Bayesian scalar-on-image regression via random image partition models: Automatic identification of regions of interest

Mica Shu Xian Teo (s1459898)

Doctor of Philosophy

School of Informatics & School of Mathematics
University of Edinburgh

March 6, 2023

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Mica Shu Xian Teo (s1459898)*)

Lay Summary

High-dimensional neuroimaging data are routinely collected to investigate different diseases and conditions. For example, Alzheimer's disease (AD) is the most prevalent cause of dementia, which is one of the main causes of death for older people. There is currently no treatment, and a brain tissue autopsy is the only way to get a definitive diagnosis, which can only be made after the patient has passed away. Various data initiavtives are taking place worldwide collecting neuroimaging, biological, and clinical data, with the aim of improving in vivo diagnosis and prediction and monitoring disease progression. Scalar-on-image regression (SIR) provides a formal statistical framework to analyse such data, in order to improve understanding of the underlying condition, the effects on the brain and improve prediction or diagnosis. This thesis builds a novel class of SIR models that employ spatial clustering to automatically extract relevant regions of interest from the image for prediction of the response. As such we provide not only a thorough overview of SIR models but also a detailed comparison of spatial clustering models, including theoretical results, prior simulations and comparisons on image segmentation tasks.

Acknowledgements

Thank you, Lord, for all the blessings you have for me, granting me good health and wisdom to complete my study. I give thanks to my pastor and pastor's wife for the spiritual guidance and good advice they have given to me. And of course, I would like to give a big thank you to those who have followed my journey thus far.

First, I would like to express my deepest appreciation to my supervisor, Dr. Sara Wade for her dedicated support and guidance throughout this research project. Her advice and encouragement have been invaluable throughout my PhD journey. I am also thankful for her support in my academic and industrial endeavours. I would also like to thank Dr. Vanda Inacio De Carvalho for her advice throughout the project.

I am also grateful to my family for all their support during this research. I could not have undertaken this journey without my sister, Qi Wen Teo who never fails to guide, motivate, and inspire me, especially during tough times.

Thanks should also go to my friends, Shenshen Ji, Sung-Lin Yeh, Yuan Chao, Gene Ping, Zeyu Zhao, Lei Tian, Brian Chan, Kendrick Lung, Dexter Loo and Paul Ang who accompany me on this journey. The PhD can be difficult. However, I can survive and enjoy the journey because of these people.

Abstract

Scalar-on-image regression aims to investigate changes in a scalar response of interest based on high-dimensional imaging data. These problems are increasingly prevalent in numerous domains, particularly in biomedical studies. For instance, they aim to utilise medical imaging data to capture and study the complex pattern of changes associated with disease to improve diagnostic accuracy. Due to the massive dimension of the images, which can often be in millions, combined with modest sample sizes, typically in the hundreds in most biomedical studies, pose serious challenges. Specifically, scalar-on-image regression belongs to the "large p, small n" paradigm, and hence, many models utilise shrinkage methods. However, neighbouring pixels in images are highly correlated, making standard regression methods, even with shrinkage, problematic due to multicollinearity and the high number of nonzero coefficients. We propose a novel Bayesian scalar-on-image regression model that utilises spatial coordinates of the pixels to group them with similar effects on the response to have a common coefficient, thus, allowing for automatic identification of regions of interest in the image for predicting the response of interest. In this thesis, we explore two classes of priors for the spatially-dependent partition process, namely, Potts-Gibbs random partition models (Potts-Gibbs) and Ewens-Pitman attraction (EPA) distribution and provide a thorough comparison of the models. In addition, Bayesian shrinkage priors are utilised to identify the covariates and regions that are most relevant for the prediction. The proposed model is illustrated using the simulated data sets and to identify brain regions of interest in Alzheimer's disease.

