\quad \sum_{j=1}^{M} \sum_{l=1}^{L} \theta_l \sqrt{\lambda_l} \Psi_l(s_j) \mathbb{E}_{q(\boldsymbol{\alpha(s)})} \left[I(|\alpha(s_j)| > \delta) \right] x_{i,j}
\end{aligned}
$$

$$
\begin{aligned}
&+ \mathbb{E}_{q(\boldsymbol{\alpha}(\boldsymbol{s}))}\left[\left(\sum_{j=1}^{M}\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\Psi_l(s_j)I(|\alpha(s_j)|>\delta)x_{i,j}\right)^2\right]\Bigg] \\
&- \mathbb{E}_{q(\sigma_\beta^2)}\left[\frac{1}{2\sigma_\beta^2}\right]\sum_{l=1}^{L}\theta_l^2 - \frac{1}{2}\mathbb{E}_{q(\sigma_\alpha^2)}\left[\frac{1}{\sigma_\alpha^2}\right]\sum_{j=1}^{M}\left[-2\mathbb{E}_{q(\boldsymbol{\alpha}(\boldsymbol{s}))}[\alpha(s_j)]\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\Psi_l(s_j)+\left(\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\Psi_l(s_j)\right)^2\right]
\end{aligned}
$$

$$
\begin{aligned}
q^*(\boldsymbol{\theta}) &= \mathcal{N}(\boldsymbol{\theta};\boldsymbol{\mu}_\theta,\boldsymbol{\Sigma}_\theta) \\
\boldsymbol{\mu}_\theta &= \boldsymbol{\Sigma}_\theta\Bigg[\mathbb{E}_{q(\sigma_\epsilon^2)}\left[\frac{1}{\sigma_\epsilon^2}\right]\left[\boldsymbol{y}-\mathbb{E}_{q(\beta_0)}\left[\beta_0\right]\mathbb{1}-\sum_{k=1}^{K}z_k\mathbb{E}_{q(\beta_k^c)}[\beta_k^c]\right]^T \\
&\quad\left[\boldsymbol{X}\mathrm{diag}\left\{\mathbb{E}_{q(\alpha(s_j))}\left[I(|\alpha(s_j)|>\delta)\right]\right\}_{j=1}^{M}\boldsymbol{\Psi}^T\mathrm{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^{L}\right] \\
&\quad+\mathbb{E}_{q(\sigma_\alpha^2)}\left[\frac{1}{\sigma_\alpha^2}\right]\mathbb{E}_{q(\boldsymbol{\alpha}(\boldsymbol{s}))}\left[\boldsymbol{\alpha}(\boldsymbol{s})\right]^T\boldsymbol{\Psi}^T\mathrm{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^{L}\Bigg]^T \\
\boldsymbol{\Sigma}_\theta &= \Bigg[\mathbb{E}_{q(\sigma_\epsilon^2)}\left[\frac{1}{\sigma_\epsilon^2}\right]\left[\boldsymbol{X}\mathrm{diag}\left\{\mathbb{E}_{q(\alpha(s_j))}\left[I(|\alpha(s_j)|>\delta)\right]\right\}_{j=1}^{M}\boldsymbol{\Psi}^T\mathrm{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^{L}\right]^T \\
&\quad\left[\boldsymbol{X}\mathrm{diag}\left\{\mathbb{E}_{q(\alpha(s_j))}\left[I(|\alpha(s_j)|>\delta)\right]\right\}_{j=1}^{M}\boldsymbol{\Psi}^T\mathrm{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^{L}\right]+\mathbb{E}_{q(\sigma_\beta^2)}\left[\frac{1}{\sigma_\beta^2}\right]\boldsymbol{I} \\
&\quad+\mathbb{E}_{q(\sigma_\alpha^2)}\left[\frac{1}{\sigma_\alpha^2}\right]\left[\boldsymbol{\Psi}^T\mathrm{diag}\{\sqrt{\lambda_l}\}_{l=1}^{L}\right]^T\left[\boldsymbol{\Psi}^T\mathrm{diag}\{\sqrt{\lambda_l}\}_{l=1}^{L}\right]\Bigg]^{-1}
\end{aligned}
$$

C.2.4 Update α

$$
\begin{aligned}
\ln\{q^*(\boldsymbol{\alpha}(\boldsymbol{s}))\} &\propto \mathbb{E}_{q(\beta_0,\boldsymbol{\beta^c},\boldsymbol{\theta},\delta,\sigma_\epsilon^2,\sigma_\beta^2,\sigma_\alpha^2,a_\epsilon,a_\beta,a_\alpha)}[\ln\{p(\boldsymbol{y},\beta_0,\boldsymbol{\beta^c},\boldsymbol{\theta},\boldsymbol{\alpha}(\boldsymbol{s}),\delta,\sigma_\epsilon^2,\sigma_\beta^2,\sigma_\alpha^2,a_\epsilon,a_\beta,a_\alpha)\}] \\
&\propto \mathbb{E}_{q(\beta_0,\boldsymbol{\beta^c},\delta,\theta,\sigma_\epsilon^2,\sigma_\alpha^2)}[\ln\{p(\boldsymbol{y}|\beta_0,\boldsymbol{\beta^c},\boldsymbol{\alpha}(\boldsymbol{s}),\boldsymbol{\theta},\delta,\sigma_\epsilon^2)p(\boldsymbol{\alpha}(\boldsymbol{s})|\boldsymbol{\theta},\sigma_\alpha^2)\}] \\
&\propto \mathbb{E}_{q(\beta_0,\boldsymbol{\beta^c},\theta,\delta,\sigma_\epsilon^2)}\left[\sum_{i=1}^{N}\ln\{\mathcal{N}(y_i|\beta_0+\sum_{k=1}^{K}z_k\beta_k^c+\sum_{j=1}^{M}\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\Psi_l(s_j)I(|\alpha(s_j)|>\delta)x_{i,j},\sigma_\epsilon^2)\}\right] \\
&+ \mathbb{E}_{q(\theta,\sigma_\alpha^2)}\left[\sum_{j=1}^{M}\ln\{\mathcal{N}(\alpha(s_j)|\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\Psi_l(s_j),\sigma_\alpha^2)\}\right] \\
&\propto \mathbb{E}_{q(\beta_0,\boldsymbol{\beta^c},\boldsymbol{\alpha}(\boldsymbol{s}),\delta,\sigma_\epsilon^2)}\Bigg[-\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{N}\Bigg(y_i-\beta_0-\sum_{k=1}^{K}z_k\mathbb{E}_{q(\beta_k^c)}[\beta_k^c] \\
&-\sum_{j=1}^{M}\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\Psi_l(s_j)I(|\alpha(s_j)|>\delta)x_{i,j}\Bigg)^2\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&- \mathbb{E}_{q(\sigma^2_\beta)} \left[\frac{1}{2\sigma^2_\beta}\right] \sum_{l=1}^{L} \theta_l^2 + \mathbb{E}_{q(\boldsymbol{\theta},\sigma^2_\alpha)} \left[-\frac{1}{2\sigma^2_\alpha} \sum_{j=1}^{M} (\alpha(s_j) - \sum_{l=1}^{L} \theta_l \sqrt{\lambda_l} \Psi_l(s_j))^2\right] \\
&\propto \mathbb{E}_{q(\beta_0,\beta^c,\theta,\delta,\sigma^2_\epsilon)} \left[-\frac{1}{2\sigma^2_\epsilon} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{k=1}^{K} z_k \mathbb{E}_{q(\beta^c_k)}[\beta^c_k] - \sum_{j=1}^{M}\sum_{l=1}^{L} \theta_l \sqrt{\lambda_l} \Psi_l(s_j) I(|\alpha(s_j)| > \delta) x_{i,j}\right)^2\right] \\
&+ \mathbb{E}_{q(\theta,\sigma^2_\alpha)} \left[-\frac{1}{2\sigma^2_\alpha} \sum_{j=1}^{M} (\alpha(s_j) - \sum_{l=1}^{L} \theta_l \sqrt{\lambda_l} \Psi_l(s_j))^2\right] \\
&\propto -\frac{1}{2}\mathbb{E}_{q(\sigma^2_\epsilon)} \left[\frac{1}{\sigma^2_\epsilon}\right] \sum_{i=1}^{N} \left\{-2(y_i - \mathbb{E}_{q(\beta_0)}[\beta_0] - \sum_{k=1}^{K} z_k \mathbb{E}_{q(\beta^c_k)}[\beta^c_k])\right. \\
&\left.\sum_{j=1}^{M}\sum_{l=1}^{L} \mathbb{E}_{q(\theta)}[\theta_l] \sqrt{\lambda_l} \Psi_l(s_j) I(|\alpha(s_j)| > \delta) x_{i,j} + \mathbb{E}_{q(\theta)} \left[\left(\sum_{j=1}^{M}\sum_{l=1}^{L} \theta_l \sqrt{\lambda_l} \Psi_l(s_j) I(|\alpha(s_j)| > \delta) x_{i,j}\right)^2\right]\right\} \\
&- \frac{1}{2}\mathbb{E}_{q(\sigma^2_\alpha)} \left[\frac{1}{\sigma^2_\alpha}\right] \sum_{j=1}^{M} \left\{\alpha(s_j)^2 - 2\alpha(s_j) \sum_{l=1}^{L} \mathbb{E}_{q(\theta)}[\theta_l] \sqrt{\lambda_l} \Psi_l(s_j)\right\} \\
&\propto I(|\alpha(s_j)| > \delta) \left[\mathbb{E}_{q(\sigma^2_\epsilon)} \left[\frac{1}{\sigma^2_\epsilon}\right] \sum_{i=1}^{N} (y_i - \mathbb{E}_{q(\beta_0)}[\beta_0] - \sum_{k=1}^{K} z_k \mathbb{E}_{q(\beta^c_k)}[\beta_k]) \mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)] x_{i,j} - \right. \\
&\left.\frac{1}{2}\mathbb{E}_{q(\sigma^2_\epsilon)} \left[\frac{1}{\sigma\epsilon^2}\right] \sum_{i=1}^{N} \left[2\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)] x_{i,j} \sum_{j' \neq j}^{M} \mathbb{E}_{q(\theta)}[\beta(s_{j'})] \mathbb{E}_{q(\alpha_{j'})}[I(|\alpha_{j'}| > \delta)] x_{i,j'} + \mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)^2] x^2_{i,j}\right]\right] \\
&- \frac{1}{2}\mathbb{E}_{q(\sigma^2_\alpha)} \left[\frac{1}{\sigma^2_\alpha}\right] \sum_{j=1}^{M} \left\{\alpha(s_j)^2 - 2\alpha(s_j) \sum_{l=1}^{L} \mathbb{E}_{q(\theta)}[\theta_l] \sqrt{\lambda_l} \Psi_l(s_j)\right\}
\end{aligned}
$$

$$
\begin{aligned}
q^*(\alpha(s_j)) = w^*_{-1,j} \ &\text{Truncated-Normal}_{(-\infty,-\delta)} \left(\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)], 1/\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma_\alpha^{-2}]\right) + \\
w^*_{0,j} \ &\text{Truncated-Normal}_{(-\delta,\delta)} \left(\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)], 1/\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma_\alpha^{-2}]\right) + \\
w^*_{1,j} \ &\text{Truncated-Normal}_{(\delta,\infty)} \left(\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)], 1/\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma_\alpha^{-2}]\right)
\end{aligned}
$$

$$
\begin{aligned}
w^*_{-1,j} &\propto \Phi\left(-\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}(\sigma_\alpha^{-2})} \left[\delta + \mathbb{E}_{q(\theta)}\{\tilde{\beta}(s_j)\}\right]\right) \ c^*_j \\
w^*_{0,j} &\propto \left[\Phi\left(\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}(\sigma_\alpha^{-2})} \left[\delta - \mathbb{E}_{q(\theta)}\{\tilde{\beta}(s_j)\}\right]\right) - \Phi\left(-\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}(\sigma_\alpha^{-2})} \left[\delta + \mathbb{E}_{q(\theta)}\{\tilde{\beta}(s_j)\}\right]\right) \ \right] \\
w^*_{1,j} &\propto \Phi\left(-\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}(\sigma_\alpha^{-2})} \left[\delta - \mathbb{E}_{q(\theta)}\{\tilde{\beta}(s_j)\}\right]\right) \ c^*_j
\end{aligned}
$$

where

$$
\begin{aligned}
c_j^* = {} & \mathbb{E}_{q(\sigma_\epsilon^2)}\left[\frac{1}{\sigma_\epsilon^2}\right]\sum_{i=1}^{N}(y_i - \mathbb{E}_{q(\beta_0)}[\beta_0] - \sum_{k=1}^{K} z_k \mathbb{E}_{q(\beta_k^c)}[\beta_k^c])\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]x_{i,j} - \\
& \frac{1}{2}\mathbb{E}_{q(\sigma_\epsilon^2)}\left[\frac{1}{\sigma_\epsilon^2}\right]\sum_{i=1}^{N}\left[2\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]x_{i,j}\sum_{j'\neq j}^{M}\mathbb{E}_{q(\theta)}[\beta(s_{j'})]\mathbb{E}_{q(\alpha_{j'})}[I(|\alpha_{j'}| > \delta)]x_{i,j'} + \mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)^2]x_{i,j}^2\right]
\end{aligned}
$$

C.2.5 Update δ

Assign δ a discrete uniform prior.

For example, given a set of possible values of δ, i.e. $\{\delta_k\}_{k=1}^K$, let

$$
\mathbb{P}(\delta = \delta_k) = \frac{1}{K},
$$

Then we can assume the q density of δ is also discrete. Then we have for $k = 1, \ldots, K$

$$
\begin{aligned}
& \log \mathbb{P}_{q^*(\delta)}(\delta = \delta_k) \\
& \propto -\frac{1}{2}\mathbb{E}_{q(\sigma_\epsilon^2)}\left[\frac{1}{\sigma_\epsilon^2}\right]\sum_{i=1}^{N}\Bigg\{-2(y_i - \mathbb{E}_{q(\beta_0)}[\beta_0] - \sum_{k=1}^{K} z_k \mathbb{E}_{q(\beta_k^c)}[\beta_k]) \\
& \sum_{j=1}^{M}\sum_{l=1}^{L}\mathbb{E}_{q(\theta)}[\theta_l]\sqrt{\lambda_l}\Psi_l(s_j)\mathbb{E}_{q(\alpha)}\{I(|\alpha(s_j)| > l_k)\}x_{i,j} + \\
& \mathbb{E}_{q(\theta,\alpha)}\left[\left(\sum_{j=1}^{M}\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\Psi_l(s_j)I(|\alpha(s_j)| > l_k)x_{i,j}\right)^2\right]\Bigg\}
\end{aligned}
$$

C.2.6 Update σ_ϵ^2

$$
\begin{aligned}
\ln\{q^*(\sigma_\epsilon^2)\} & \propto \mathbb{E}_{q(\beta_0, \boldsymbol{\beta^c}, \boldsymbol{\theta}, \boldsymbol{\alpha(s)}, \delta, \sigma_\beta^2, \sigma_\alpha^2, a_\epsilon, a_\beta, a_\alpha)}[\ln\{p(\boldsymbol{y}, \beta_0, \boldsymbol{\beta^c}, \boldsymbol{\theta}, \boldsymbol{\alpha(s)}, \delta, \sigma_\epsilon^2, \sigma_\beta^2, \sigma_\alpha^2, a_\epsilon, a_\beta, a_\alpha)\}] \\
& \propto \mathbb{E}_{q(\beta_0, \boldsymbol{\beta^c}, \theta, \alpha, \delta, a_\epsilon)}\Big[\ln\{p(\boldsymbol{y}|\beta_0, \boldsymbol{\beta^c}, \boldsymbol{\theta}, \boldsymbol{\alpha(s)}, \delta, \sigma_\epsilon^2)p(\sigma_\epsilon^2|a_\epsilon)\}\Big]
\end{aligned}
$$

$$
\begin{aligned}
&\propto \mathbb{E}_{q(\beta_0,\boldsymbol{\beta^c},\theta,\alpha,\delta)} \left[\sum_{i=1}^{N} \ln\{\mathcal{N}(y_i|\beta_0 + \sum_{k=1}^{K} z_k\beta_k^c + \sum_{j=1}^{M}\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\Psi_l(s_j)I(|\alpha(s_j)|>\delta)x_{i,j}, \sigma_\epsilon^2)\}\right] \\
&+ \mathbb{E}_{q(a_\epsilon)}\left[\ln\left\{\text{Inverse-Gamma}\left(\sigma_\epsilon^2; \frac{1}{2}, \frac{1}{a_\epsilon}\right)\right\}\right] \\
&\propto -\frac{N}{2}\ln\{\sigma_\epsilon^2\} - \frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{N}\Bigg\{ y_i^2 + \mathbb{E}_{q(\beta_0)}[\beta_0^2] + \mathbb{E}_{q(\alpha,\theta,\delta)}\left[\left(\sum_{j=1}^{M}\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\Psi_l(s_j)I(|\alpha(s_j)|>\delta)x_{i,j}\right)^2\right] \\
&+ \sum_{k=1}^{K}\mathbb{E}_{q(\beta_k^c)}[\beta_k^{c2}]z_k^2 - 2(y_i - \mathbb{E}_{q(\beta_0)}[\beta_0] - \sum_{k=1}^{K} z_k\mathbb{E}_{q(\beta_k^c)}[\beta_k^c]) \\
&\mathbb{E}_{q(\alpha,\theta,\delta)}\left[\sum_{j=1}^{M}\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\Psi_l(s_j)I(|\alpha(s_j)|>\delta)x_{i,j}\right] - 2y_i(\mathbb{E}_{q(\beta_0)}[\beta_0] + \sum_{k=1}^{K} z_k\mathbb{E}_{q(\beta_k^c)}[\beta_k^c])\Bigg\} \\
&- \left(\frac{1}{2}+1\right)\ln\{\sigma_\epsilon^2\} - \frac{\mathbb{E}_{q(a_\epsilon)}\left[\frac{1}{a_\epsilon}\right]}{\sigma_\epsilon^2}
\end{aligned}
$$

$$
\begin{aligned}
q^*(\sigma_\epsilon^2) &= \text{Inverse-Gamma}(\sigma_\epsilon^2; b_\epsilon, c_\epsilon) \\
b_\epsilon &= \frac{N}{2} + \frac{1}{2} \\
c_\epsilon &= \Bigg[\mathbb{E}_{q(a_\epsilon)}\left[\frac{1}{a_\epsilon}\right] + \frac{1}{2}\sum_{i=1}^{N}\Big\{ y_i^2 + \mathbb{E}_{q(\beta_0)}[\beta_0^2] \\
&\quad + \mathbb{E}_{q(\alpha,\theta,\delta)}\left[\left(\sum_{j=1}^{M}\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\Psi_l(s_j)I(|\alpha(s_j)|>\delta)x_{i,j}\right)^2\right] \\
&\quad + \sum_{k=1}^{K}\mathbb{E}_{q(\beta_k^c)}[\beta_k^{c2}]z_k^2 - 2(y_i - \mathbb{E}_{q(\beta_0)}[\beta_0] - \sum_{k=1}^{K} z_k\mathbb{E}_{q(\beta_k^c)}[\beta_k^c]) \\
&\quad \mathbb{E}_{q(\alpha,\theta,\delta)}\left[\sum_{j=1}^{M}\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\Psi_l(s_j)I(|\alpha(s_j)|>\delta)x_{i,j}\right] - 2y_i(\mathbb{E}_{q(\beta_0)}[\beta_0] + \sum_{k=1}^{K} z_k\mathbb{E}_{q(\beta_k^c)}[\beta_k^c])\Big\}\Bigg]
\end{aligned}
$$

C.2.7 Update σ_β^2

$$
\begin{aligned}
\ln\{q^*(\sigma_\beta^2)\} &\propto \mathbb{E}_{q(\beta_0,\boldsymbol{\theta},\boldsymbol{\alpha(s)},\delta,\sigma_\epsilon^2,\sigma_\alpha^2,a_\epsilon,a_\beta,a_\alpha)}[\ln\{p(\boldsymbol{y},\beta_0,\boldsymbol{\theta},\boldsymbol{\alpha(s)},\delta,\sigma_\epsilon^2,\sigma_\beta^2,\sigma_\alpha^2,a_\epsilon,a_\beta,a_\alpha)\}] \\
&\propto \mathbb{E}_{q(\theta,a_\beta)}\left[\ln\left\{p(\theta|\sigma_\beta^2)p(\sigma_\beta^2|a_\beta)\right\}\right] \\
&\propto -\frac{L}{2}\ln\{\sigma_\beta^2\} - \frac{1}{2\sigma_\beta^2}\sum_{l=1}^{L}\mathbb{E}_{q(\theta)}[\theta_l^2] - \left(\frac{1}{2}+1\right)\ln\{\sigma_\beta^2\} - \frac{\mathbb{E}_{q(a_\beta)}\left[\frac{1}{a_\beta}\right]}{\sigma_\beta^2}
\end{aligned}
$$

$$
\begin{aligned}
q^*(\sigma_\beta^2) &= \text{Inverse-Gamma}(\sigma_\beta^2; b_\beta, c_\beta) \\
b_\beta &= \frac{L}{2} + \frac{1}{2} \\
c_\beta &= \left[\frac{1}{2}\sum_{l=1}^{L} \mathbb{E}_{q(\theta)}[\theta_l^2] + \mathbb{E}_{q(a_\beta)}\left[\frac{1}{a_\beta}\right]\right]
\end{aligned}
$$