Keywords: Bayesian; Gibbs-type priors; Potts model; Clustering; Generalised Swendsen-Wang; High-dimensional imaging data

Contents

Table of notation

Notation	Description
n	Number of samples
p	Number of predictors
q	Number of fixed effects
L	Number of components
M	Number of clusters
M_j	Number of clusters up to $j-1$ steps
O	Number of nested clusters
D	Maximal degree
y_i	Outcome measure
$\boldsymbol{w}_i = (w_{i1}, \ldots, w_{iq})^T \in \mathbb{R}^q$	q-dimensional vector of covariates
$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^p$	p-dimensional image predictor
$\boldsymbol{x}_i^* = (x_{i1}^*, \cdots\ x_{im}^*)$	Extracted features from the ith image for the mth region
$\tilde{\boldsymbol{x}}_i$	$(\boldsymbol{w}_i, \boldsymbol{x}_i^*)$
$\boldsymbol{s}_j = (s_{j1}, s_{j2})^T \in \mathbb{R}^2$	Spatial location
$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_q)^T \in \mathbb{R}^q$	q-dimensional fixed effects vector
$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$	Coefficient image
$\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_M^*)^T$	Unique coefficient values
$\tilde{\boldsymbol{\beta}}$	$(\boldsymbol{\mu}, \boldsymbol{\beta}^*)$
$\boldsymbol{\eta}^* = (\eta_1^*, \ldots, \eta_M^*)^T$	Local shrinkage parameters
ϵ_i	Error
σ^2	Noise variance
π_n	Partition of n units
$C_1, \ldots, C_M$	Clusters
$A_1, \ldots, A_O$	Nested clusters
$C_{m,j}$	Subset of indices contained in the mth cluster after j steps
z_i	Cluster label
r_{jk}	Bond variable
$j \sim k$	Represent j and k are neighbours
S_m	Number of connected neighbour pairs in cluster m
$S_{m,j}$	Number of neighbours of j in cluster m
S	Total number of connected neighbour pairs

Notation	Description
(α, δ, τ)	Parameters for the EPA distribution (Concentration, discount, temperature)
$\Psi = (\psi_1, \ldots, \psi_p)$	Permutation
$\mathcal{S}(\cdot)$	Similarity function
$\mathcal{F}(\cdot)$	Decay function
$\boldsymbol{D}$	Distance matrix
$\mathcal{A}(z_1, \cdots z_p)$	Site-wise term for MRF
$\mathcal{B}(z_1, \cdots z_p)$	Interaction term for MRF
υ	Smoothing parameter
ϕ	Parameters related to the Gibbs-type random partition model
α	Parameters for the DP (Concentration)
(α, δ)	Parameters for the PY (Concentration, discount)
(λ, γ)	Parameter for the MFM
ψ	Parameter related to the prior on the number of clusters
Σ_{β^*}	Hyperparameter for $\boldsymbol{\beta}^*$
$(\boldsymbol{m}_\mu, \Sigma_\mu)$	Hyperparameters for $\boldsymbol{\mu}$
(a_σ, b_σ)	Hyperparameters for σ
(a_η, b_η)	Hyperparameters for $\boldsymbol{\eta}^*$
(a_o, b_o)	Hyperparameters for b_η
(ζ_{jk}, κ)	Tuning parameters of the GSW sampler
$x^{(a)}_{(b)}$	Pochhammer symbol with increment b
$x^{(a)}$	Rising factorial
$x_{(a)}$	Falling factorial
$a_p \gtrsim b_p$	Lim sup $a_p/b_p \geq 1$
$\simeq$	Asymptotic expressions

Chapter 1

Introduction

Through advances in data acquisition, vast amounts of high-dimensional imaging data are collected to study phenomena in many fields. Such data are common in biomedical studies to understand a disease or condition of interest (Craddock et al., 2009; Fan et al., 2008; Shi et al., 2014a; Van Walderveen et al., 1998), and in other fields such as psychology (Davatzikos et al., 2005; Sun et al., 2009), social sciences (Ferwerda et al., 2016; Hum et al., 2011; Kim and Kim, 2018; Samany, 2019), economics (Henderson et al., 2009; Naik et al., 2016, 2017), climate sciences (O'Neill, 2013; O'Neill et al., 2013), environmental sciences (Debois et al., 2013; Gundlach-Graham et al., 2015; Maloof et al., 2020) and more. While extracting features from the images based on predefined regions of interest favours interpretation and eases computational and statistical issues, changes may occur in only part of a region or span multiple structures. In order to capture the complex spatial pattern of changes and improve accuracy and understanding of the underlying phenomenon, sophisticated approaches are required that utilise the entire high-dimensional imaging data. However, the massive dimension of the images, which is often in the millions, combined with the relatively small sample size, which at best is usually in the hundreds, poses serious challenges.