C.2.8 Update σ_α^2

$$
\begin{aligned}
\ln\{q^*(\sigma_\alpha^2)\} &\propto \mathbb{E}_{q(\beta_0,\boldsymbol{\theta},\boldsymbol{\alpha}(\boldsymbol{s}),\delta,\sigma_\epsilon^2,\sigma_\beta^2,a_\epsilon,a_\beta,a_\alpha)}[\ln\{p(\boldsymbol{y},\beta_0,\boldsymbol{\theta},\boldsymbol{\alpha}(\boldsymbol{s}),\delta,\sigma_\epsilon^2,\sigma_\beta^2,\sigma_\alpha^2,a_\epsilon,a_\beta,a_\alpha)\}] \\
&\propto \mathbb{E}_{q(\alpha,\theta,\delta,a_\alpha)}\left[\ln\left\{p(\alpha|\theta,\sigma_\alpha^2)p(\sigma_\alpha^2|a_\alpha)\right\}\right] \\
&\propto -\frac{M}{2}\ln\{\sigma_\alpha^2\} - \frac{1}{2\sigma_\alpha^2}\sum_{j=1}^{M}\Bigg\{\mathbb{E}_{q(\alpha)}[\alpha(s_j)^2] - 2\mathbb{E}_{q(\alpha)}[\alpha(s_j)]\sum_{l=1}^{L}\mathbb{E}_{q(\theta)}[\theta_l]\sqrt{\lambda_l}\Psi_l(s_j) \\
&+\mathbb{E}_{q(\theta)}\left[\left(\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\Psi_l(s_j)\right)^2\right]\Bigg\} - \left(\frac{1}{2}+1\right)\ln\{\sigma_\alpha^2\} - \frac{\mathbb{E}_{q(a_\alpha)}\left[\frac{1}{a_\alpha}\right]}{\sigma_\alpha^2}
\end{aligned}
$$

$$
\begin{aligned}
q^*(\sigma_\alpha^2) &= \text{Inverse-Gamma}(\sigma_\alpha^2; b_\alpha, c_\alpha) \\
b_\alpha &= \frac{M}{2} + \frac{1}{2} \\
c_\alpha &= \Bigg[\mathbb{E}_{q(a_\alpha)}\left[\frac{1}{a_\alpha}\right] + \frac{1}{2}\sum_{j=1}^{M}\Bigg\{\mathbb{E}_{q(\alpha)}[\alpha(s_j)^2] - 2\mathbb{E}_{q(\alpha)}[\alpha(s_j)]\sum_{l=1}^{L}\mathbb{E}_{q(\theta)}[\theta_l]\sqrt{\lambda_l}\Psi_l(s_j) \\
&+\mathbb{E}_{q(\theta)}\left[\left(\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\Psi_l(s_j)\right)^2\right]\Bigg\}\Bigg]
\end{aligned}
$$

C.2.9 Update a_ϵ

$$
\begin{aligned}
\ln\{q^*(a_\epsilon)\} &\propto \mathbb{E}_{q(\beta_0,\boldsymbol{\theta},\boldsymbol{\alpha}(\boldsymbol{s}),\delta,\sigma_\epsilon^2,\sigma_\beta^2,\sigma_\alpha^2,a_\beta,a_\alpha)}[\ln\{p(\boldsymbol{y},\beta_0,\boldsymbol{\theta},\boldsymbol{\alpha}(\boldsymbol{s}),\delta,\sigma_\epsilon^2,\sigma_\beta^2,\sigma_\alpha^2,a_\epsilon,a_\beta,a_\alpha)\}] \\
&\propto \mathbb{E}_{q(\sigma_\epsilon^2)}\left[\ln\left\{p(\sigma_\epsilon^2|a_\epsilon)p(a_\epsilon)\right\}\right] \\
&\propto \frac{1}{2}\ln\left\{\frac{1}{a_\epsilon}\right\} - \frac{1}{a_\epsilon}\mathbb{E}_{q(\sigma_\epsilon^2)}\left[\frac{1}{\sigma_\epsilon^2}\right] + \left(\frac{1}{2}+1\right)\ln\left\{\frac{1}{a_\epsilon}\right\} - \frac{\frac{1}{s_\epsilon^2}}{a_\epsilon}
\end{aligned}
$$

$$\begin{aligned}
q^*(a_\epsilon) &= \text{Inverse-Gamma}(a_\epsilon; d_\epsilon, e_\epsilon)\\
d_\epsilon &= 1\\
e_\epsilon &= \mathbb{E}_{q(\sigma^2_\epsilon)}\left[\frac{1}{\sigma^2_\epsilon}\right] + \frac{1}{s^2_\epsilon}
\end{aligned}$$

C.2.10 Update a_β

$$\begin{aligned}
\ln\{q^*(a_\beta)\} &\propto \mathbb{E}_{q(\beta_0,\boldsymbol{\theta},\boldsymbol{\alpha(s)},\delta,\sigma^2_\epsilon,\sigma^2_\beta,\sigma^2_\alpha,a_\epsilon,a_\alpha)}[\ln\{p(\boldsymbol{y},\beta_0,\boldsymbol{\theta},\boldsymbol{\alpha(s)},\delta,\sigma^2_\epsilon,\sigma^2_\beta,\sigma^2_\alpha,a_\epsilon,a_\beta,a_\alpha)\}]\\
&\propto \mathbb{E}_{q(\sigma^2_\beta)}\left[\ln\left\{p(\sigma^2_\beta|a_\beta)p(a_\beta)\right\}\right]\\
&\propto \frac{1}{2}\ln\left\{\frac{1}{a_\beta}\right\} - \frac{1}{a_\beta}\mathbb{E}_{q(\sigma^2_\beta)}\left[\frac{1}{\sigma^2_\beta}\right] + \left(\frac{1}{2}+1\right)\ln\left\{\frac{1}{a_\beta}\right\} - \frac{\frac{1}{s^2_\beta}}{a_\beta}
\end{aligned}$$

$$\begin{aligned}
q^*(a_\beta) &= \text{Inverse-Gamma}(a_\beta; d_\beta, e_\beta)\\
d_\beta &= 1\\
e_\beta &= \mathbb{E}_{q(\sigma^2_\beta)}\left[\frac{1}{\sigma^2_\beta}\right] + \frac{1}{s^2_\beta}
\end{aligned}$$

C.2.11 Update a_α

$$\begin{aligned}
\ln\{q^*(a_\alpha)\} &\propto \mathbb{E}_{q(\beta_0,\boldsymbol{\theta},\boldsymbol{\alpha(s)},\delta,\sigma^2_\epsilon,\sigma^2_\beta,\sigma^2_\alpha,a_\epsilon,a_\beta)}[\ln\{p(\boldsymbol{y},\beta_0,\boldsymbol{\theta},\boldsymbol{\alpha(s)},\delta,\sigma^2_\epsilon,\sigma^2_\beta,\sigma^2_\alpha,a_\epsilon,a_\beta,a_\alpha)\}]\\
&\propto \mathbb{E}_{q(\sigma^2_\alpha)}\left[\ln\left\{p(\sigma^2_\alpha|a_\alpha)p(a_\alpha)\right\}\right]\\
&\propto \frac{1}{2}\ln\left\{\frac{1}{a_\alpha}\right\} - \frac{1}{a_\alpha}\mathbb{E}_{q(\sigma^2_\alpha)}\left[\frac{1}{\sigma^2_\alpha}\right] + \left(\frac{1}{2}+1\right)\ln\left\{\frac{1}{a_\alpha}\right\} - \frac{\frac{1}{s^2_\alpha}}{a_\alpha}
\end{aligned}$$

$$\begin{aligned}
q^*(a_\alpha) &= \text{Inverse-Gamma}(a_\alpha; d_\alpha, e_\alpha)\\
d_\alpha &= 1\\
e_\alpha &= \mathbb{E}_{q(\sigma^2_\alpha)}\left[\frac{1}{\sigma^2_\alpha}\right] + \frac{1}{s^2_\alpha}
\end{aligned}$$

C.2.12 ELBO

$$
\begin{aligned}
\mathcal{L}(q) \geq\ & \mathbb{E}_{q(\theta,\alpha,\delta,\sigma_\epsilon^2,\sigma_\beta^2,\sigma_\alpha^2,a_\epsilon,a_\beta,a_\alpha} \left[\ln\left\{p\left(\boldsymbol{y},\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{\alpha}(\boldsymbol{s}),\sigma_\epsilon^2,\sigma_\beta^2,\sigma_\alpha^2,a_\epsilon,a_\beta,a_\alpha\right)\right\}\right] \\
& - \mathbb{E}_{q(\theta,\alpha,\delta,\sigma_\epsilon^2,\sigma_\beta^2,\sigma_\alpha^2,a_\epsilon,a_\beta,a_\alpha} \left[\left\{q\left(\boldsymbol{\theta},\boldsymbol{\alpha}(\boldsymbol{s}),\sigma_\epsilon^2,\sigma_\beta^2,\sigma_\alpha^2,a_\epsilon,a_\beta,a_\alpha\right)\right\}\right] \\
& = \sum_{i=1}^{N}\left\{-\frac{1}{2}\ln\{2\pi\} - \frac{1}{2}\mathbb{E}_{q(\sigma_\epsilon^2)}\left[\ln\{\sigma_\epsilon^2\}\right]\right. \\
& -\frac{1}{2}\mathbb{E}_{q(\sigma_\epsilon^2)}\left[\frac{1}{\sigma_\epsilon^2}\right]\left[y_i^2 + \mathbb{E}_{q(\beta_0)}\left[\beta_0^2\right] + \mathbb{E}_{q(\theta,\alpha,\delta)}\left[\left(\sum_{j=1}^{M}\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\Psi_l(s_j)I(|\alpha(s_j)|>\delta)x_{i,j}\right)^2\right]\right] \\
& \left.-2\left[y_i - \mathbb{E}_{q(\beta_0)}[\beta_0]\right]\mathbb{E}_{q(\theta,\alpha,\delta)}\left[\sum_{j=1}^{M}\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\Psi_l(s_j)I(|\alpha(s_j)|>\delta)x_{i,j}\right] - 2y_i\mathbb{E}_{q(\beta_0)}[\beta_0]\right\} \\
& -\frac{1}{2}\ln\{2\pi\} - \frac{1}{2}\ln\{\sigma_{\beta_0}^2\} - \frac{1}{2\sigma_{\beta_0}^2}\mathbb{E}_{q(\beta_0)}[\beta_0^2] \\
& + \sum_{l=1}^{L}\left\{-\frac{1}{2}\ln\{2\pi\} - \frac{1}{2}\mathbb{E}_{q(\sigma_\beta^2)}\left[\ln\{\sigma_\beta^2\}\right] - \frac{1}{2}\mathbb{E}_{q(\sigma_\beta^2)}\left[\frac{1}{\sigma_\beta^2}\right]\mathbb{E}_{q(\theta)}\left[\theta_l^2\right]\right\} \\
& + \sum_{j=1}^{M}\left\{-\frac{1}{2}\ln\{2\pi\} - \frac{1}{2}\mathbb{E}_{q(\sigma_\alpha^2)}\left[\ln\{\sigma_\alpha^2\}\right]\right. \\
& -\frac{1}{2}\mathbb{E}_{q(\sigma_\alpha^2)}\left[\frac{1}{\sigma_\alpha^2}\right]\left[\mathbb{E}_{q(\alpha)}[\alpha(s_j)^2] - 2\mathbb{E}_{q(\alpha)}[\alpha(s_j)]\sum_{l=1}^{L}\mathbb{E}_{q(\theta)}[\theta_l]\sqrt{\lambda_l}\Psi_l(s_j)\right. \\
& \left.\left.+\mathbb{E}_{q(\theta)}\left[\left(\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\Psi_l(s_j)\right)^2\right]\right]\right\} \\
& + \frac{1}{2}\mathbb{E}_{q(a_\epsilon)}\left[\ln\left\{\frac{1}{a_\epsilon}\right\}\right] + \left(\frac{1}{2}+1\right)\mathbb{E}_{q(\sigma_\epsilon^2)}[\ln\{\sigma_\epsilon^2\}] - \frac{\mathbb{E}_{q(a_\epsilon)}\left[\frac{1}{a_\epsilon}\right]}{\mathbb{E}_{q(\sigma_\epsilon^2)}[\sigma_\epsilon^2]} \\
& + \frac{1}{2}\mathbb{E}_{q(a_\beta)}\left[\ln\left\{\frac{1}{a_\beta}\right\}\right] + \left(\frac{1}{2}+1\right)\mathbb{E}_{q(\sigma_\beta^2)}[\ln\{\sigma_\beta^2\}] - \frac{\mathbb{E}_{q(a_\beta)}\left[\frac{1}{a_\beta}\right]}{\mathbb{E}_{q(\sigma_\beta^2)}[\sigma_\beta^2]} \\
& + \frac{1}{2}\mathbb{E}_{q(a_\alpha)}\left[\ln\left\{\frac{1}{a_\alpha}\right\}\right] + \left(\frac{1}{2}+1\right)\mathbb{E}_{q(\sigma_\alpha^2)}[\ln\{\sigma_\alpha^2\}] - \frac{\mathbb{E}_{q(a_\alpha)}\left[\frac{1}{a_\alpha}\right]}{\mathbb{E}_{q(\sigma_\alpha^2)}[\sigma_\alpha^2]} \\
& - \left(\frac{1}{2}+1\right)\mathbb{E}_{q(a_\epsilon)}[\ln\{a_\epsilon\}] - \frac{\frac{1}{s_\epsilon^2}}{\mathbb{E}_{q(a_\epsilon)}[a_\epsilon]} \\
& - \left(\frac{1}{2}+1\right)\mathbb{E}_{q(a_\beta)}[\ln\{a_\beta\}] - \frac{\frac{1}{s_\beta^2}}{\mathbb{E}_{q(a_\beta)}[a_\beta]} \\
& - \left(\frac{1}{2}+1\right)\mathbb{E}_{q(a_\alpha)}[\ln\{a_\alpha\}] - \frac{\frac{1}{s_\alpha^2}}{\mathbb{E}_{q(a_\alpha)}[a_\alpha]} \\
& + \frac{1}{2}\ln\{\Sigma_{\beta_0}\} + \frac{1}{2\Sigma_{\beta_0}}\left\{\mathbb{E}_{q(\beta_0)}[\beta_0^2] - 2\mathbb{E}_{q(\beta_0)}[\beta_0]\mu_{\beta_0} + \mu_{\beta_0}^2\right\}
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{L}{2}\ln\{2\pi\}+\frac{L}{2}\ln\{\det\{\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\}\}+\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbb{E}_{q(\theta)}[\boldsymbol{\theta}\boldsymbol{\theta}^T]-2\mathbb{E}_{q(\theta)}[\boldsymbol{\theta}]\boldsymbol{\mu}_\theta^T+\boldsymbol{\mu}_\theta\boldsymbol{\mu}_\theta^T\right)\\
&+\sum_{j=1}^{M}\left\{\frac{1}{2}\mathbb{E}_{q(\sigma_\alpha^2)}[\ln\{\sigma_\alpha^2\}]+w_{-1,j}\mathbb{E}_{q(\theta,\sigma_\alpha^2)}\left[\Phi\left(\frac{-\delta-\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\Psi_l(s_j)}{\sigma_\alpha}\right)\right]\right.\\
&+w_{0,j}\mathbb{E}_{q(\theta,\sigma_\alpha)}\left[\ln\left\{\Phi\left(\frac{\delta-\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\Psi_l(s_j)}{\sigma_\alpha}\right)-\Phi\left(\frac{-\delta-\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\Psi_l(s_j)}{\sigma_\alpha}\right)\right\}\right]\\
&+w_{1,j}\mathbb{E}_{q(\theta,\sigma_\alpha^2)}\left[1-\Phi\left(\frac{\delta-\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\Psi_l(s_j)}{\sigma_\alpha}\right)\right]\\
&\left.+\sum_{k\in\{-1,0,1\}}w_{k,j}\left[\mathbb{E}_{q(\alpha)}[\alpha(s_j)^2]-2\mathbb{E}_{q(\alpha)}[\alpha(s_j)]\mu_{\alpha(s_j)}+\mu_{\alpha(s_j)}^2\right]\right\}\\
&-b_\epsilon\ln\{c_\epsilon\}+\ln\{\Gamma(b_\epsilon)\}-(b_\epsilon+1)\mathbb{E}_{q(\sigma_\epsilon^2)}\left[\ln\left\{\frac{1}{\sigma_\epsilon^2}\right\}\right]+\frac{c_\epsilon}{\mathbb{E}_{q(\sigma_\epsilon^2)}\left[\sigma_\epsilon^2\right]}\\
&-b_\beta\ln\{c_\beta\}+\ln\{\Gamma(b_\beta)\}-(b_\beta+1)\mathbb{E}_{q(\sigma_\beta^2)}\left[\ln\left\{\frac{1}{\sigma_\beta^2}\right\}\right]+\frac{c_\beta}{\mathbb{E}_{q(\sigma_\beta^2)}\left[\sigma_\beta^2\right]}\\
&-b_\alpha\ln\{c_\alpha\}+\ln\{\Gamma(b_\alpha)\}-(b_\alpha+1)\mathbb{E}_{q(\sigma_\epsilon^2)}\left[\ln\left\{\frac{1}{\sigma_\alpha^2}\right\}\right]+\frac{c_\alpha}{\mathbb{E}_{q(\sigma_\alpha^2)}\left[\sigma_\alpha^2\right]}\\
&-d_\epsilon\ln\{e_\epsilon\}+\ln\{\Gamma(d_\epsilon)\}-(d_\epsilon+1)\mathbb{E}_{q(a_\epsilon)}\left[\ln\left\{\frac{1}{a_\epsilon}\right\}\right]+\frac{e_\epsilon}{\mathbb{E}_{q(a_\epsilon)}[a_\epsilon]}\\
&-d_\beta\ln\{e_\beta\}+\ln\{\Gamma(d_\beta)\}-(d_\beta+1)\mathbb{E}_{q(a_\beta)}\left[\ln\left\{\frac{1}{a_\beta}\right\}\right]+\frac{e_\beta}{\mathbb{E}_{q(a_\beta)}[a_\beta]}\\
&-d_\alpha\ln\{e_\alpha\}+\ln\{\Gamma(d_\alpha)\}-(d_\alpha+1)\mathbb{E}_{q(a_\alpha)}\left[\ln\left\{\frac{1}{a_\alpha}\right\}\right]+\frac{e_\alpha}{\mathbb{E}_{q(a_\alpha)}[a_\alpha]}
\end{aligned}
$$

C.2.13 Expectations

$$\mathbb{E}_{q(\beta_0)}[\beta_0^2]=Var(\beta_0)+\mathbb{E}_{q(\beta_0)}[\beta_0]^2=\Sigma_{\beta_0}+\mu_{\beta_0}^2$$

$$\mathbb{E}_{q(\theta)}[\theta_l^2]=Var(\theta_l)+\mathbb{E}_{q(\theta)}[\theta_l]^2=\Sigma_{\theta_l}+\mu_{\theta_l}^2$$

$$\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)^2]=[\boldsymbol{\lambda}\cdot\boldsymbol{\psi}(\boldsymbol{s_j})]^T\left[\boldsymbol{\Sigma}_\theta+\boldsymbol{\mu}_\theta\boldsymbol{\mu}_\theta^T\right][\boldsymbol{\lambda}\cdot\boldsymbol{\psi}(\boldsymbol{s_j})]$$

$$\mathbb{E}_{q(\alpha)}(I|\alpha(s_j)|>\delta)=\mathbb{P}_{q(\alpha)}(\alpha(s_j)>\delta)+\mathbb{P}_{q(\alpha)}(\alpha(s_j)<-\delta)=w_{1,j}+w_{-1,j}$$

Note that $\tilde{\beta}(s_j)=\sum_{l=1}^{L}\theta_l\sqrt{\lambda_l}\Phi_l(s_j)$. Then

$$\mathbb{E}_{q(\theta)}\{\tilde{\beta}(s_j)\}=\sum_{l=1}^{L}\mathbb{E}_{q(\theta)}(\theta_l)\sqrt{\lambda_l}\Phi_l(s_j)$$

$$\begin{aligned}\mathbb{E}_{q(\theta)}\{\beta^2(s_j)\} &= \sum_{l=1}^{L}\mathbb{E}_{q(\theta)}(\theta_l^2)\lambda_l\Phi_l^2(s_j) + 2\sum_{l<l'}\mathbb{E}_{q(\theta)}(\theta_l)\mathbb{E}_{q(\theta)}(\theta_{l'})\sqrt{\lambda_l}\Phi_{l'}(s_j)\sqrt{\lambda_l}\Phi_{l'}(s_j)\\ &= \sum_{l=1}^{L}\mathrm{Var}_{q(\theta)}(\theta_l)\lambda_l\Phi_l^2(s_j) + \left(\sum_{l=1}^{L}\mathbb{E}_{q(\theta)}(\theta_l)\sqrt{\lambda_l}\Phi_{l'}(s_j)\right)^2\end{aligned}$$

Now

$$\mathbb{E}_{q(\alpha,\theta,\delta)}\left(\sum_{j=1}^{M}\tilde{\beta}(s_j)I(|\alpha(s_j)|>\delta)x_{i,j}\right) = \sum_{j=1}^{M}\mathbb{E}_{q(\theta)}\{\tilde{\beta}(s_j)\}I(|\alpha(s_j)|>\delta)x_{i,j}$$

and

$$\begin{aligned}&\mathbb{E}_{q(\alpha,\theta,\delta)}\left[\left(\sum_{j=1}^{M}\tilde{\beta}(s_j)I(|\alpha(s_j)|>\delta)x_{i,j}\right)^2\right]\\ &= \sum_{j=1}^{M}\mathrm{Var}_{q(\alpha,\theta,\delta)}\{\tilde{\beta}(s_j)I(|\alpha(s_j)>\delta|)\}x_{i,j}^2 + \left[\sum_{j=1}^{M}\mathbb{E}_{q(\alpha,\theta,\delta)}\{\tilde{\beta}(s_j)I(|\alpha(s_j)>\delta|)\}x_{i,j}\right]^2\end{aligned}$$

where

$$\begin{aligned}&\mathrm{Var}_{q(\alpha,\theta,\delta)}\{\tilde{\beta}(s_j)I(|\alpha(s_j)>\delta|)\}\\ &= \mathbb{E}_{q(\alpha,\theta,\delta)}\{\beta^2(s_j)I(|\alpha(s_j)>\delta|)\} - \left[\mathbb{E}_{q(\alpha,\theta,\delta)}\{\tilde{\beta}(s_j)I(|\alpha(s_j)>\delta|)\}\right]^2\\ &= \mathbb{E}_{q(\theta)}\{\beta^2(s_j)\}\mathbb{E}_{q(\alpha,\delta)}\{I(|\alpha(s_j)>\delta|)\} - \mathbb{E}_{q(\theta)}\{\tilde{\beta}(s_j)\}^2\mathbb{E}_{q(\alpha,\delta)}\{I(|\alpha(s_j)>\delta|)\}^2\end{aligned}$$

$$\begin{aligned}&\mathbb{E}_{q(\theta)}\left[\left[\boldsymbol{\Psi}^T\mathrm{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^{L}\boldsymbol{\theta}\right]^T\left[\boldsymbol{\Psi}^T\mathrm{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^{L}\boldsymbol{\theta}\right]\right]\\ &= \mathbb{E}_{q(\theta)}\left[\mathrm{tr}\left(\left[\boldsymbol{\Psi}^T\mathrm{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^{L}\right]^T\left[\boldsymbol{\Psi}^T\mathrm{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^{L}\right]\boldsymbol{\theta}\boldsymbol{\theta}^T\right)\right]\\ &= \mathrm{tr}\left(\left[\boldsymbol{\Psi}^T\mathrm{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^{L}\right]^T\left[\boldsymbol{\Psi}^T\mathrm{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^{L}\right]\mathbb{E}_{q(\theta)}\left[\boldsymbol{\theta}\boldsymbol{\theta}^T\right]\right)\\ &= \mathrm{tr}\left(\left[\boldsymbol{\Psi}^T\mathrm{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^{L}\right]^T\left[\boldsymbol{\Psi}^T\mathrm{diag}\left\{\sqrt{\lambda_l}\right\}_{l=1}^{L}\right]\left(\boldsymbol{\Sigma}_\theta + \boldsymbol{\mu}_\theta\boldsymbol{\mu}_\theta^T\right)\right)\end{aligned}$$

$$\mathbb{E}_{q(\sigma_\epsilon^2)}\left[\frac{1}{\sigma_\epsilon^2}\right] = \frac{1}{\mathbb{E}_{q(\sigma_\epsilon^2)}[\sigma_\epsilon^2]} = \frac{b_\epsilon}{c_\epsilon}$$

$$\mathbb{E}_{q(\sigma_\beta^2)}\left[\frac{1}{\sigma_\beta^2}\right] = \frac{1}{\mathbb{E}_{q(\sigma_\beta^2)}[\sigma_\beta^2]} = \frac{b_\beta}{c_\beta}$$

$$\mathbb{E}_{q(\sigma^2_\alpha)}\left[\frac{1}{\sigma^2_\alpha}\right] = \frac{1}{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]} = \frac{b_\alpha}{c_\alpha}$$

$$\mathbb{E}_{q(a_\epsilon)}\left[\frac{1}{a_\epsilon}\right] = \frac{1}{\mathbb{E}_{q(a_\epsilon)}[a_\epsilon]} = \frac{d_\epsilon}{e_\epsilon}$$

$$\mathbb{E}_{q(a_\beta)}\left[\frac{1}{a_\beta}\right] = \frac{1}{\mathbb{E}_{q(a_\beta)}[a_\beta]} = \frac{d_\beta}{e_\beta}$$

$$\mathbb{E}_{q(a_\alpha)}\left[\frac{1}{a_\alpha}\right] = \frac{1}{\mathbb{E}_{q(a_\alpha)}[a_\alpha]} = \frac{d_\alpha}{e_\alpha}$$

$$\mathbb{E}_{q(\sigma^2_\epsilon)}\left[\ln\{\sigma^2_\epsilon\}\right] = \ln\{c_\epsilon\} - \psi(b_\epsilon)$$

$$\mathbb{E}_{q(\sigma^2_\beta)}\left[\ln\{\sigma^2_\beta\}\right] = \ln\{c_\beta\} - \psi(b_\beta)$$

$$\mathbb{E}_{q(\sigma^2_\alpha)}\left[\ln\{\sigma^2_\alpha\}\right] = \ln\{c_\alpha\} - \psi(b_\alpha)$$

where $\psi(\cdot)$ is the digamma function.

$$\begin{aligned}
\mathbb{E}_{q(\alpha)}[\alpha(s_j)^2] &= Var(\alpha(s_j)) + \mathbb{E}_{q(\alpha)}[\alpha(s_j)]^2 \\
&= w_{-1,j}\left(Var(\alpha(s_j)|\alpha(s_j) < -\delta) + \mathbb{E}_{q(\alpha)}[\alpha(s_j)|\alpha(s_j) < -\delta]^2\right) \\
&+ w_{0,j}\left(Var(\alpha(s_j)||\alpha(s_j)| < \delta) + \mathbb{E}_{q(\alpha)}[\alpha(s_j)||\alpha(s_j)| < \delta]^2\right) \\
&+ w_{1,j}\left(Var(\alpha(s_j)|\alpha(s_j) > \delta) + \mathbb{E}_{q(\alpha)}[\alpha(s_j)|\alpha(s_j) > \delta]^2\right) + \mathbb{E}_{q(\alpha(s_j))}[\alpha(s_j)]^2 \\
\mathbb{E}_{q(\alpha(s_j))}[\alpha(s_j)] &= w_{-1,j}\mathbb{E}_{q(\alpha(s_j))}[\alpha(s_j)|\alpha(s_j) < -\delta] + w_{0,j}\mathbb{E}_{q(\alpha(s_j))}[\alpha(s_j)||\alpha(s_j)| < \delta] \\
&+ w_{1,j}\mathbb{E}_{q(\alpha(s_j))}[\alpha(s_j)|\alpha(s_j) > \delta]
\end{aligned}$$

$$\phi(\xi) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\xi^2\right\}$$

$$\mathbb{E}_{q(\alpha)}[\alpha(s_j)|\alpha(s_j) < -\delta] = \mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)] - \sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]} \frac{\phi\left(\frac{-\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)}{\Phi\left(\frac{-\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)}$$

$$\mathbb{E}_{q(\alpha)}[\alpha(s_j)||\alpha(s_j)| < \delta] = \mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)] - \sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]} \frac{\phi\left(\frac{\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right) - \phi\left(\frac{-\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)}{\Phi\left(\frac{\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right) - \phi\left(\frac{-\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)}$$

$$\mathbb{E}_{q(\alpha)}[\alpha(s_j)|\alpha(s_j) > \delta] = \mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)] + \sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]} \frac{\phi\left(\frac{\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)}{1-\Phi\left(\frac{\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)}$$

$$\begin{aligned} Var(\alpha(s_j)|\alpha(s_j) < -\delta) = \mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha] \Bigg[1 - \left(\frac{-\delta - \mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right) \frac{\phi\left(\frac{-\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)}{\Phi\left(\frac{-\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)} \\ - \left(\frac{\phi\left(\frac{-\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)}{\Phi\left(\frac{-\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)}\right)^2 \Bigg] \end{aligned}$$

$$Var(\alpha(s_j)||\alpha(s_j)| < \delta) = \mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]\,[1- \frac{\left(\frac{\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)\phi\left(\frac{\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)-\left(\frac{-\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)\phi\left(\frac{-\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)}{\Phi\left(\frac{\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)-\Phi\left(\frac{-\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)} - \left.\left(\frac{\phi\left(\frac{\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)-\phi\left(\frac{-\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)}{\Phi\left(\frac{\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)-\Phi\left(\frac{-\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)}\right)^2\right]$$

$$Var(\alpha(s_j)|\alpha(s_j) > \delta) = \mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]\left[1+\left(\frac{\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)\frac{\phi\left(\frac{\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)}{1-\Phi\left(\frac{\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)} - \left(\frac{\phi\left(\frac{\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)}{1-\Phi\left(\frac{\delta-\mathbb{E}_{q(\theta)}[\tilde{\beta}(s_j)]}{\sqrt{\mathbb{E}_{q(\sigma^2_\alpha)}[\sigma^2_\alpha]}}\right)}\right)^2\right]$$

C.2.14 Computational Complexity of RTGP-VI

Quantity	Dimensions	Space Complexity
$q(\boldsymbol{\theta})$	$L \times 1$	$\mathcal{O}(M \times L^2 + N \times M)$
$q(\beta_0)$	1×1	$\mathcal{O}(N \times M + L \times M)$
$q(\boldsymbol{\alpha}(\boldsymbol{s}))$	$M \times 1$	$\mathcal{O}(N \times M + L \times M)$
$q(\delta)$	1×1	$\mathcal{O}(N \times M + L \times M)$
$q(\sigma_\epsilon^2)$	1×1	$\mathcal{O}(N \times M + L \times M)$
$q(\sigma_\alpha^2)$	1×1	$\mathcal{O}(L \times M)$
$q(\sigma_\beta^2)$	1×1	$\mathcal{O}(L)$
$q(a_\epsilon)$	1×1	$\mathcal{O}(1)$
$q(a_\alpha)$	1×1	$\mathcal{O}(1)$
$q(a_\beta)$	1×1	$\mathcal{O}(1)$
ELBO	1×1	$\mathcal{O}(N \times M)$

Table C.1: Computational complexity of RTGP-VI algorithm for each variational update and the calculation of the ELBO.

C.3 Simulation Study: Volumetric Data Representation

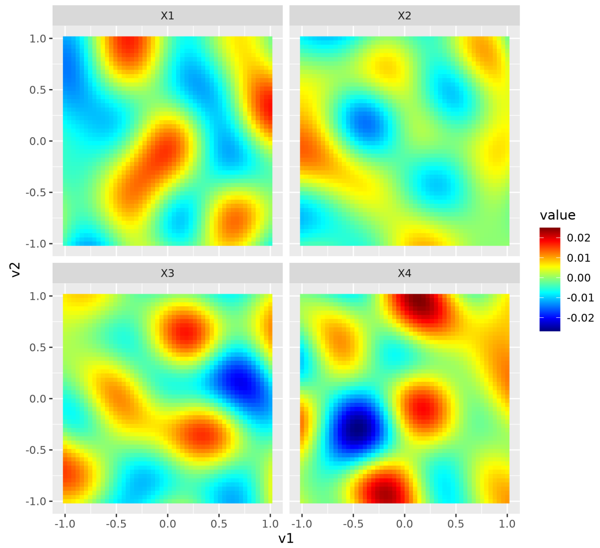
The relaxed-thresholded Gaussian process prior is a general class of priors that can be applied to both cortical surface-based and volume-based data; however, the kernel function as well as the distance function needs to be adjusted accordingly to respect the change in data representation. For the covariance function of the GP (in the volumetric data case), we choose a modified squared exponential kernel function, which is defined as:

$$\kappa(s, s') = \exp\left\{-a\left(||s||^2 + ||s'||\right)^2 - b||s - s'||^2\right\}, \tag{C.3}$$

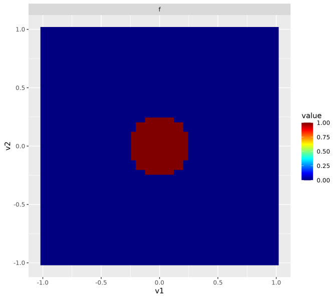
for $a > 0$ and $b > 0$, where the norm $||s||^2 = s^T s$. The hyperparameters of the kernel determine the properties of the spatial process where a controls the decay rate of $\text{Var}\{\boldsymbol{\beta}(\boldsymbol{s})\}$ compared to the maximum marginal variance $\text{Var}\{\boldsymbol{\beta}(\mathbf{0})\} = \sigma_\beta^2$ and b controls the smoothness of $\boldsymbol{\beta}(\boldsymbol{s})$. The distance between spatial locations in the correlation function is measured by Euclidean distance. Using the modified squared exponential kernel is computationally fast as the R package "`BayesGPfit`" produces a fast eigendecomposition of the specified kernel through Hermite polynomials. The rest of the model specifications for a scalar-on-image regression with a RTGP prior on the spatially varying coefficients remains the same as defined in Section 4.2.1.

We simulate data by defining an effect of interest, a circle with pre-defined radius in the middle of a 2D image with dimensions 50 × 50 rendering 5% of the voxels out of a total of $M = 2,500$ as active. Figure C.1 provides an example of the simulation setup where a) displays 4 examples of an input image generated with a GP, b) shows the true parameter map with a circular effect, and c) indicates where in the image an effect is significant. The scalar output y is generated by plugging in the above quantities in $y_i = \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}(\boldsymbol{s})\epsilon_i$ for all subjects $i = 1, \ldots, N$ with intercept $\beta_0 = 2$ and noise drawn from a random Normal distribution $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ where $\sigma = \{0.2, 1\}$ depending on the signal-to-noise ratio.

a) Input image $\boldsymbol{x}_i$



b) True beta map $\boldsymbol{\beta}(\boldsymbol{s})$



c) True significance map

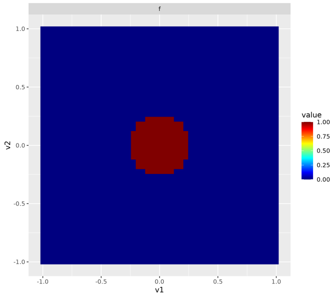


Figure C.1: Overview of simulation setup for 2D volumetric data with a) an example of an input image drawn from a GP, b) the true image coefficient map $\boldsymbol{\beta}(\boldsymbol{s})$, and c) the true binary significance map showing the areas which have an effect in our simulation study.

We compare the parameter, predictive and inference results for simulation settings with varying sample sizes $N = \{500; 1,000; 2,000\}$ and varying signal-to-noise ratios $SNR = \{1/0.2; 3/1\}$ for our method RTGP-Gibbs (RTGP estimated via a Gibbs sampler) and RTGP-VI (RTGP estimated via variational inference) to the baseline methods which are a Gaussian process regression with a Normal prior on the basis coefficients (GPR + Normal) and a scalar-on-image regression with a soft-thresholded Gaussian process on the image coefficients (STGP). The model with a STGP prior on the coefficients we estimate with the R package "`STGP`"[1] provided by the authors of the work (Kang et al., 2018).

---

[1] `https://public.websites.umich.edu/~jiankang/software/STGP.html`

Parameter Estimates	**Bias**		**MSE**	
N=500	SNR = 1/0.2	SNR = 3/1	SNR = 1/0.2	SNR = 3/1
GPR + Normal	0.0010	0.0061	0.0144	0.1404
STGP	0.0013	0.0040	0.0067	0.0625
RTGP-Gibbs	-0.0031	-0.0160	0.0047	0.0524
RTGP-VI	0.0004	0.0064	0.0052	0.1078
N=1,000	SNR = 1/0.2	SNR = 3/1	SNR = 1/0.2	SNR = 3/1
GPR + Normal	0.0007	0.0040	0.0139	0.1337
STGP	0.0012	0.0038	0.0066	0.0614
RTGP-Gibbs	-0.0017	-0.0106	0.0048	0.0493
RTGP-VI	0.0004	0.0063	0.0052	0.1002
N=2,000	SNR = 1/0.2	SNR = 3/1	SNR = 1/0.2	SNR = 3/1
GPR + Normal	-0.0003	0.0008	0.0136	0.1272
STGP	0.0008	0.0031	0.0065	0.0600
RTGP-Gibbs	-0.0020	-0.0054	0.0040	0.0458
RTGP-VI	0.0014	0.0048	0.0062	0.0903

Table C.2: Evaluation of parameter estimate results with bias and MSE for a simulation study setting with varying sample sizes $N = \{500; 1,000; 2,000\}$ and signal-to-noise rations $SNR = \{1/0.2; 3/1\}$ for the methods GPR+Normal, STGP, RTGP-Gibbs, and RTGP-VI.

Prediction	**R^2 (train)**		**MSE (train)**		**R^2 (test)**		**MSE (test)**	
N=500	SNR = 1/0.2	SNR = 3/1	SNR = 1/0.2	SNR = 3/1	SNR = 1/0.2	SNR = 3/1	SNR = 1/0.2	SNR = 3/1
GPR + Normal	0.9377	0.8455	0.0378	0.9205	0.9316	0.8290	0.0424	1.0413
STGP	0.9343	0.8396	0.0398	0.9553	0.9356	0.8365	0.0399	0.9947
RTGP-Gibbs	0.9345	0.8397	0.0399	0.9682	0.9354	0.8364	0.0400	1.0149
RTGP-VI	0.9361	0.8311	0.0400	1.0292	0.9340	0.8298	0.0404	1.0645
N=1,000	SNR = 1/0.2	SNR = 3/1	SNR = 1/0.2	SNR = 3/1	SNR = 1/0.2	SNR = 3/1	SNR = 1/0.2	SNR = 3/1
GPR + Normal	0.9374	0.8421	0.0390	0.9708	0.9317	0.8302	0.0417	1.0367
STGP	0.9362	0.8394	0.0398	0.9875	0.9335	0.8352	0.0405	1.0060
RTGP-Gibbs	0.9364	0.8392	0.0397	0.9930	0.9340	0.8352	0.0403	1.0104
RTGP-VI	0.9361	0.8334	0.0400	1.0451	0.9340	0.8295	0.0404	1.0581
N=2,000	SNR = 1/0.2	SNR = 3/1	SNR = 1/0.2	SNR = 3/1	SNR = 1/0.2	SNR = 3/1	SNR = 1/0.2	SNR = 3/1
GPR + Normal	0.9347	0.8350	0.0396	0.9965	0.9326	0.8266	0.0405	1.0214
STGP	0.9340	0.8339	0.0400	1.0036	0.9337	0.8294	0.0398	1.0048
RTGP-Gibbs	0.9335	0.8337	0.0404	1.0055	0.9331	0.8292	0.0401	1.0067
RTGP-VI	0.9336	0.8308	0.0403	1.0332	0.9336	0.8266	0.0398	1.0330

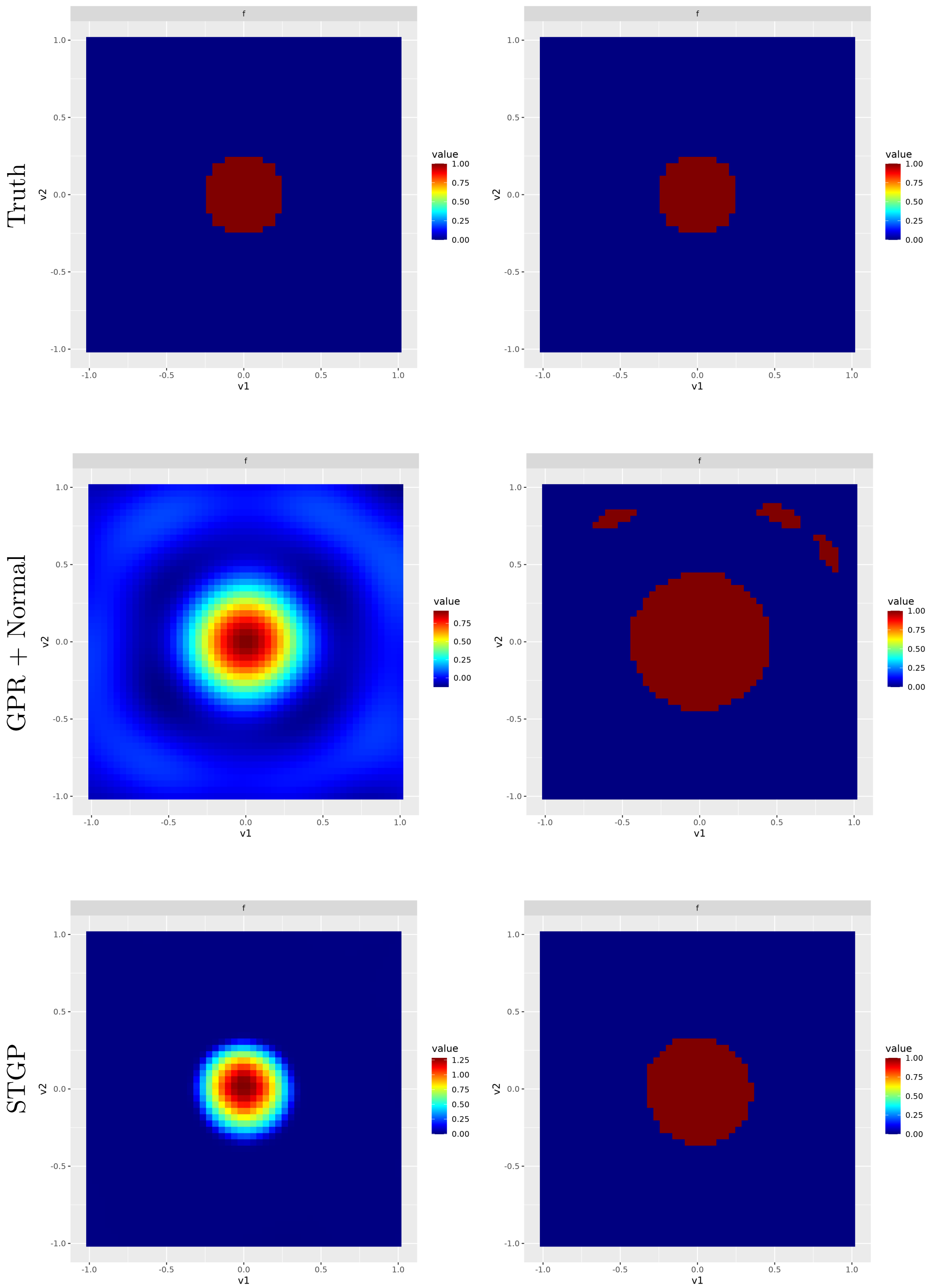
Table C.3: Evaluation of predictive results with R^2 and MSE for a simulation study setting with varying sample sizes $N = \{500; 1,000; 2,000\}$ and signal-to-noise rations $SNR = \{1/0.2; 3/1\}$ for the methods GPR+Normal, STGP, RTGP-Gibbs, and RTGP-VI.