In the statistical literature, this is framed as a scalar-on-image regression (SIR) problem (Reiss et al., 2011), so-called as the responses are scalars as in a typical regression, but the covariate is the entire image. SIR belongs to the "large p, small n" paradigm (Bernardo et al., 2003); indeed, the dimension of the image can be massive, particularly for brain images, and thus, many SIR models utilise shrinkage methods that additionally incorporate the spatial information in the image (Goldsmith et al., 2014). In the SIR problem, the covariates represent the image value at a single pixel/voxel, i.e. a very tiny part of the brain, and the effect on the response is most often weak, unreliable and uninterpretable. Moreover, neighbouring pixels/voxels are highly correlated, making standard regression methods, even with shrinkage, problematic due to multicollinearity.

To overcome these difficulties, we develop novel SIR models that group pixels/voxels with similar effects on the response to have a common coefficient within the SIR model, through the use of spatial random partition models. As opposed to the Ising-DP model proposed by Li et al. (2015) (refer to Section 2.1.3 for more details), the proposed model employs spatial random partition models for clustering voxels by utilising the spatial coordinates of the voxels to encourage that groups represent spatially contiguous regions. Importantly, this allows for the automatic identification of regions of interest and integrates regression/classification and identification of regions into a single model-based framework. Thus, the clusters represent brain regions which are inherently defined to be the most discriminative based on the chosen regression/classification model. This not only improves the signal and eases interpretability, but also reduces the computational burden by drastically decreasing the image dimension and addressing the multicollinearity problem.

In particular, the novel models are developed using Bayesian nonparametric (BNP) spatial random partition models; BNP is an exciting and expanding field characterised by flexible models that adapt to the complexity of the model to the data (Gershman and Blei, 2012). Advantages of the proposed BNP approach include allowing the data to determine the number of regions; quantification of uncertainty in the diagnosis and other unknowns, such as the number of regions; and incorporation of prior knowledge from previous studies or expertise of doctors and clinicians. We focus on and provide a thorough comparison of two BNP spatial random partition models: the Ewens-Pitman attraction (EPA) distribution (Dahl et al., 2017) and the Potts-Gibbs random partition (Potts-Gibbs) models, leading to the development of two novel SIR models, namely SIR EPA and SIR Potts-Gibbs models.

Deep learning models such as fully connected neural networks (FNNs) (Amoroso et al., 2018; Zhou et al., 2019), deep polynomial network (DPN) (Shi et al., 2017), convolutional neural networks (CNNs) (Islam and Zhang, 2017; Lin et al., 2018), auto-encoders (Ju et al., 2017), deep belief networks (DBNs) (Shen et al., 2019) and recurrent neural networks (RNNs) (Liu et al., 2018) have been utilised to detect or predict diseases or conditions based on imaging data. Despite remarkable success in medical diagnosis in various applications, deep learning with imaging data still has a number of limitations. Compared to deep learning, our approach provides a number of advantages, namely interpretability (offers various graphics and tools to summarize and interpret results, such as coefficient maps, automatically defined regions of interests, posterior inclusion maps) and uncertainty quantification as well as the possibility to include knowledge from previous studies or expertise of doctors and clinicians. These advantages are particularly important in biomedical settings and other safety-critical applications, where interpretability and well-calibrated uncertainty quantification are crucial.