Inference	**TPR**		**TDR**		**FPR**		**FDR**	
N=500	SNR = 1/0.2	SNR = 3/1	SNR = 1/0.2	SNR = 3/1	SNR = 1/0.2	SNR = 3/1	SNR = 1/0.2	SNR = 3/1
GPR + Normal	1.0000	1.0000	0.2915	0.3258	0.1287	0.1095	0.7085	0.6742
STGP	1.0000	1.0000	0.5607	0.5404	0.0410	0.0447	0.4393	0.4596
RTGP-Gibbs	0.9863	0.9935	0.9123	0.8239	0.0053	0.0120	0.0877	0.1761
RTGP-VI	0.9992	1.0000	0.8436	0.5015	0.0100	0.0522	0.1564	0.4985
N=1,000	SNR = 1/0.2	SNR = 3/1	SNR = 1/0.2	SNR = 3/1	SNR = 1/0.2	SNR = 3/1	SNR = 1/0.2	SNR = 3/1
GPR + Normal	1.0000	1.0000	0.2577	0.3054	0.1544	0.1190	0.7423	0.6946
STGP	1.0000	1.0000	0.5590	0.5367	0.0414	0.0457	0.4410	0.4633
RTGP-Gibbs	0.9952	0.9960	0.8698	0.8278	0.0082	0.0113	0.1302	0.1722
RTGP-VI	0.9992	1.0000	0.8436	0.5400	0.0100	0.0452	0.1564	0.4600
N=2,000	SNR = 1/0.2	SNR = 3/1	SNR = 1/0.2	SNR = 3/1	SNR = 1/0.2	SNR = 3/1	SNR = 1/0.2	SNR = 3/1
GPR + Normal	1.0000	1.0000	0.2078	0.2549	0.2003	0.1552	0.7922	0.7451
STGP	1.0000	1.0000	0.5763	0.5631	0.0384	0.0407	0.4237	0.4369
RTGP-Gibbs	0.9831	0.9952	0.9598	0.8619	0.0024	0.0088	0.0402	0.1381
RTGP-VI	1.0000	1.0000	0.8069	0.5928	0.0126	0.0359	0.1931	0.4072

Table C.4: Evaluation of inference results with TPR, TDR, FPR, and FDR for a simulation study setting with varying sample sizes $N = \{500; 1,000; 2,000\}$ and signal-to-noise rations $SNR = \{1/0.2; 3/1\}$ for the methods GPR+Normal, STGP, RTGP-Gibbs, and RTGP-VI.

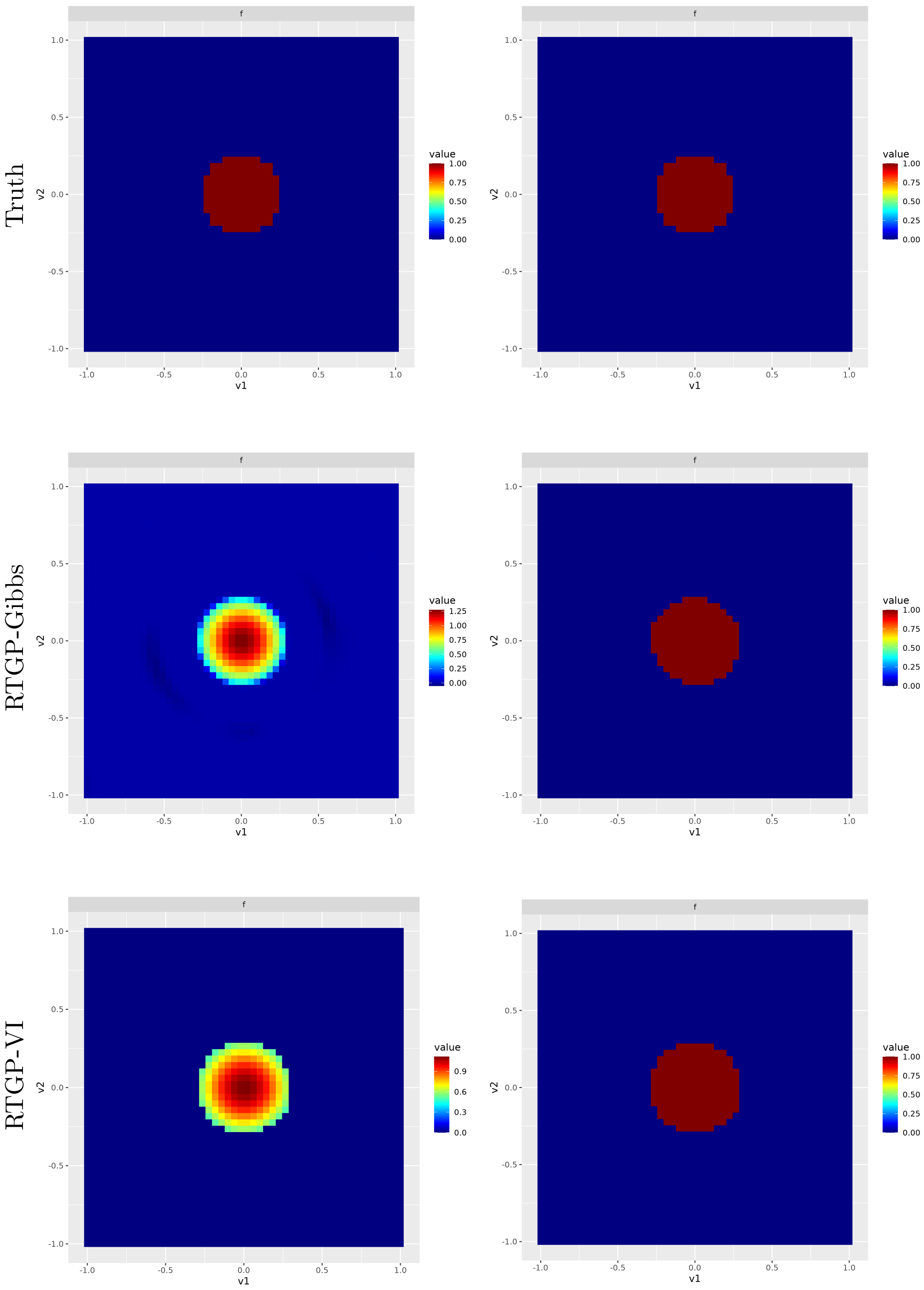
N = 500	SNR=1/0.2	SNR=3/1
GPR + Normal	00:00:00	00:00:00
STGP	00:06:25	00:06:13
RTGP-Gibbs	09:14:15	09:23:20
RTGP-VI	00:03:28	00:03:12
N = 1,000	SNR=1/0.2	SNR=3/1
Horseshoe	00:00:00	00:00:00
STGP	00:12:48	00:13:33
RTGP-Gibbs	13:42:33	13:51:50
RTGP-VI	00:06:28	00:06:32
N = 2,000	SNR=1/0.2	SNR=3/1
Horseshoe	00:00:00	00:00:00
STGP	00:32:50	00:33:05
RTGP-Gibbs	18:28:03	18:09:41
RTGP-VI	00:14:12	00:15:33

Table C.5: Run time evaluation of the models, GPR + Normal, STGP ($n_{\text{iter}} = 5,000$, $n_{\text{burn-in}} = 2,000$), RTGP-Gibbs ($n_{\text{iter}} = 10,000$, $n_{\text{burn-in}} = 5,000$), and RTGP-VI model (in hours).



(a) Parameter map **(b)** Binary significance map

Figure C.2: Comparison of (a) parameter maps and (b) binary significance maps of baseline methods (GPR + Normal, STGP) with truth maps with a simulation study setting of $N = 500$, $M = 2,500$, $L = 66$, signal $= 1$ and noise $= 0.2$.



(a) Parameter map

(b) Binary significance map

Figure C.3: Comparison of (a) parameter maps and (b) binary significance maps estimated via RTGP model (Gibbs, VI) with truth maps with a simulation study setting of $N = 500$, $M = 2,500$, $L = 66$, signal $= 1$ and noise $= 0.2$.

C.4 Image-on-Scalar Regression: Surface-based Mass-univariate Approach

In our work, we are interested in scalar-on-image regression problems, specifically the analysis of how task-based fMRI data is predictive of intelligence scores in the ABCD study; however, in the literature the question has been studied from the reverse perspective as an image-on-scalar regression problem, see Makowski et al. (2023) for an evaluation of the analysis. We repeat the analysis by performing a simple Normal linear regression at each vertex for the dataset described in Section 4.3.2.1 associating the test statistics resulting from a first-level fMRI analysis for the emotional n-back task with a 2- vs 0-back contrast with intelligence scores and confounding variables.

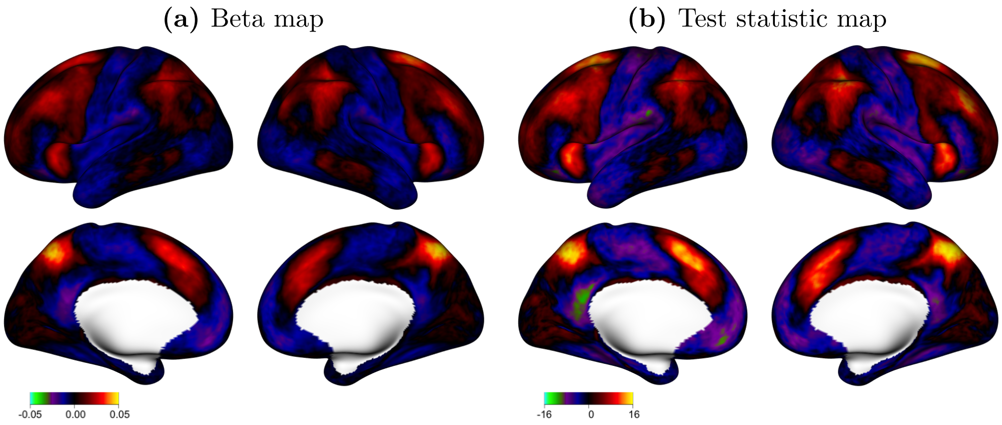

(a) Beta map **(b)** Test statistic map

Figure C.4: (a) Beta and (b) test statistic map displaying the estimated coefficients and test statistic values for the covariate "intelligence scores" in the mass-univariate analysis.

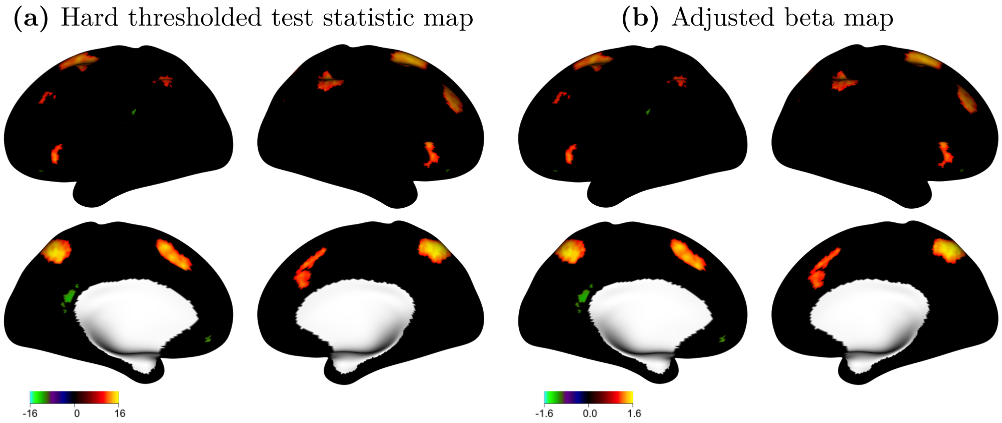

(a) Hard thresholded test statistic map **(b)** Adjusted beta map

Figure C.5: (a) Test statistic map (hard thresholded at a value of 10) and (b) adjusted beta map (test statistic map divided by 10) displaying the artificially created hard thresholded effect maps used in the subsampled surface-based simulation study as a true beta map (before subsampling).

C.5 Further Results on Simulation Study: Surface-based Representation

C.5.1 Simulation Study: Baseline Models

In the surface-based simulation study we compare our model RTGP to six baseline models: Ridge, LASSO, BR + Normal, BR + Horseshoe, GPR + Normal, and GPR + Horseshoe.

- **Ridge regression:** $\boldsymbol{y} = \beta_0 + \boldsymbol{X}\boldsymbol{\beta}(\boldsymbol{s}) + \eta \sum_{j=1}^{M} \beta(s_j)^2 + \boldsymbol{\epsilon}$ where image coefficients $\boldsymbol{\beta}(\boldsymbol{s})$ are regularised by considering the sum of squares of the magnitude of the coefficients in the cost function. We perform 100-fold cross validation to determine the regularisation parameter η.

- **LASSO regression:** $\boldsymbol{y} = \beta_0 + \boldsymbol{X}\boldsymbol{\beta}(\boldsymbol{s}) + \eta \sum_{j=1}^{M} |\beta(s_j)| + \boldsymbol{\epsilon}$ are regularised by considering the absolute value of the magnitude of the coefficients in the cost function. We perform 100-fold cross validation to determine the regularisation parameter η.

- **BR + Normal:** $\boldsymbol{y} = \beta_0 + \boldsymbol{X}\boldsymbol{\beta}(\boldsymbol{s}) + \boldsymbol{\epsilon}$ with an independent Normal prior on each of the image coefficients $\beta(s_j) \sim \mathcal{N}(0, \tau^2)$ for each vertex $j = 1, \ldots, M$.

- **BR + Horseshoe:** $\boldsymbol{y} = \beta_0 + \boldsymbol{X}\boldsymbol{\beta}(\boldsymbol{s}) + \boldsymbol{\epsilon}$ with an independent Horseshoe prior on each of the image coefficients $\beta(s_j) \sim \mathcal{N}(0, \eta_j^2\tau^2)$ and $\eta_j \sim \mathrm{Half-Cauchy}(0, 1)$ for each vertex $j = 1, \ldots, M$.

- **GPR + Normal:** $\boldsymbol{y} = \beta_0 + (\boldsymbol{X}(\boldsymbol{\Psi} \odot \boldsymbol{\lambda}))\boldsymbol{\theta} + \boldsymbol{\epsilon}$ with an independent Normal prior on each of the basis coefficients $\theta_l \sim \mathcal{N}(0, \tau^2)$ for each basis $l = 1, \ldots, L$. We use the same bases $\boldsymbol{\Psi} \odot \boldsymbol{\lambda}$ determined via the kernel decomposition needed for the RTGP model to transform the input images $\boldsymbol{X}$ for this baseline

Gaussian process regression model.

- **GPR + Horseshoe:** $\boldsymbol{y} = \beta_0 + (\boldsymbol{X}(\boldsymbol{\Psi} \odot \boldsymbol{\lambda}))\boldsymbol{\theta} + \boldsymbol{\epsilon}$ with an independent Horseshoe prior on each of the basis coefficients $\theta_l \sim \mathcal{N}(0, \eta_j^2\tau^2)$ and $\eta_j \sim \text{Half} - \text{Cauchy}(0,1)$ for each basis $l = 1, \ldots, L$. We use the same bases $\boldsymbol{\Psi} \odot \boldsymbol{\lambda}$ determined via the kernel decomposition needed for the RTGP model to transform the input images $\boldsymbol{X}$ for this baseline Gaussian process regression model.

C.5.2 RTGP: Variational Inference vs. Gibbs Sampler

N=500	**Bias**	**MSE**
RTGP-Gibbs	5.19 (1.55)	3.35 (0.21)
RTGP-VI	6.56 (1.97)	3.23 (0.69)
N=1,000	**Bias**	**MSE**
RTGP-Gibbs	5.23 (0.31)	3.30 (0.18)
RTGP-VI	3.74 (0.52)	2.45 (0.33)
N=2,000	**Bias**	**MSE**
RTGP-Gibbs	4.08 (0.24)	2.16 (0.06)
RTGP-VI	3.22 (0.26)	2.09 (0.18)

Table C.6: Evaluation of parameter estimate results with absolute bias and MSE mean (standard error) for a simulation study setting with varying sample sizes $N = \{500; 1,000; 2,000\}$ for our model RTGP estimated via variational inference (RTGP-VI) and estimated via a Gibbs sampler (RTGP-Gibbs). The values are multiplied by a scaling factor of 10^2 for clarity.

N=500	**R^2 (train)**	**MSE (train)**	**R^2 (test)**	**MSE (test)**
RTGP-Gibbs	50.45 (1.34)	3.95 (0.32)	40.03 (2.38)	4.88 (0.31)
RTGP-VI	49.89 (1.41)	3.92 (0.31)	39.38 (2.33)	4.86 (0.30)
N=1,000	**R^2 (train)**	**MSE (train)**	**R^2 (test)**	**MSE (test)**
RTGP-Gibbs	43.95 (2.55)	4.10 (0.18)	42.10 (1.82)	4.29 (0.14)
RTGP-VI	44.05 (2.30)	4.08 (0.19)	41.75 (2.23)	4.32 (0.17)
N=2,000	**R^2 (train)**	**MSE (train)**	**R^2 (test)**	**MSE (test)**
RTGP-Gibbs	45.09 (1.65)	4.03 (0.14)	42.43 (1.79)	4.25 (0.11)
RTGP-VI	42.22 (1.60)	4.20 (0.14)	42.66 (2.32)	4.26 (0.20)

Table C.7: Evaluation of prediction results with R^2 (in%) and predictive MSE (calculated for the training and test set) mean (standard error) for a simulation study setting with varying sample sizes $N = \{500; 1,000; 2,000\}$ for our model RTGP estimated via variational inference (RTGP-VI) and estimated via a Gibbs sampler (RTGP-Gibbs). The MSE values are multiplied by a scaling factor of 10^2 for clarity.

N=500	**TPR**	**TDR**	**FPR**	**FDR**
RTGP-Gibbs	80.03 (5.99)	58.82 (10.89)	5.29 (4.92)	41.83 (12.34)
RTGP-VI	75.22 (7.03)	56.85 (17.96)	6.23 (5.78)	44.15 (17.96)
N=1,000	**TPR**	**TDR**	**FPR**	**FDR**
RTGP-Gibbs	82.89 (9.30)	59.78 (8.56)	2.54 (1.05)	41.01 (8.56)
RTGP-VI	79.56 (5.81)	60.72 (8.05)	2.86 (1.08)	39.28 (8.05)
N=2,000	**TPR**	**TDR**	**FPR**	**FDR**
RTGP-Gibbs	89.52 (1.08)	61.45 (2.50)	2.55 (0.52)	34.91 (2.50)
RTGP-VI	79.33 (3.19)	66.70 (3.20)	2.10 (0.33)	33.30 (3.20)

Table C.8: Evaluation of inference results with TPR, TDR, FPR, and FDR mean (standard error) for a simulation study setting with varying sample sizes $N = \{500; 1,000; 2,000\}$ for our model RTGP estimated via variational inference (RTGP-VI) and estimated via a Gibbs sampler (RTGP-Gibbs). The values are multiplied by a scaling factor of 10^2 for clarity.

C.5.3 Evaluation of Parameter Estimates (Effect vs No Effect)

N=500	**Bias (Total)**	Bias (Effect)	Bias (No Effect)	**MSE (Total)**	MSE (Effect)	MSE (No Effect)
RTGP	**6.56** (1.97)	**35.87** (3.02)	5.01 (2.17)	**3.23** (0.69)	**19.70** (4.36)	2.78 (2.17)
GPR + Normal	11.77 (0.40)	52.89 (2.15)	9.60 (0.45)	3.37 (0.14)	33.13 (1.95)	1.80 (0.45)
GPR + Horseshoe	12.48 (1.67)	58.39 (2.91)	10.06 (1.77)	3.74 (0.58)	39.37 (3.56)	1.86 (1.77)
BR + Normal	10.81 (0.19)	58.13 (2.10)	8.32 (0.24)	3.28 (0.11)	40.69 (2.76)	1.31 (0.24)
BR + Horseshoe	8.11 (0.31)	119.26 (6.84)	2.26 (0.25)	19.14 (4.16)	374.95 (85.44)	**0.40** (0.25)
Ridge	11.50 (0.33)	56.54 (4.04)	9.13 (0.48)	3.47 (0.16)	37.95 (4.52)	1.65 (0.48)
LASSO	7.89 (0.43)	136.53 (7.80)	**1.12** (0.46)	22.24 (4.74)	407.34 (102.82)	1.96 (0.46)
N=1,000	**Bias (Total)**	Bias (Effect)	Bias (No Effect)	**MSE (Total)**	MSE (Effect)	MSE (No Effect)
RTGP	**3.74** (0.52)	**37.88** (3.78)	1.94 (0.53)	**2.45** (0.33)	**22.57** (4.65)	1.39 (0.53)
GPR + Normal	10.31 (0.50)	53.05 (2.33)	8.06 (0.60)	3.04 (0.09)	33.32 (2.61)	1.44 (0.60)
GPR + Horseshoe	11.04 (0.60)	54.22 (2.80)	8.77 (0.69)	3.21 (0.16)	34.67 (3.07)	1.56 (0.69)
BR + Normal	9.78 (0.29)	55.92 (1.67)	7.35 (0.37)	2.97 (0.06)	37.83 (2.21)	1.14 (0.37)
BR + Horseshoe	7.79 (0.85)	115.73 (22.39)	2.11 (0.42)	22.38 (15.89)	442.47 (318.66)	**0.26** (0.42)
Ridge	10.51 (0.35)	53.32 (2.73)	8.25 (0.45)	3.17 (0.10)	34.41 (2.98)	1.52 (0.45)
LASSO	7.38 (0.32)	128.07 (7.44)	**1.02** (0.33)	19.11 (3.80)	346.31 (78.92)	1.88 (0.33)
N=2,000	**Bias (Total)**	Bias (Effect)	Bias (No Effect)	**MSE (Total)**	MSE (Effect)	MSE (No Effect)
RTGP	**3.22** (0.26)	**35.62** (3.07)	1.51 (0.21)	**2.09** (0.18)	**20.51** (2.75)	1.12 (0.21)
GPR + Normal	9.66 (0.22)	51.98 (2.51)	7.43 (0.33)	2.83 (0.05)	31.49 (2.80)	1.32 (0.33)
GPR + Horseshoe	11.44 (1.13)	49.83 (2.22)	9.42 (1.17)	3.18 (0.33)	29.60 (2.50)	1.78 (1.18)
BR + Normal	9.33 (0.20)	53.30 (1.49)	7.01 (0.26)	2.80 (0.03)	34.11 (1.99)	1.15 (0.26)
BR + Horseshoe	7.75 (0.41)	120.57 (8.17)	1.80 (0.16)	21.33 (6.57)	418.89 (128.92)	**0.40** (0.16)
Ridge	10.22 (0.23)	50.05 (1.52)	8.12 (0.30)	3.04 (0.06)	30.70 (1.82)	1.58 (0.30)
LASSO	7.17 (0.28)	127.05 (5.29)	**0.86** (0.17)	17.31 (3.85)	322.53 (75.64)	1.24 (0.17)

Table C.9: Evaluation of parameter estimate results with bias and MSE mean (standard error) for total number of vertices, in areas of true effect, and in areas of no effect for a simulation study setting with varying sample sizes $N = \{500; 1,000; 2,000\}$ for our model RTGP and the baseline models GPR + Normal, GPR + Horseshoe, BR + Normal, BR + Horseshoe, Ridge and LASSO. The values are multiplied by a scaling factor of 10^2 for clarity.

C.5.4 Computational Run Time

	N = 500	**N = 1,000**	**N = 2,000**
Ridge	00:00:01	00:00:02	00:00:04
LASSO	00:00:01	00:00:03	00:00:05
BR + Normal	00:01:03	00:03:21	00:11:31
BR + Horseshoe	00:01:12	00:04:01	00:12:23
GPR + Normal	00:00:01	00:00:01	00:00:01
GPR + Horseshoe	00:00:01	00:00:01	00:00:02
RTGP-VI	00:10:55	00:30:12	00:45:21
RTGP-Gibbs	04:44:02	09:39:45	18:45:21

Table C.10: Run time evaluation across varying sample sizes of the baseline models, Ridge, LASSO, BR + Normal ($n_{\text{iter}} = 5,000$, $n_{\text{burn-in}} = 2,000$), BR + Horseshoe ($n_{\text{iter}} = 5,000$, $n_{\text{burn-in}} = 2,000$), GPR + Normal ($n_{\text{iter}} = 5,000$, $n_{\text{burn-in}} = 2,000$), and GPR + Horseshoe ($n_{\text{iter}} = 5,000$, $n_{\text{burn-in}} = 2,000$), and our models RTGP-Gibbs ($n_{\text{iter}} = 10,000$, $n_{\text{burn-in}} = 5,000$) and RTGP-VI (in hours).

C.5.5 Sensitivity Analysis: Initialisation

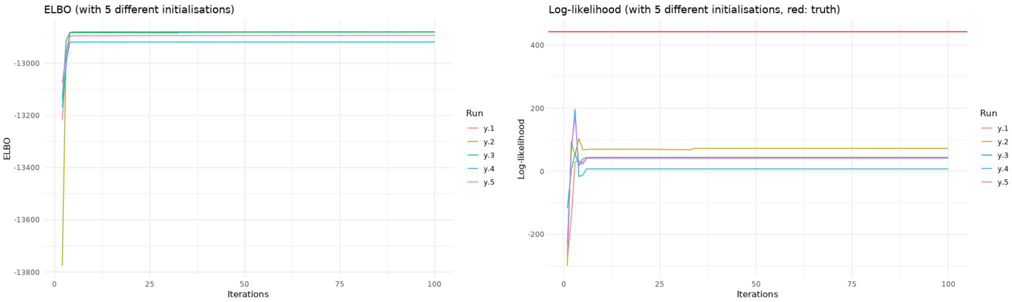


Figure C.6: Sensitivity analysis of initialisation of parameters by randomly perturbing the GPR + Horseshoe model estimates with a Normal noise draw and re-estimating the RTGP with 5 of those different initialisations. (Left) ELBO plot which shows monotonously increasing ELBO values across iterations for 5 initialisation and (right) log-likelihood plot which shows log-likelihood values across iterations for 5 initialisations in addition to the true value (in red).

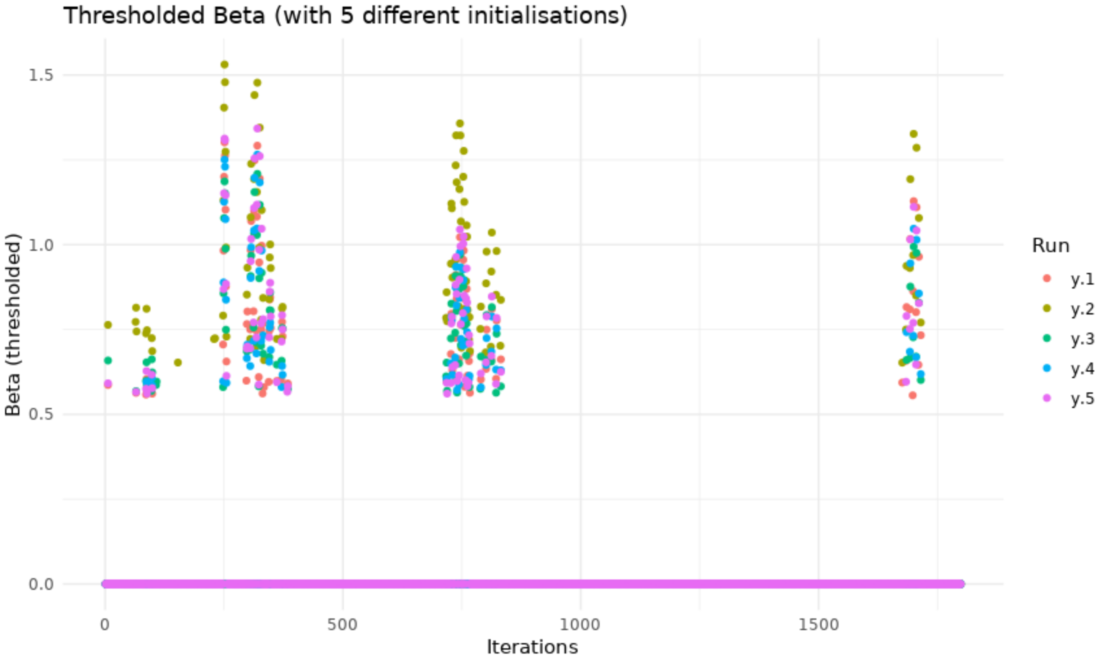


Figure C.7: Sensitivity analysis of initialisation of parameters by randomly perturbing the GPR + Horseshoe model estimates with a Normal noise draw and re-estimating the RTGP with 5 of those different initialisations. Scatterplot for RTGP model estimates with 5 different initialisations.

	Run 1	Run 2	Run 3	Run 4	Run 5
ELBO	-12879	-12880	-12880	-12918	-12893
Log-likelihood	43.6990	72.6133	43.8297	7.7484	41.3881
Threshold $\hat{\delta}$	0.5517	0.6513	0.5615	0.5619	0.5587
Intercept $\hat{\beta}_0$	2.0041	2.0028	2.0040	2.0053	2.0048
Std $\hat{\sigma}_\beta$	0.2461	0.3349	0.2488	0.2556	0.2791
Std $\hat{\sigma}_\alpha$	0.0011	0.0011	0.0011	0.0011	0.0011
Std $\hat{\sigma}_\epsilon$	0.1436	0.1415	0.1436	0.1462	0.1438
# of active vertices	95	99	86	75	86

Table C.11: Comparison of estimated quantities across 5 different initialisation runs.

Inference	TPR	TDR	FPR	FDR
Run 1	0.7333	0.6947	0.0170	0.3053
Run 2	0.7556	0.6939	0.0176	0.3061
Run 3	0.6778	0.7093	0.0146	0.2907
Run 4	0.6222	0.7467	0.0111	0.2533
Run 5	0.6556	0.6860	0.0158	0.3140

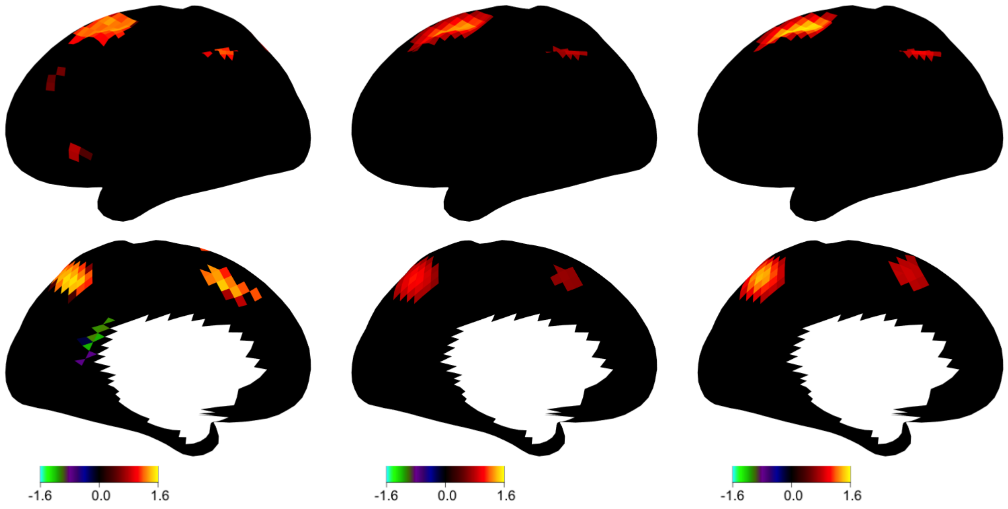
Table C.12: Comparison of inference results across 5 different initialisation runs.

Parameters	Bias	MSE
Run 1	-0.0055	0.0231
Run 2	0.0043	0.0221
Run 3	-0.0108	0.0235
Run 4	-0.0144	0.0253
Run 5	-0.0088	0.0251

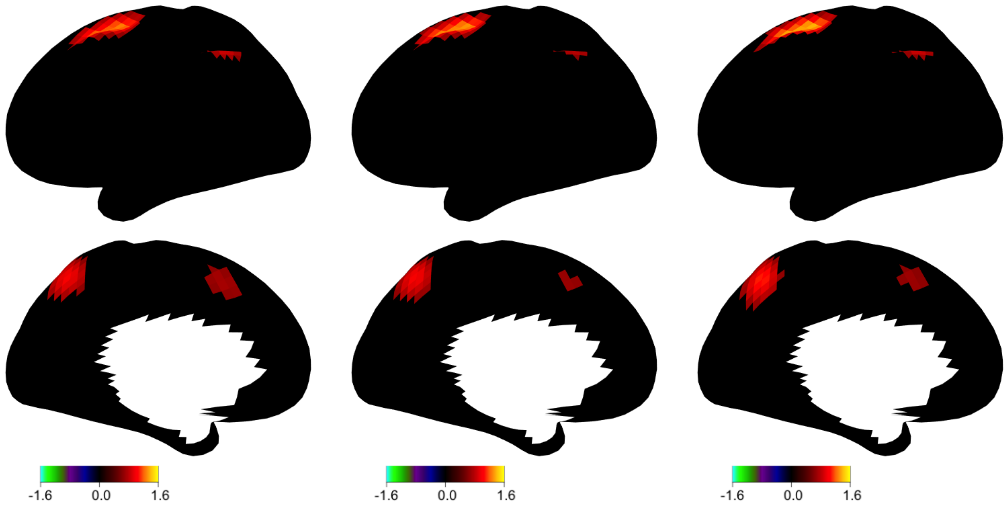
Table C.13: Comparison of parameter estimate results across 5 different initialisation runs.

Prediction	R^2 (train)	MSE (train)	R^2 (test)	MSE (test)
Run 1	0.4014	0.0412	0.4372	0.0444
Run 2	0.4232	0.0400	0.4376	0.0428
Run 3	0.4135	0.0412	0.4378	0.0446
Run 4	0.3891	0.0427	0.4392	0.0461
Run 5	0.4030	0.0413	0.4388	0.0448

Table C.14: Comparison of prediction results across 5 different initialisation runs.



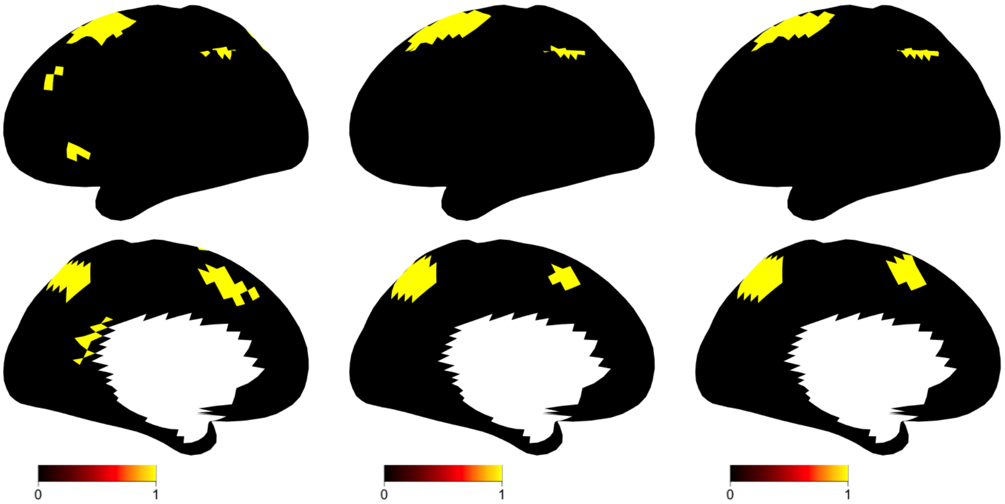
(a) True beta **(b)** Run 1 **(c)** Run 2



(d) Run 3 **(e)** Run 4 **(f)** Run 5

Figure C.8: Comparison of true beta map with the estimated beta maps of the RTGP model with the 5 different initialisations.

(a) True beta **(b)** Run 1 **(c)** Run 2



(d) Run 3 **(e)** Run 4 **(f)** Run 5

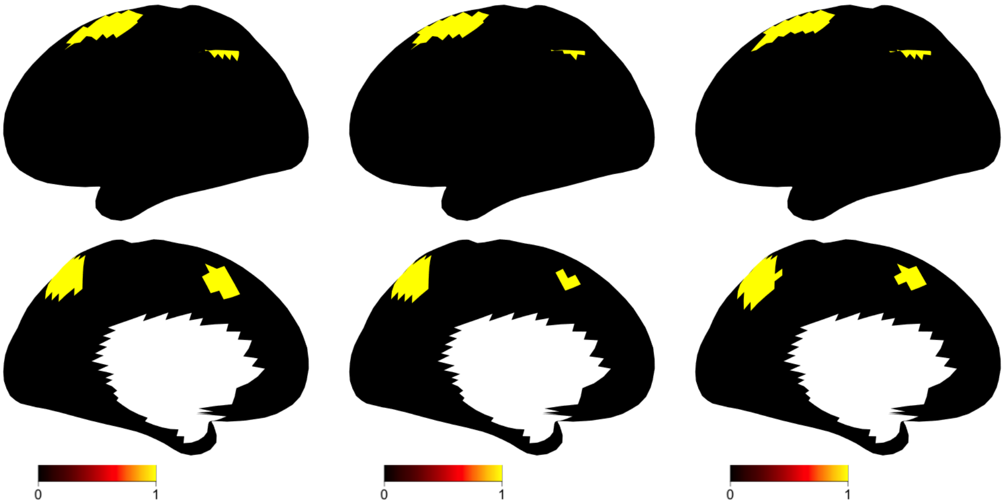


Figure C.9: Comparison of true binary significance map with the estimated binary significance maps of the RTGP model with the 5 different initialisations.

C.5.6 Sensitivity Analysis: Data Removal

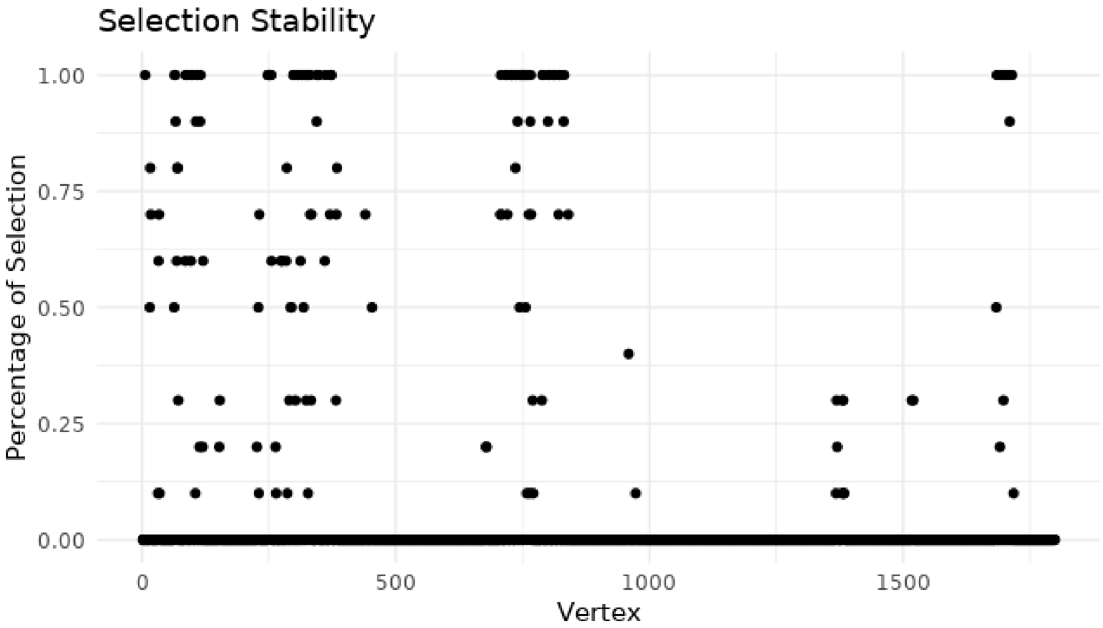


Figure C.10: Compare selection stability of estimating RTGP model after removing 10% of the training data (average of 10 CV runs).

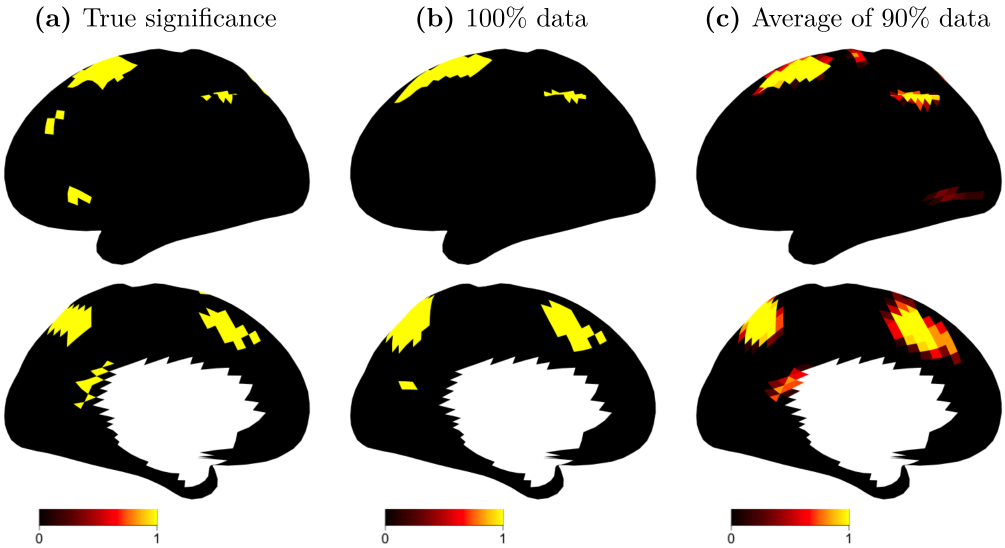

(a) True significance **(b)** 100% data **(c)** Average of 90% data

Figure C.11: Comparison of true binary significance map with the estimated binary significance map of the RTGP model with 100% of the training data and the average probability of activation across 10 CV runs when removing 10% of the training data and re-estimating the RTGP model.