1.1 Motivating application

Alzheimer's disease (AD) is a damaging brain disease and an increasing burden on society. In 2019, 50 million people worldwide were living with dementia, which is set to reach 152 million by 2050 (International, 2019). The number of people forecasted to develop dementia is increasing at a fast rate worldwide. In the UK, over half a million people have been diagnosed with dementia in 2019 and the female-to-male ratio is 1.67 ([1.52–1.85]) (Nichols et al., 2022). By 2025, the number of individuals with dementia will reach 1 million, and by 2050, it will increase to 2 million. The cost of dementia to governments, social services and individuals has reached staggering figures, with £34.7 billion reported in the UK in 2019 (Wittenberg et al., 2019). According to World Health Organization (WHO), AD is the most prevalent form of dementia. It is estimated that AD accounts for around 60% to 70% of all dementia cases.

Unfortunately, a definite diagnosis of the disease is typically unknown until an autopsy, as it requires histopathologic examination of brain tissue, an invasive procedure. In practice, clinical diagnosis is based on a patient's history and symptoms, behavioural and cognitive tests, and visual examination of neuroimages, if available. Several studies have followed patients to autopsy to estimate the accuracy of a clinical diagnosis; the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) criteria, which are widely used for clinical diagnosis, have an average sensitivity of 81% and an average specificity of 70% for the diagnosis of "probable" AD and an average sensitivity of 93% and an average specificity of 48% for the diagnosis of "possible" AD (Knopman et al., 2001). It is now widely recognised that AD biomarkers based on neuroimaging or biological data can improve diagnosis, particularly in the early stages of the disease when treatments are most likely to be effective. Indeed, new diagnostic criteria based on a revision of NINCDS-ADRDA criteria to include positivity to imaging or fluid biomarkers have been recently proposed for earlier diagnosis of AD (Dubois et al., 2007). Prestia et al. (2015) investigate the revised criteria based on different combinations of well established biomarkers, including biomarkers extracted from amyloid positron emission tomography (Aβ-PET), fluorodeoxyglucose-positron emission tomography (FDG-PET) and structural magnetic resonance images (sMRI). However, many studies investigating the diagnostic accuracy of disease-based neuroimaging data focus on biomarkers from predefined regions of interest (ROIs); this approach has had some successful results, depending on the ROIs used and the severity of the disease for the observed subjects (Convit et al., 2000; Wolf et al., 2001). However, the changes due to the disease or conditions in the brain associated with the disease may occur in only part of the specified brain structure or span multiple structures.

Vast amounts of clinical, biological and neuroimaging data to study AD are being collected through projects such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the UK Biobank. The overall goal of this thesis is to utilise this vast and diverse data to improve diagnosis and understanding of the disease, particularly in the early stages of the disease when a diagnosis is most critical and any proposed drugs or therapies are most likely to be most effective. However, several key challenges need to be addressed for the analysis of neuroimaging data. First, the complex spatial dependence in brain imaging data makes it difficult to comprehend the structure and understand which regions are important. Second, neuroimaging data is inherently noisy and has low signal-to-noise ratios, thus making it harder to accurately detect a difference between groups or to identify subtle changes. The third challenge to highlight is the limited number of subjects. Coupled with the high-dimension of the feature and comparatively small sample sizes, it can be difficult to obtain a reliable analysis and generalise the research findings. We develop scalable Bayesian SIR models for the automatic identification of brain regions to diagnose AD that aim to capture the complex pattern of spatial patterns associated with AD, with the goal of improving diagnostic accuracy, particularly in the early stages, when changes can be very subtle.

1.2 Contributions

In this thesis, we make the following contributions:

1. We develop novel Bayesian scalar-on-image regression models, SIR EPA and SIR Potts-Gibbs models, which exhibit spatial dependence by leveraging the spatial coordinates of the pixels, and demonstrate their application to neuroimaging studies. The SIR model is extended based on a generalized linear model framework to account for different types of outcomes (e.g. continuous, binary, ordinal and counts) and a state-of-the-art inference scheme is developed and implemented based on the generalized Swendsen-Wang (GSW) algorithm, that takes advantages of the spatial coordinates for efficient split-merge moves.

2. We provide a thorough review and comparison of a number of SIR models, followed by a review of dependent random partition models. These form the two main components of our proposed models.