Parameters	Bias	MSE		
100%	0.0099	0.0213		
90% (threshold at 90%)	-0.0006	0.0214		
Prediction	R^2 (train)	MSE (train)	R^2 (test)	MSE (test)
100%	0.4308	0.0400	0.4350	0.0438
90% (threshold at 90%)	0.4184	0.0401	0.4313	0.0431
Inference	TPR	TDR	FPR	FDR
100%	0.8822	0.6852	0.0199	0.3148
90% (threshold at 90%)	0.6889	0.7561	0.0117	0.2439

Table C.15: Comparison of parameter estimates, prediction, and inference results of RTGP model estimated with 100% of the training data and with average of 10 datasets where 10% of the data have been removed and inference is assessed with thresholding the selection stability at 90%.

C.5.7 Sensitivity Analysis: Variance Hyperparameters

Estimates	$\hat{\sigma}_\epsilon$	$\hat{\sigma}_\beta$	$\hat{\sigma}_\alpha$	
$a_\alpha/a_\beta/a_\epsilon = 0.01$	0.1395	0.3597	0.0033	
$a_\alpha/a_\beta/a_\epsilon = 0.001$	0.1414	0.3674	0.0011	
$a_\alpha/a_\beta/a_\epsilon = 0.0001$	0.1413	0.3877	0.0003	
Parameters	Bias	MSE		
$a_\alpha/a_\beta/a_\epsilon = 0.01$	0.0076	0.0190		
$a_\alpha/a_\beta/a_\epsilon = 0.001$	0.0099	0.0213		
$a_\alpha/a_\beta/a_\epsilon = 0.0001$	0.0031	0.0224		
Prediction	R^2 (train)	MSE (train)	R^2 (test)	MSE (test)
$a_\alpha/a_\beta/a_\epsilon = 0.01$	0.4403	0.0388	0.4377	0.0421
$a_\alpha/a_\beta/a_\epsilon = 0.001$	0.4308	0.0400	0.4358	0.0438
$a_\alpha/a_\beta/a_\epsilon = 0.0001$	0.4245	0.0399	0.4382	0.0430
Inference	TPR	TDR	FPR	FDR
$a_\alpha/a_\beta/a_\epsilon = 0.01$	0.8222	0.6916	0.0193	0.3084
$a_\alpha/a_\beta/a_\epsilon = 0.001$	0.8222	0.6852	0.0199	0.3148
$a_\alpha/a_\beta/a_\epsilon = 0.0001$	0.7222	0.7303	0.0140	0.2697

Table C.16: Comparison of estimates, parameter estimates, prediction, and inference results of RTGP model estimated with different variance hyperparameter settings.

Estimates	$\hat{\sigma}_\epsilon$	$\hat{\sigma}_\beta$	$\hat{\sigma}_\alpha$	
Inverse-Gamma	0.1414	0.3674	0.0011	
Half-Cauchy	0.1455	0.3901	0.0009	
Parameters	Bias	MSE		
Inverse-Gamma	0.0099	0.0213		
Half-Cauchy	0.0119	0.0217		
Prediction	R^2 (train)	MSE (train)	R^2 (test)	MSE (test)
Inverse-Gamma	0.4308	0.0400	0.4358	0.0438
Half-Cauchy	0.4341	0.0399	0.4371	0.0427
Inference	TPR	TDR	FPR	FDR
Inverse-Gamma	0.8222	0.6852	0.0199	0.3148
Half-Cauchy	0.8222	0.6491	0.0234	0.3509

Table C.17: Comparison of estimates, parameter estimates, prediction, and inference results of RTGP model estimated with different variance prior distributions (Inverse-Gamma, Half-Cauchy).

C.5.8 Inference: FDR Control

$$\frac{\sum_{j=1}^{M} I\{\pi(s_j) > \delta\}\gamma(s_j)}{\sum_{j=1}^{M} I(\pi(s_j) > \delta} \approx \alpha \tag{C.4}$$

- $\gamma(s_j)$: True binary significance result at location s_j.

- $\pi(s_j)$: Expected probability $\mathbb{E}[I(|\alpha(s_j)| > \delta)]$.

- δ: Threshold in RTGP model.

- α: Significance threshold to FDR control.

For comparison, we FDR control RTGP through equation and compare it to our baseline approaches, GPR + Normal and GPR + Horseshoe, which we also FDR control at the same $\alpha = 31\%$.

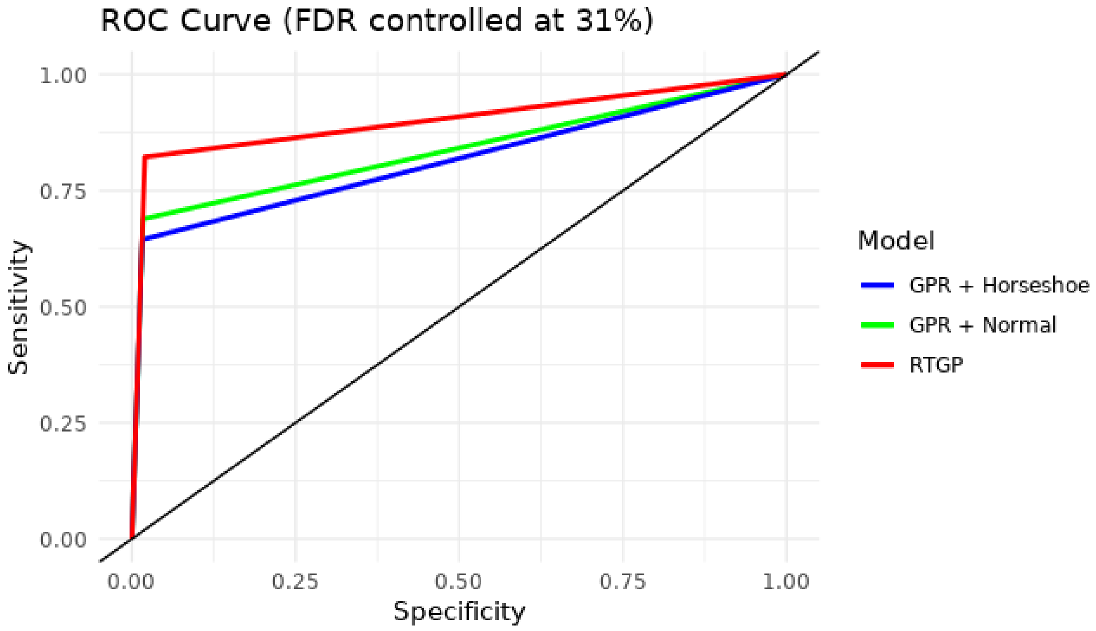


Figure C.12: ROC plot comparing the sensitivity with the specificity of RTGP, GPR + Normal, and GPR + Horseshoe when FDR controlling all methods at $\alpha = 31\%$.

Inference	TPR	TDR	FPR	FDR
GPR + Normal	0.6889	0.6813	0.0170	0.3187
GPR + Horseshoe	0.6444	0.6824	0.0158	0.3176
RTGP	0.8222	0.6852	0.0199	0.3148

Table C.18: Comparison of the inference results of RTGP, GPR + Normal, and GPR + Horseshoe when FDR controlling all methods at $\alpha = 31\%$.

C.5.9 Cluster Size Comparisons

We use the Human Connectome Workbench to identify clusters from significance maps, specifically the "`wb_command - cifti-find-clusters`"[2], where identify the clusters from the binary significance maps. Hence, the cluster-defining threshold is 0. Binary significance is determined in the RTGP model with $\mathbb{E}[I(|\alpha(s_j)| > \delta)] > 0.5$ for all vertices $j = 1, \ldots, M$ and in the GPR + Normal and GPR + Horseshoe models binary significance is determined by checking if 0 is included in the Bayesian 95%-HPDI credible interval. Note that this particular cluster command does not sort the clusters by location or size. Hence, the cluster number in each map does not match the cluster number in another map from a different method. For calculating the Rand index, we do match the numbers of the 4 biggest clusters occurring in the true cluster size map with the clusters appearing in the same spatial region (if they have at least one overlapping vertex) for the cluster maps of RTGP, GPR + Normal, and GPR + Horseshoe. We use the R package "`fossil`" to calculate the Rand Index (RI) and the Adjusted Rand Index (ARI) to measure the similarity of two cluster maps and in the case of the adjusted version also accounting for the chance grouping of the elements in each cluster map.

---

[2] `https://www.humanconnectome.org/software/workbench-command/-cifti-find-clusters`

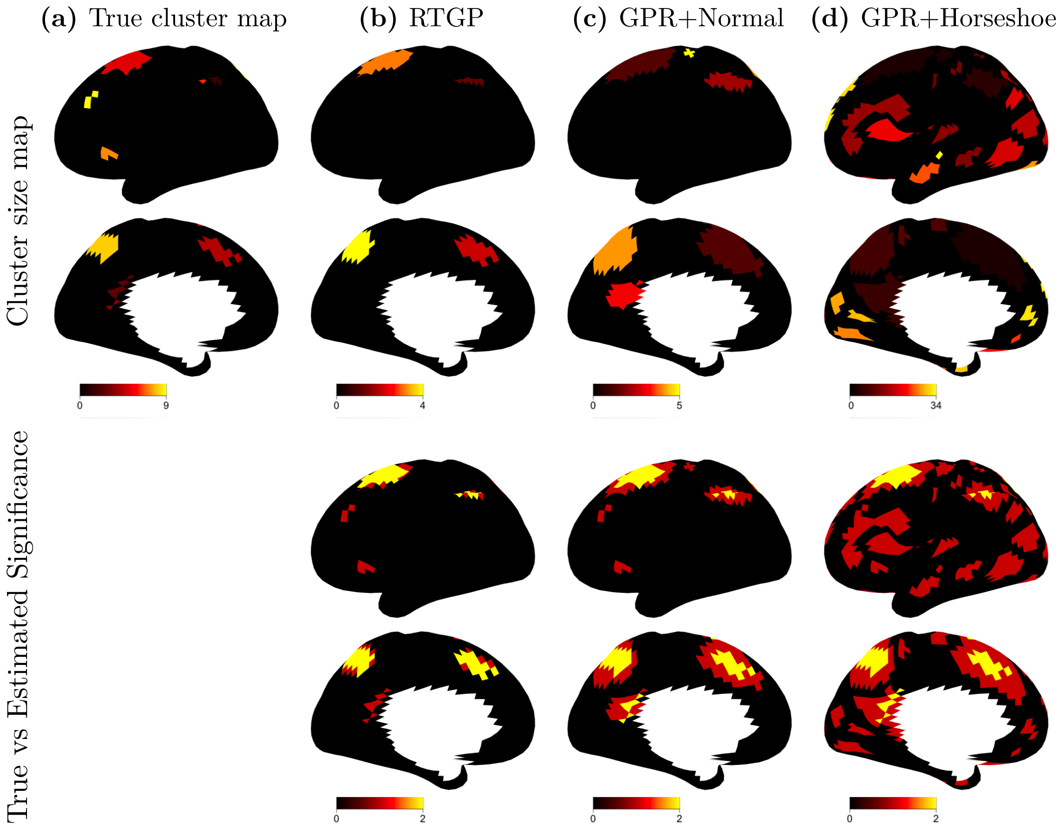

(a) True cluster map **(b)** RTGP **(c)** GPR+Normal **(d)** GPR+Horseshoe

Figure C.13: Comparison of cluster size maps for the true significance map and the cluster size maps resulting from RTGP, GPR + Normal, and GPR + Horseshoe in the top row. The bottom row shows the overlap of the true significance map with the respective significance map from each method.

	Truth	GPR + Normal	GPR + Horseshoe	RTGP
Cluster 1	13	102	4	23
Cluster 2	4	50	133	14
Cluster 3	1	17	56	37
Cluster 4	11	57	35	27
Cluster 5	29	5	65	
Cluster 6	1		7	
Cluster 7	3		5	
Cluster 8	25		2	
Cluster 9	3		6	
Cluster 10			3	
Cluster 11			7	
...			...	
Cluster 34			1	
Total	90	231	551	108

Table C.19: Comparison of identified clusters and their respective cluster sizes. Note that the cluster numbers do not match location across the different methods.

	RI	ARI
GPR + Normal	84.31%	97.94%
GPR + Horseshoe	57.02%	55.82%
RTGP	95.05%	99.13%

Table C.20: Comparison of Rand Index and Adjusted Rand Index across different methods measuring the similarity between the true identified cluster maps and the estimated cluster maps.

C.6 Further ABCD Study Application Results

C.6.1 Correlation: Intelligence Scores and Task-based fMRI (Emotional n-back Task)

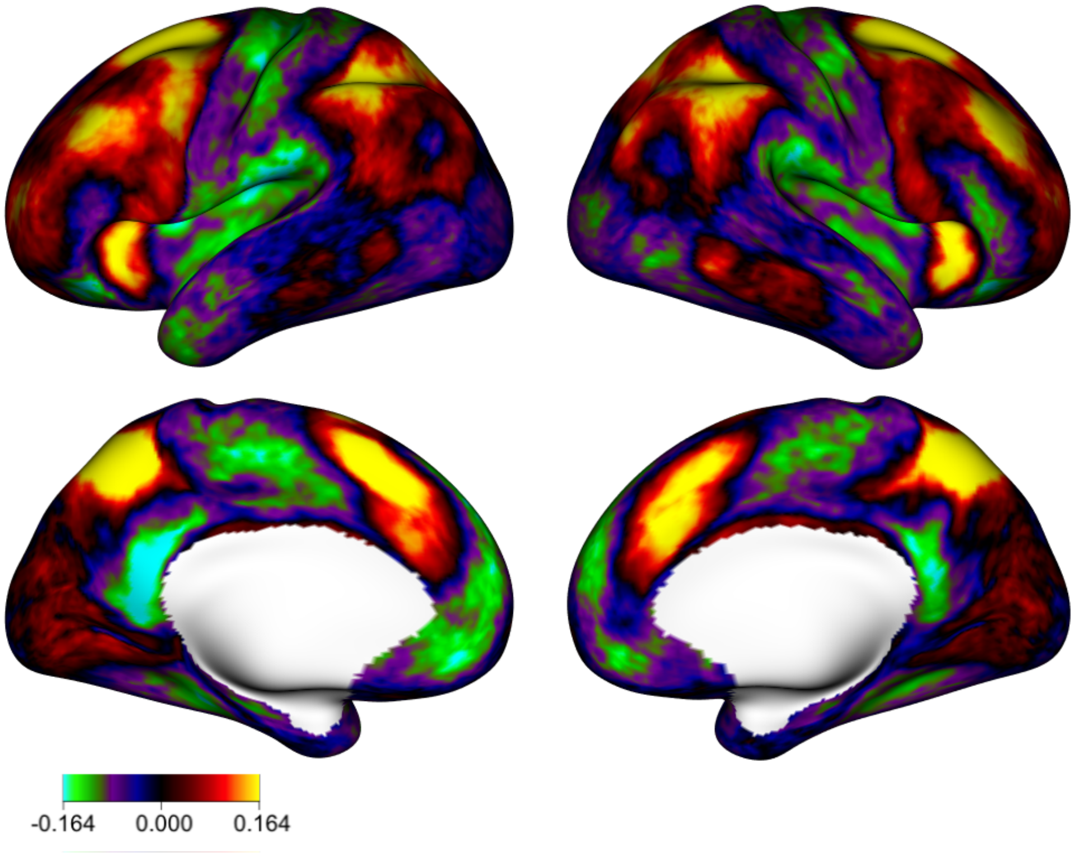


Figure C.14: Correlation map between the output variable intelligence scores and the input which is the test statistics of the first-level fMRI analysis of the 2- vs. 0-back contrast of the emotional n-back fMRI task.

C.6.2 Eigendecomposition of Kernel Matrix

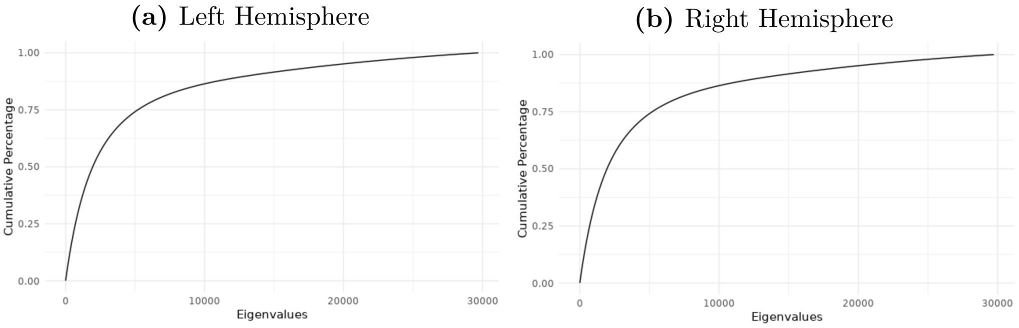


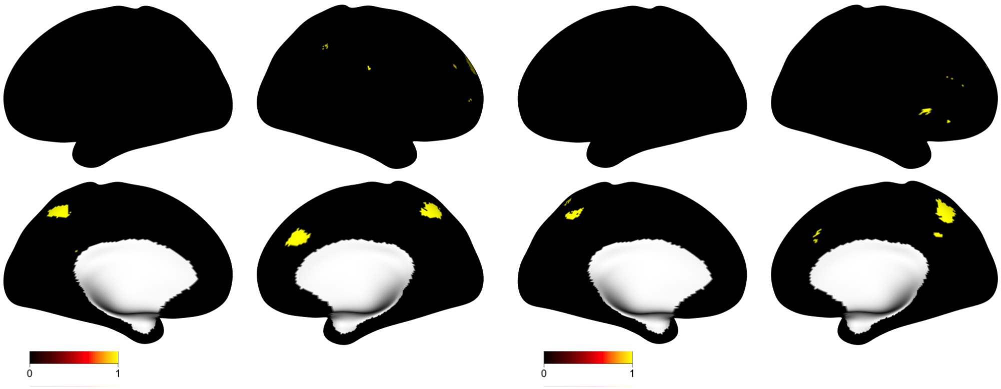
Figure C.15: Comparison of the total percentage of variation captured by the cumulative eigenvalues resulting from the eigendecomposition of a squared exponential kernel ($\boldsymbol{\xi} = (\phi = 0.077, \nu = 2)^T$) where the correlation matrix was determined with the geodesic distance on the sphere.

Total Variation	# of Bases (left)	# of Bases (right)
90%	13,181	13,200
80%	6,694	6,698
70%	4,159	4,162
60%	2,808	2,809
50%	1,950	1,951
40%	1,347	1,348
30%	893	894
20%	536	536
10%	244	244

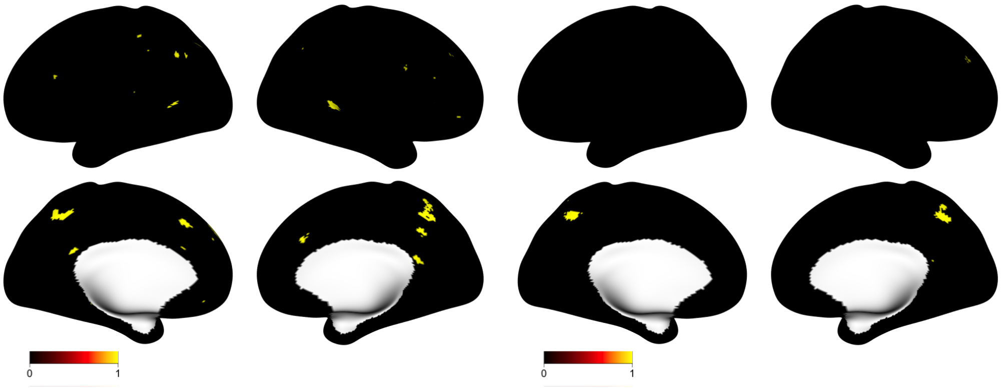
Table C.21: Number of eigenfunctions needed to capture $x\%$ of total variation with the eigendecomposition of the kernel for the left and right hemisphere of the cortex.

C.6.3 Kernel Hyperparameter Stability: GPR + Horseshoe

(a) $\boldsymbol{\xi} = (\phi = 0.01, \nu = 1.2)$ **(b)** $\boldsymbol{\xi} = (\phi = 0.05, \nu = 1.2)$



(c) $\boldsymbol{\xi} = (\phi = 0.1, \nu = 1.2)$ **(d)** $\boldsymbol{\xi} = (\phi = 0.2, \nu = 1.2)$



(e) $\boldsymbol{\xi} = (\phi = 0.3, \nu = 1.2)$ **(f)** $\boldsymbol{\xi} = (\phi = 0.4, \nu = 1.2)$

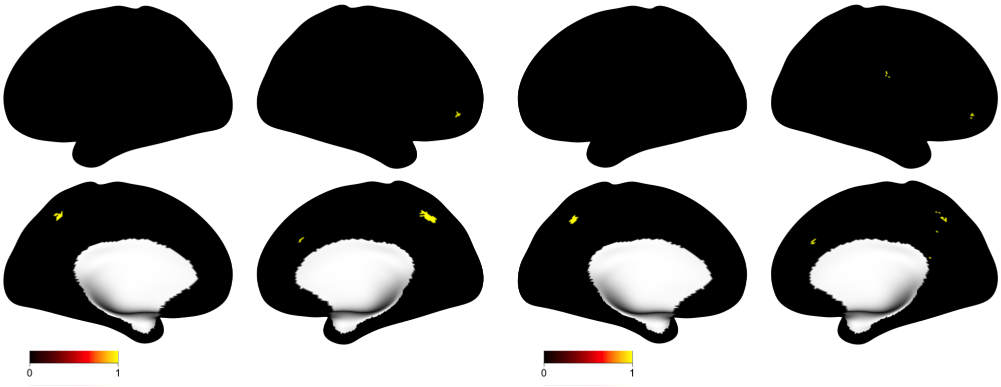
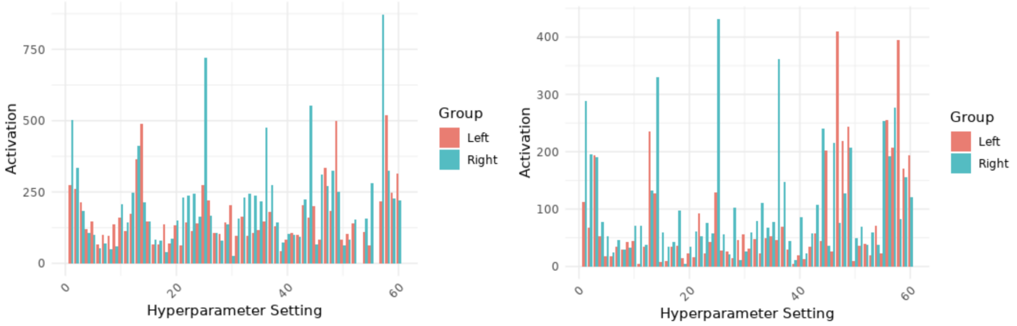
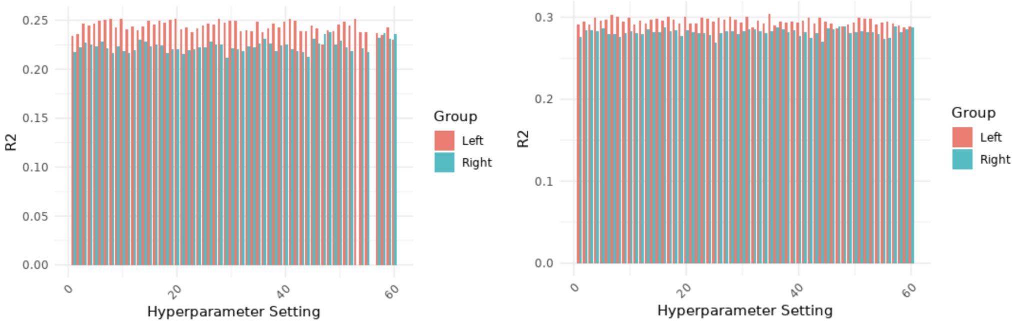


Figure C.16: Comparison of activation patterns / binary significance maps across different kernel hyperparameters for the GPR + Horseshoe model.

(a) Activation (no confounds) **(b)** Activation (with confounds)



(c) R^2 (no confounds) **(d)** R^2 (with confounds)



(e) MSE (no confounds) **(f)** MSE (with confounds)

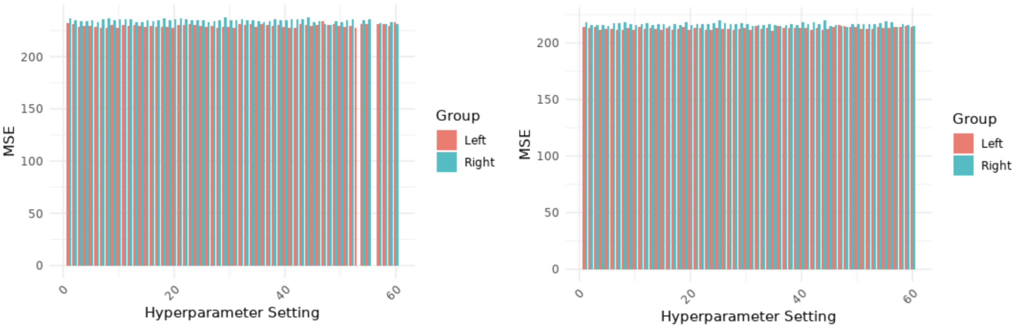
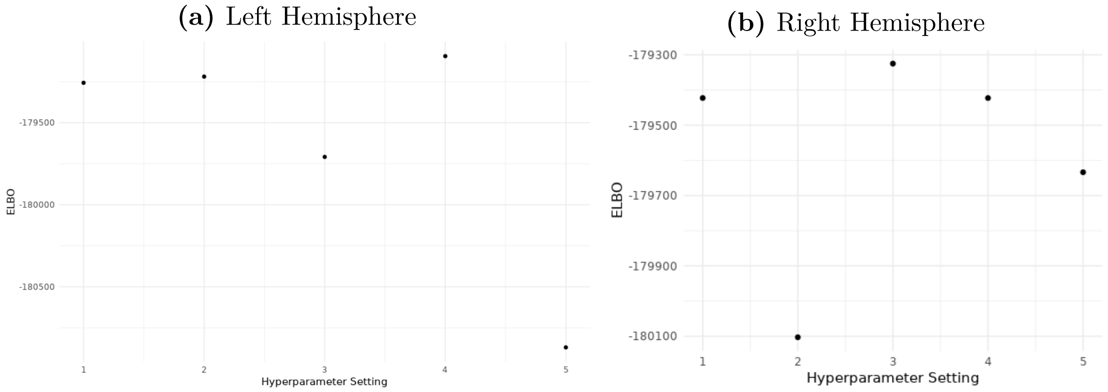


Figure C.17: Comparison of activation patterns, R^2 (test), and predictive MSE (test) across different kernel hyperparameters (60 combinations of $\phi = \{0.01, 0.05, 0.1, 0.2, \ldots, 1\}$ and $\nu = \{1.2, 1.4, \ldots, 2\}$) for the GPR + Horseshoe model and for models that include or exclude the confounding variables in the analysis.

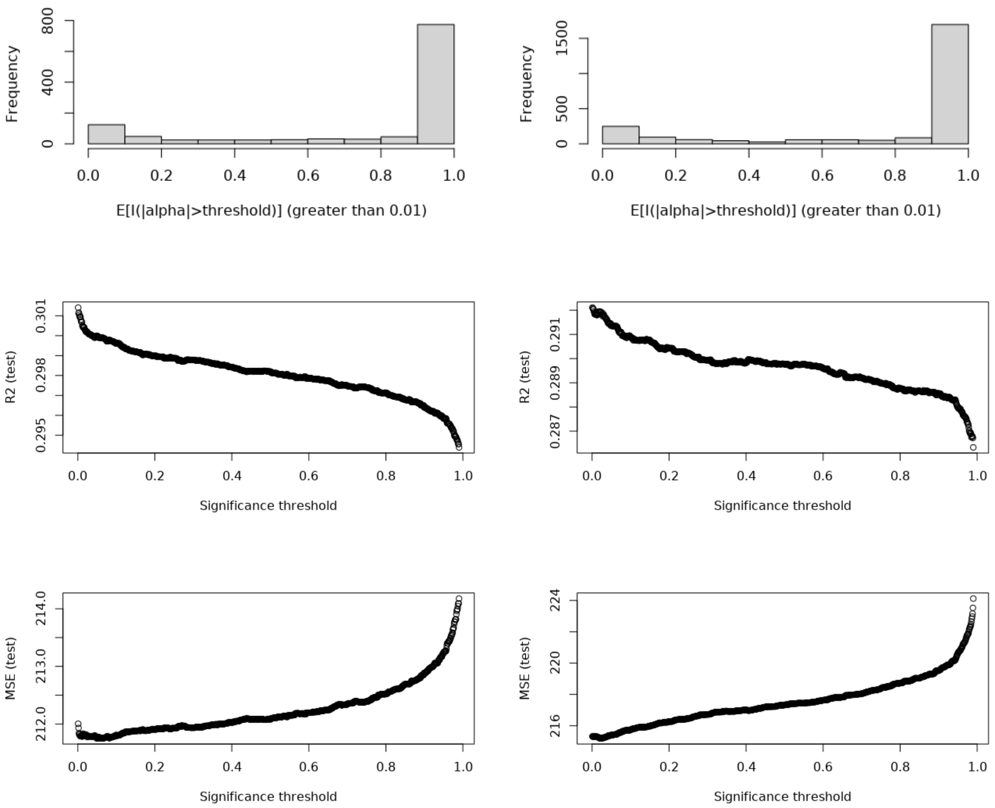
C.6.4 Kernel Hyperparameter Selection: RTGP Model


(a) Left Hemisphere

(b) Right Hemisphere

Figure C.18: Comparison of kernel hyperparameter settings $\phi = \{0.01, 0.05, 0.1, 0.2, 0.3\}$ and $\nu = 1.2$ by plotting the ELBO as a measure for model comparison between the different settings. The highest yielding ELBO for the RTGP model for the left hemisphere is $\phi = 0.2$ and for the right hemisphere is $\phi = 0.05$.

C.6.5 Selection of Significance Threshold



(a) Left Hemisphere **(b)** Right Hemisphere

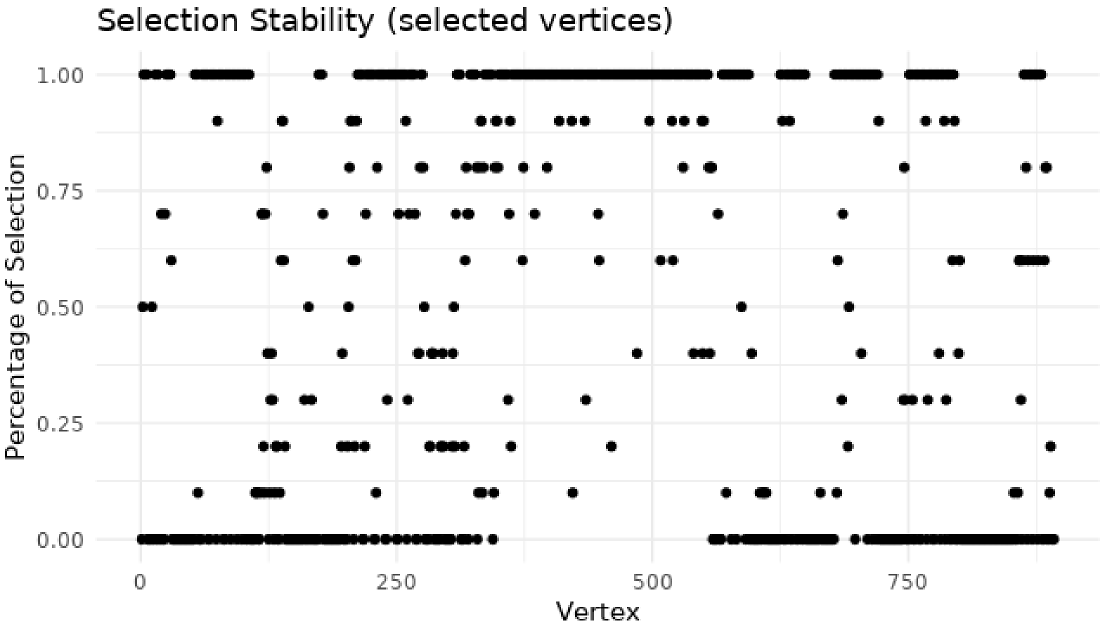
Figure C.19: Comparison of (top) expected values $\mathbb{E}[I(|\alpha(s_j)| > \delta)]$ for all vertices $j = 1, \dots, M$, (b) predictive R^2 (test), and (c) predictive MSE (test) for various significance thresholds. Normally, we apply the threshold of 0.5 based on the median probability model (Barbieri and Berger, 2004); however, one can choose more liberal or conservative thresholds if desired. The threshold with the lowest predictive MSE is at $\alpha_{sig} = 0.05$ for the left hemisphere and $\alpha_{sig} = 0.025$ for the right hemisphere.

Left Hemisphere	R^2 (train)	MSE (train)	R^2 (test)	MSE (test)
Ridge	0.6666	110.9729	0.1833	252.4871
LASSO	0.3545	188.6745	0.2834	216.8615
BR + Normal	1.0000	0.0201	0.0442	532.4117
BR + Horseshoe	0.4995	149.9042	0.2748	218.8589
GPR + Normal	0.2625	222.9945	0.1601	286.7683
GPR + Horseshoe	0.3109	198.3787	0.2997	211.6691
RTGP ($\alpha_{sig} = 0.05$)	0.2790	209.0666	0.3000	211.7499
Right Hemisphere	R^2 (train)	MSE (train)	R^2 (test)	MSE (test)
Ridge	0.6674	109.7410	0.1935	246.4008
LASSO	0.3618	186.8528	0.2843	216.6652
BR + Normal	1.0000	0.0067	0.0263	526.8691
BR + Horseshoe	0.5200	144.3401	0.2488	227.0925
GPR + Normal	0.2674	222.8687	0.1518	292.7009
GPR + Horseshoe	0.3159	196.8422	0.2839	215.9624
RTGP ($\alpha_{sig} = 0.025$)	0.2941	204.4201	0.2919	215.1883

Table C.22: Comparison of predictive results for training and test datasets for RTGP model (thresholded with significance threshold that yields the lowest predictive MSE) and various baseline methods.

C.6.6 Selection Stability

(a) Selection Stability



(b) Overlap **(c)** 100% Data **(d)** Averaged 90% Data

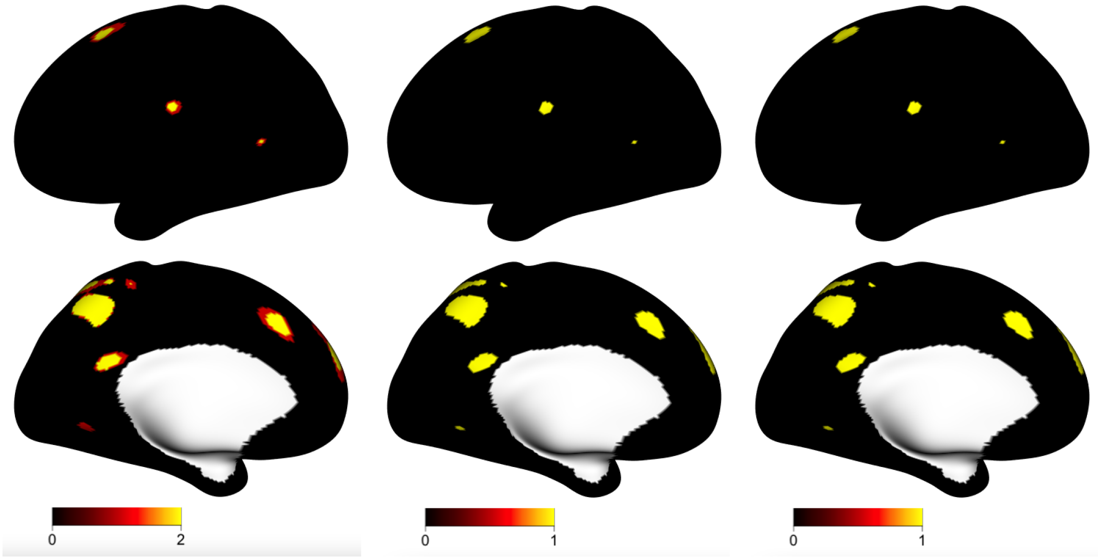


Figure C.20: (a) Selection stability for selected vertices (vertices that were active in the 100% data analysis), (b) overlap map between the 100% data binary significance results (red) and the averaged 10-fold CV 90% data analysis (yellow) where binary significance is determined by counting a vertex as significant if it has 90% selection stability in the 10-fold CV, (c) the binary significance map for the 100% data analysis, and (d) the binary significance map of the 10-fold CV 90% data analysis.

	R^2 (train)	MSE (train)	R^2 (test)	MSE (test)
Mean (left)	0.2348	228.0353	0.2163	230.6964
Sd (left)	0.0029	1.3614	0.0297	16.2700
Mean (right)	0.2604	222.0986	0.2420	228.8603
Sd (right)	0.0028	1.2347	0.0271	13.2318

Table C.23: Cross-validation of ABCD study application by taking splitting the training data into 10 partitions and computing the R^2 and MSE of the prediction results. We average the results of the 10 partitions and display the variation of the predictive performance alongside the mean.

References

Albert, James H and Siddhartha Chib (1993). “Bayesian analysis of binary and polychotomous response data”. In: *Journal of the American statistical Association* 88.422, pp. 669–679.

Alfaro-Almagro, Fidel et al. (2018). “Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank”. In: *Neuroimage* 166, pp. 400–424.

Alfaro-Almagro, Fidel et al. (2021). “Confound modelling in UK Biobank brain imaging”. In: *NeuroImage* 224, p. 117002.

Andersen, Michael Riis, Ole Winther, and Lars Kai Hansen (2014). “Bayesian inference for structured spike and slab priors”. In: *Advances in Neural Information Processing Systems* 2, pp. 1745–1753.

Bagot, K.S. et al. (2018). “Current, future and potential use of mobile and wearable technologies and social media data in the ABCD study to increase understanding of contributors to child health”. In: *Developmental cognitive neuroscience* 32, pp. 121–129.

Bagot, K.S. et al. (2022). “Youth screen use in the ABCD® study”. In: *Developmental Cognitive Neuroscience* 57, p. 101150.

Barbieri, Maria M. and James O. Berger (2004). “Optimal predictive model selection”. In: *Annals of Statistics* 32.3, pp. 870–897.

Barbieri, Maria M. et al. (2021). “The Median Probability Model and Correlated Variables”. In: *Bayesian Analysis* 16.4, pp. 1085–1112.

Barch, Deanna M. et al. (2013). “Function in the human connectome: Task-fMRI and individual differences in behavior”. In: *NeuroImage* 80, pp. 169–189.

Benjamini, Yoav and Yosef Hochberg (1995). “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 57.1, pp. 289–300.

Berlucchi, Giovanni (1983). “Two hemispheres but one brain”. In: *Behavioral and Brain Sciences* 6.1, pp. 171–172.

Besag, Julian (1974). “Spatial interaction and the statistical analysis of lattice systems”. In: *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 36.2, pp. 192–225.

— (1986). “On the statistical analysis of dirty pictures”. In: *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 48.3, pp. 259–279.

Bishop, Christopher M (1998). “Variational learning in graphical models and neural networks”. In: *ICANN 98: Proceedings of the 8th International Conference on Artificial Neural Networks, Skövde, Sweden, 2–4 September 1998 8*. Springer, pp. 13–22.

— (2006). *Pattern Recognition and Machine Learning*. Springer New York.

Blei, David M, Alp Kucukelbir, and Jon D McAuliffe (2017). “Variational inference: A review for statisticians”. In: *Journal of the American statistical Association* 112.518, pp. 859–877.

Bordin, Valentina et al. (2021). “Integrating large-scale neuroimaging research datasets: harmonisation of white matter hyperintensity measurements across Whitehall and UK Biobank datasets”. In: *Neuroimage* 237, pp. 118–189.

Brodoehl, Stefan et al. (2020). “Surface-based analysis increases the specificity of cortical activation patterns and connectivity results”. In: *Scientific Reports* 10.

Brook, D. (1964). “On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems”. In: *Biometrika* 51.3-4.

Brown, M.A. and R.C. Semelka (2011). *MRI: Basic Principles and Applications.* Wiley. URL: `https://books.google.co.uk/books?id=oYOIHi3YkuMC`.

Burt, Joshua B. et al. (2020). “Generative modeling of brain maps with spatial autocorrelation”. In: *NeuroImage* 220, p. 117038.

Buxton, Richard B. et al. (2004). “Modeling the hemodynamic response to brain activation”. In: *NeuroImage* 23, pp. 220–233.

Cai, Qingpo, Jian Kang, and Tianwei Yu (2020). “Bayesian Network Marker Selection via the Thresholded Graph Laplacian Gaussian Prior”. In: *Bayesian Analysis* 15.1, p. 79.

Carbonetto, Peter and Matthew Stephens (2012). “Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies”. In: *Bayesian Analysis* 7.1, pp. 73–108.

Carvalho, Carlos M., Nicholas G. Polson, and James G. Scott (2010). “The Horseshoe estimator for sparse signals”. In: *Biometrika* 97.2, pp. 465–480.

Casey, Betty Jo et al. (2018). “The Adolescent Brain Cognitive Development (ABCD) study: imaging acquisition across 21 sites”. In: *Developmental Cognitive Neuroscience* 32, pp. 43–54.

Cressie, Noel A. (1993). *Statistics for Spatial Data.*

Dahlke, Frank et al. (2021). “Characterisation of MS phenotypes across the age span using a novel data set integrating 34 clinical trials (NO. MS cohort): Age is a key contributor to presentation”. In: *Multiple Sclerosis Journal* 27.13, pp. 2062–2076.

Dale, Anders M, Bruce Fischl, and Martin I Sereno (1999). “Cortical surface-based analysis: I. Segmentation and surface reconstruction”. In: *NeuroImage* 9.2, pp. 179–194.

Datta, Abhirup et al. (2016). “Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets”. In: *Journal of the American Statistical Association* 111.514, pp. 800–812.

Debette, S. and H. S. Markus (2010). “The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis”. In: *BMJ* 341.

Deshpande, Sameer K, Veronika Ročková, and Edward I George (2019). “Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso”. In: *Journal of Computational and Graphical Statistics* 28.4, pp. 921–931.

Dick, Anthony Steven et al. (2021). “Meaningful associations in the adolescent brain cognitive development study”. In: *NeuroImage* 239, p. 118262.

Durante, Daniele and Tommaso Rigon (2019). “Conditionally conjugate mean-field variational Bayes for logistic models”. In: *Statistical Science* 34.3, pp. 472–485.

Feczko, Eric et al. (2021). “Adolescent Brain Cognitive Development (ABCD) community MRI collection and utilities”. In: *bioRxiv.*

Firth, David (1993). “Bias reduction of maximum likelihood estimates”. In: *Biometrika* 80.1, pp. 27–38.

Fischl, Bruce (2012). “FreeSurfer”. In: *NeuroImage* 62.2, pp. 774–781.

Fischl, Bruce, Martin I Sereno, and Anders M Dale (1999a). “Cortical surface-based analysis: II: inflation, flattening, and a surface-based coordinate system”. In: *NeuroImage* 9.2, pp. 195–207.

Fischl, Bruce et al. (1999b). “High-resolution intersubject averaging and a coordinate system for the cortical surface”. In: *Human Brain Mapping* 8.4, pp. 272–284.

Fong, Edwin, Simon Lyddon, and Chris Holmes (2019). “Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap”. In: *International Conference on Machine Learning*. PMLR, pp. 1952–1962.

Friston, K. J. et al. (1996). “Detecting activations in PET and fMRI: levels of inference and power”. In: *NeuroImage* 4, pp. 223–235.

Friston, K.J. et al. (2000). “To smooth or not to smooth?: bias and efficiency in fMRI time-series analysis”. In: *NeuroImage* 12.2, pp. 196–208.

Furrer, Reinhard, Marc G Genton, and Douglas Nychka (2012). “Covariance tapering for interpolation of large spatial datasets”. In: *Journal of Computational and Graphical Statistics* 21.3, pp. 823–824.

Ge, Tian et al. (2014). “Analysis of multiple sclerosis lesions via spatially varying coefficients”. In: *The Annals of Applied Statistics* 8.2.

Gelfand, Alan E. and Penelope Vounatsou (2003). “Proper multivariate conditional autoregressive models for spatial data analysis”. In: *Biostatistics* 4.1, pp. 11–15.

Geman, Stuart and Donald Geman (1984). “Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, pp. 721–741.

George, Edward I and Robert E McCulloch (1993). “Variable selection via Gibbs sampling”. In: *Journal of the American Statistical Association* 88.423, pp. 881–889.

— (1997). “Approaches for Bayesian variable selection”. In: *Statistica Sinica*, pp. 339–373.

Girolami, Mark and Ben Calderhead (2011). “Riemann manifold langevin and hamiltonian monte carlo methods”. In: *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 73.2, pp. 123–214.

Glasser, Matthew F. et al. (2013). “The minimal preprocessing pipelines for the Human Connectome Project”. In: *NeuroImage* 80, pp. 105–124.

Goldsmith, J., L. Huang, and C. Crainiceanu (2014). “Smooth scalar-on-image regression via spatial Bayesian variable selection”. In: *Journal of Computational and Graphical Statistics* 23.1, pp. 46–64.

Griffanti, Ludovica et al. (2016). “BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities”. In: *NeuroImage* 141, pp. 191–205.

Griffanti, Ludovica et al. (2018). “Classification and characterization of periventricular and deep white matter hyperintensities on MRI: A study in older adults”. In: *NeuroImage* 170.

Guo, Cui, Jian Kang, and Timothy D. Johnson (2022). “A spatial Bayesian latent factor model for image-on-image regression”. In: *Biometrics* 78.1, pp. 72–84.

Hagler, Donald J. et al. (2019). “Image processing and analysis methods for the Adolescent Brain Cognitive Development Study”. In: *NeuroImage* 202.

Happ, Clara, Sonja Greven, and Volker J Schmid (2018). “The impact of model assumptions in scalar-on-image regression”. In: *Statistics in Medicine* 37.28, pp. 4298–4317.

Hastings, W. K. (1970). “Monte Carlo sampling methods using Markov chains and their applications”. In: *Biometrika* 57.1, pp. 97–109.

Hazra, Arnab et al. (2019). “A spatio-temporal model for longitudinal image-on-image regression”. In: *Statistics in Biosciences* 11.1, pp. 22–46.

Henson, Richard et al. (1999). “The slice-timing problem in event-related fMRI”. In: *5th International Conference on Functional Mapping of the Human Brain (HBM'99) and Educational Brain Mapping Course, June 22 - 26, 1999, Düsseldorf, Germany.* 9.

Hernández-Lobato, Daniel, J. Miguel Hernández-Lobato, and P. Dupont (2013). “Generalized spike-and-slab priors for Bayesian group feature selection using expectation propagation”. In: *JMLR* 14, pp. 1891–1945.

Higdon, D.M., J. Swall, and J. Kern (1999). “Non-stationary spatial modeling”. In: *Bayesian Statistics 6 – Proceedings of the Sixth Valencia Meeting*, pp. 761–768.

Hoffman, Matthew and David Blei (2015). “Stochastic structured variational inference”. In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics.* Vol. 38. Proceedings of Machine Learning Research. PMLR, pp. 361–369.

Huang, Lei et al. (2013). “Bayesian scalar-on-image regression with application to association between intracranial DTI and cognitive outcomes”. In: *NeuroImage* 83, pp. 210–223.

Jaakkola, Tommi S. and Michael I. Jordan (2000). “Bayesian parameter estimation via variational methods”. In: *Statistics and Computing* 10.1, pp. 25–37.

Jenkinson, Mark and Michael Chappell (2018). *Introduction to neuroimaging analysis.* Oxford University Press.

Johnson, Valen E and David Rossell (2010). “On the use of non-local prior densities in Bayesian hypothesis tests”. In: *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 72.2, pp. 143–170.

Jordan, Michael I. et al. (1999). “Introduction to variational methods for graphical models”. In: *Machine Learning* 37.2, pp. 183–233.

Jousse, Florent et al. (2021). “Geodesic squared exponential kernel for non-rigid shape registration”. In: *16th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2021, Jodhpur, India, December 15-18, 2021.* IEEE, pp. 1–8.

Jung, Rex E. and Richard J. Haier (2007). “The Parieto-Frontal Integration Theory (P-FIT) of intelligence: converging neuroimaging evidence”. In: *The Behavioral and brain sciences* 30.2, pp. 135–154.

Kang, Jian, Brian J. Reich, and Ana Maria Staicu (2018). “Scalar-on-image regression via the soft-thresholded Gaussian process”. In: *Biometrika* 105.1, pp. 165–184.

Karama, Sherif et al. (2011). “Cortical thickness correlates of specific cognitive performance accounted for by the general factor of intelligence in healthy children aged 6 to 18”. In: *NeuroImage* 55.4, pp. 1443–1453.

Karcher, Nicole R and Deanna M Barch (2021). “The ABCD study: understanding the development of risk for mental and physical health outcomes”. In: *Neuropsychopharmacology* 46.1, pp. 131–142.

Katzfuss, Matthias and Joseph Guinness (2021). “A general framework for Vecchia approximations of Gaussian Processes”. In: *Statistical Science* 36.1, pp. 124–141.

Kennedy, James T. et al. (2022). “Reliability and stability challenges in ABCD task fMRI data”. In: *NeuroImage* 252.

Kindalova, Petya, Ioannis Kosmidis, and Thomas E. Nichols (2021). “Voxel-wise and spatial modelling of binary lesion masks: Comparison of methods with a realistic simulation framework”. In: *NeuroImage* 236.

Kosmidis, Ioannis (2021). *brglm2: Bias Reduction in Generalized Linear Models*. R package version 0.8.0. URL: `https://CRAN.R-project.org/package=brglm2`.

Kosmidis, Ioannis, Euloge Clovis Kenne Pagui, and Nicola Sartori (2020). “Mean and median bias reduction in generalized linear models”. In: *Statistics and Computing* 30.1, pp. 43–59.

Li, Fan and Nancy R Zhang (2010). “Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics”. In: *Journal of the American Statistical Association* 105.491, pp. 1202–1214.

Li, Fan et al. (2015). “Spatial Bayesian variable selection and grouping for high-dimensional scalar-on-image regression”. In: *The Annals of Applied Statistics* 9.2, pp. 687–713.

Li, Moyan (2023). “Statistical inference on large-scale and complex data via Gaussian Process”. PhD thesis.

Li, Xinyi, Li Wang, and Huixia Judy Wang (2020). “Sparse learning and structure identification for ultrahigh-dimensional image-on-scalar regression”. In: *Journal of the American Statistical Association* 116.536, pp. 1994–2008.

Liu, Bingyuan et al. (2022). “Robust high-dimensional regression with coefficient thresholding and its application to imaging data analysis”. In: *Journal of the American Statistical Association*.

Luciana, M. et al. (2018). “Adolescent neurocognitive development and impacts of substance use: Overview of the Adolescent Brain Cognitive Development (ABCD) baseline neurocognition battery”. In: *Developmental Cognitive Neuroscience* 32, pp. 67–79.

Lyddon, Simon (2018). “Model misspecification and general Bayesian bootstraps”. PhD thesis. University of Oxford. URL: `https://ora.ox.ac.uk/objects/uuid:2f4dbdc4-0f06-46b1-a4a0-ce8ce71f349e`.

Lyddon, Simon, Stephen Walker, and Chris C. Holmes (2018). “Nonparametric learning from Bayesian models with randomized objective functions”. In: *Advances in Neural Information Processing Systems* 31.

Makowski, Carolina et al. (2023). “Reports of the death of brain-behavior associations have been greatly exaggerated”. In: *bioRxiv*, p. 2023.06.16.545340.

Marcus, Daniel et al. (2011). “Informatics and data mining tools and strategies for the Human Connectome Project”. In: *Frontiers in Neuroinformatics* 5.

Mardia, K. V. (1988). “Multi-dimensional multivariate Gaussian Markov random fields with application to image processing”. In: *Journal of Multivariate Analysis* 24.2, pp. 265–284.

Marek, Scott et al. (2022). “Reproducible brain-wide association studies require thousands of individuals”. In: *Nature 2022 603:7902* 603.7902, pp. 654–660.

Matérn, Bertil (2013). “Spatial Variation: Stochastic models and their application to some problems in forest surveys and other sampling investigations”. PhD thesis.

Mejia, Amanda F et al. (2020). “A Bayesian general linear modeling approach to cortical surface fMRI data analysis”. In: *Journal of the American Statistical Association* 115.530, pp. 501–520.

Menacher, Anna et al. (2023). "Bayesian Lesion Estimation with a Structured Spike-and-Slab Prior". In: *Journal of the American Statistical Association* 0, pp. 1–23. URL: `https://doi.org/10.1080/01621459.2023.2278201`.

Metropolis, Nicholas et al. (1953). "Equation of state calculations by fast computing machines". In: *The Journal of Chemical Physics* 21.6, pp. 1087–1092.

Miller, Alan (2002). *Subset selection in regression.* CRC Press.

Miller, Karla L. et al. (2016). "Multimodal population brain imaging in the UK Biobank prospective epidemiological study". In: *Nature Neuroscience* 19.11, pp. 1523–1536.

Minka, Thomas E. (2013). "Expectation propagation for approximate Bayesian inference". In: *arXiv.*

Mitchell, Toby J and John J Beauchamp (1988). "Bayesian variable selection in linear regression". In: *Journal of the American Statistical Association* 83.404, pp. 1023–1032.

Morris, Emily L., Kevin He, and Jian Kang (2022). "Scalar on network regression via boosting". In: *The Annals of Applied Statistics* 16.4, pp. 2755–2773.

Muetzel, Ryan L. et al. (2015). "White matter integrity and cognitive performance in school-age children: A population-based neuroimaging study". In: *NeuroImage* 119, pp. 119–128.

Narr, Katherine L. et al. (2007). "Relationships between IQ and regional cortical gray matter thickness in healthy adults". In: *Cerebral Cortex* 17.9, pp. 2163–2171.

Nelsen, Roger B. (1999). "An Introduction to Copulas". In: *An Introduction to Copulas.*

Newton, Michael A., Nicholas G. Polson, and Jianeng Xu (2021). "Weighted Bayesian bootstrap for scalable posterior distributions". In: *Canadian Journal of Statistics* 49.2, pp. 421–437.

Newton, Michael A. and Adrian E. Raftery (1994). "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap". In: *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 56.1, pp. 3–48.

Nie, Lizhen and Veronika Ročková (2022). "Bayesian Bootstrap Spike-and-Slab LASSO". In: *Journal of the American Statistical Association.*

Ormerod, John T et al. (2017). "Bayesian hypothesis tests with diffuse priors: Can we have our cake and eat it too?" In: *arXiv preprint arXiv:1710.09146.*

Oxtoby, Neil P. et al. (2019). "ABCD Neurocognitive Prediction Challenge 2019: Predicting individual residual fluid intelligence scores from cortical grey matter morphology". In: *Lecture Notes in Computer Science*, pp. 114–123.

Penke, L. et al. (2012). "Brain white matter tract integrity as a neural foundation for general intelligence". In: *Molecular Psychiatry 2012 17:10* 17.10, pp. 1026–1030.

Pham, Damon D, John Muschelli, and Amanda F Mejia (2022). "ciftiTools: A package for reading, writing, visualizing, and manipulating CIFTI files in R". In: *NeuroImage* 250, p. 118877.

Piironen, Juho and Aki Vehtari (2017). "Sparsity information and regularization in the horseshoe and other shrinkage priors". In: *Electronic Journal of Statistics* 11.2, pp. 5018–5051.

Pol, Hilleke E.Hulshoff et al. (2006). "Genetic contributions to human brain morphology and intelligence". In: *The Journal of Neuroscience* 26.40, pp. 10235–10242.

Poldrack, Russell A., Jeanette A. Mumford, and Thomas E. Nichols (2011). *Handbook of Functional MRI Data Analysis.* Cambridge University Press.

Poline, J. B. and B. M. Mazoyer (1993). "Analysis of individual positron emission tomography activation maps by detection of high signal-to-noise-ratio pixel clusters". In: *Journal of Cerebral Blood Flow and Metabolism* 13 (3), pp. 425–437.

Polson, Nicholas G and James G Scott (2010). "Shrink globally, act locally: Sparse Bayesian regularization and prediction". In: *Bayesian Statistics* 9.501-538, p. 105.

Polson, Nicholas G, James G Scott, and Jesse Windle (2013). "Bayesian inference for logistic models using Pólya–Gamma latent variables". In: *Journal of the American Statistical Association* 108.504, pp. 1339–1349.

Prins, Niels D. and Philip Scheltens (2015). "White matter hyperintensities, cognitive impairment and dementia: an update". In: *Nature Reviews Neurology* 11.3, pp. 157–165.

Reiss, Philip T. and R. Todd Ogden (2010). "Functional generalized linear models with images as predictors". In: *Biometrics* 66.1, pp. 61–69.

Reiss, Philip T. et al. (2015). "Wavelet-domain regression and predictive inference in psychiatric neuroimaging". In: *Annals of Applied Statistics* 9.2, pp. 1076–1101.

Rockafellar, R.T. (1972). *Convex Analysis*. Princeton. URL: `https://books.google.co.uk/books?id=QdUmtAEACAAJ`.

Ročková, Veronika and Edward I George (2014). "EMVS: The EM approach to Bayesian variable selection". In: *Journal of the American Statistical Association* 109.506, pp. 828–846.

— (2018). "The Spike-and-Slab LASSO". In: *Journal of the American Statistical Association* 113.521, pp. 431–444.

Rostrup, E. et al. (2012). "The spatial distribution of age-related white matter changes as a function of vascular risk factors—Results from the LADIS study". In: *NeuroImage* 60.3, pp. 1597–1607.

Roy, Arkaprava et al. (2021). "Spatial shrinkage via the product independent Gaussian process prior". In: *Journal of Computational and Graphical Statistics* 30.4, pp. 1068–1080.

Rubin, Donald B. (1981). "The Bayesian Bootstrap". In: *The Annals of Statistics* 9.1, pp. 130–134.

Sharma, Roopa and Sandeep Sekhon (2021). "White Matter Lesions". In: *StatPearls Publishing*.

Shi, Ran and Jian Kang (2015). "Thresholded Multiscale Gaussian Processes with Application to Bayesian Feature Selection for Massive Neuroimaging Data". In: *arXiv*.

Shin, Minsuk, Anirban Bhattacharya, and Valen E. Johnson (2018). "Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings". In: *Statistica Sinica* 28.2, pp. 1053–1078.

Sladky, Ronald et al. (2011). "Slice-timing effects and their correction in functional MRI". In: *NeuroImage* 58.2, pp. 588–594.

Smola, Alexander Johannes (1998). "Learning with Kernels". PhD thesis.

Spanos, Pol D., Michael Beer, and John Red-Horse (2007). "Karhunen–Loéve expansion of stochastic processes with a modified exponential covariance kernel". In: *Journal of Engineering Mechanics* 133.7, pp. 773–779.

Stein, Michael L. (2014). "Limitations on low rank approximations for covariance matrices of spatial data". In: *Spatial Statistics* 8, pp. 1–19.

Stein, Michael Leonard. (1999). *Interpolation of spatial data : some theory for kriging*. Springer.

Stingo, Francesco C. et al. (2010). "A Bayesian graphical modeling approach to microRNA regulatory network inference". In: *The Annals of Applied Statistics* 4.4, pp. 2024–2048.

Sudlow, Cathie et al. (2015). “UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age”. In: *PLoS medicine* 12.3.

Sweeney, E. M. et al. (2013). “Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal MRI”. In: *American Journal of Neuroradiology* 34.1, pp. 68–73.

Tadesse, Mahlet and Marina Vannucci (2022). *Handbook of Bayesian variable selection.* First edition. Chapman Hall/CRC Press.

Tavor, I. et al. (2016). “Task-free MRI predicts individual differences in brain activity during task performance”. In: *Science* 352.6282, pp. 216–220.

Titsias, Michalis and Miguel Lázaro-Gredilla (2011). “Spike and slab variational inference for multi-task and multiple kernel learning”. In: *Advances in Neural Information Processing Systems* 24.

Uban, Kristina A et al. (2018). “Biospecimens and the ABCD study: Rationale, methods of collection, measurement and early data”. In: *Developmental Cognitive Neuroscience* 32, pp. 97–106.

Van Essen, D. C. et al. (2012). “The Human Connectome Project: A data acquisition perspective”. In: *NeuroImage* 62.4.

Van Essen, David C et al. (2013). “The WU-Minn Human Connectome Project: An overview”. In: *Neuroimage* 80, pp. 62–79.

Vecchia, A. V. (1988). “Estimation and model identification for continuous spatial processes”. In: *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 50.2, pp. 297–312.

Veldsman, Michele et al. (2020). “Spatial distribution and cognitive impact of cerebrovascular risk-related white matter hyperintensities”. In: *NeuroImage: Clinical* 28.

Wainwright, Martin J. and Michael I. Jordan (2008). “Graphical Models, Exponential Families, and Variational Inference”. In: *Foundations and Trends in Machine Learning* 1.1–2, pp. 1–305.

Wardlaw, Joanna M, Maria C Valdés Hernández, and Susana Muñoz-Maniega (2015). “What are white matter hyperintensities made of? Relevance to vascular cognitive impairment”. In: *Journal of the American Heart Association* 4.6, e001140.

Wardlaw, Joanna M et al. (2013). “Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration”. In: *The Lancet Neurology* 12.8, pp. 822–838.

Whiteman, Andrew (2022). “Bayesian Analysis of Neuroimage Data Using Gaussian Process Priors”. In: URL: `https://deepblue.lib.umich.edu/handle/2027.42/174640`.

Whiteman, Andrew S., Timothy D. Johnson, and Jian Kang (2023). “Bayesian inference for group-level cortical surface image-on-scalar-regression with Gaussian process priors”. In: *arXiv.*

Wu, Ben, Ying Guo, and Jian Kang (2022a). “Bayesian Spatial Blind Source Separation via the Thresholded Gaussian Process”. In: *Journal of the American Statistical Association.*

Wu, Luhuan, Geoff Pleiss, and John Cunningham (2022b). “Variational Nearest Neighbor Gaussian Process”. In: *arXiv.*

Wu, Yunan et al. (2022c). “A multicohort geometric deep learning study of age dependent cortical and subcortical morphologic interactions for fluid intelligence prediction”. In: *Scientific reports* 12.1, p. 17760.

Yao, Yuling et al. (2018). “Yes, but Did It Work?: Evaluating Variational Inference”. In: *PMLR*, pp. 5581–5590.

Yu, Chunshui et al. (2008). “White matter tract integrity and intelligence in patients with mental retardation and healthy adults”. In: *NeuroImage* 40.4, pp. 1533–1541.

Yu, Shan et al. (2021). “Multivariate Spline Estimation and Inference for Image-On-Scalar Regression”. In: *Statistica Sinica*, pp. 1463–1487.

Zeng, Zijian, Meng Li, and Marina Vannucci (2022). “Bayesian Image-on-Scalar Regression with a Spatial Global-local Spike-and-slab Prior”. In: *Bayesian Analysis*, pp. 1–26.

Zhang, Daiwei, Tianci Liu, and Jian Kang (2023). “Density regression and uncertainty quantification with Bayesian deep noise neural networks”. In: *Stat* 12.1.

Zhao, Yize, Ben Wu, and Jian Kang (2023). “Bayesian interaction selection model for multimodal neuroimaging data analysis”. In: *Biometrics* 79.2, pp. 655–668.