3. We give a detailed discussion of the random image partition models: the EPA distribution and the Potts-Gibbs models, including prior simulations to study the main properties of the partition structure implied by each random image partition model chosen, particularly the number of clusters, the sizes of the clusters, and the number of connected neighbours (reflecting spatial connectivity). In addition, we derive properties, such as the predictive distribution and prior expected number of clusters, of a particular model of interest within the class of Potts-Gibbs models, namely the Potts-mixture of finite mixtures (Potts-MFM).

4. We include a number of simulated experiments to compare the proposed SIR models with existing ones, along with an application for the diagnosis of AD based on hippocampus surface statistics extracted from sMRI.

1.3 Outline of thesis

The organisation of the thesis is as follows. Chapter 2 reviews related literature, including SIR and random image partition models with covariates. Chapter 3 provides a detailed discussion of the random partition image partition models, namely the EPA distribution and the Potts-Gibbs models, which we incorporate in the proposed models. Chapter 4 outlines the development of the proposed models: EPA SIR and Potts-Gibbs SIR models, as well as the proposed inference schemes. Chapter 5 illustrates the proposed models through simulation studies as well as real data from a neuroimaging application. We finish in Chapter 6 with some discussion and thoughts on future research directions.

Throughout the thesis, bold uppercase characters and bold lowercase characters are used to denote matrices and vectors, respectively. On the other hand, lowercase letters represent scalars.

Chapter 2

Literature Review

In this chapter, we lay the groundwork for our proposed models: EPA SIR and Potts-Gibbs SIR models by providing a thorough review of the important processes and distributions that form the basis of our proposed models: SIR and random image partition models. First, we review the SIR models, followed by a review of dependent random partition models.

2.1 Scalar-on-image regression (SIR)

Scalar-on-image regression (SIR) is an example of high-dimensional regression that aims to derive a more complete picture of the association between a high-dimensional imaging predictor and a scalar outcome measure. It has been applied to study issues in, for example, disease diagnosis (Feng et al., 2020; Goldsmith et al., 2014; Huang et al., 2013; Palma et al., 2020; Wang et al., 2017), psychiatry (Reiss et al., 2015), social science (Kang et al., 2018), and more.

Formally, SIR is a statistical linear method used to study and analyse the relationship between a scalar outcome measure and two or three-dimensional predictor images under a single regression model (Goldsmith et al., 2014; Huang et al., 2013; Kang et al., 2018; Li et al., 2015). For each data point, $i = 1, \ldots, n$, we have

$$y_i = \boldsymbol{w}_i^T \boldsymbol{\mu} + \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \overset{i.i.d}{\sim} \mathrm{N}\left(0, \sigma^2\right), \tag{2.1}$$

where y_i is a scalar continuous outcome measure, $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{iq})^T \in \mathbb{R}^q$ is a q-dimensional vector of covariates, and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^p$ is a p-dimensional image predictor. Each x_{ij} indicates the value of the image at a single unit (pixel or voxel) with spatial location $\boldsymbol{s}_j = (s_{j1}, s_{j2})^T \in \mathbb{R}^2$ for $j = 1, \ldots, p$. We define $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_q)^T \in \mathbb{R}^q$ as a q-dimensional fixed effects vector and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ (with $\beta_j := \beta(\boldsymbol{s}_j)$) as the spatially varying coefficient image described on the same lattice as $\boldsymbol{x}_i$. The $\{\epsilon_i\}$ are independent and identically distributed (i.i.d.) errors with mean zero and variance σ^2, representing the variation within the experiment.

In contrast to SIR, the most classical approach to investigate the association between an imaging predictor and outcome of interest is the voxel-wise or mass-univariate regression method (Ashburner and Friston, 2000; Smith et al., 2006), which fits a general linear model (GLM) to imaging data on the individual level and then statistical parametric maps of test statistics and p-values are produced to identify the regions of the coefficient image that are significant. This approach is parallelizable across voxels but unrealistically assumes voxels are mutually independent. While corrections are applied to account for multiple testing, there is no sharing of information across neighbouring voxels.

Alternatively, low-rank models can be employed as statistical models to induce spatial smoothing in a single spatial surface. In spatial statistics, low-rank models approximate the spatial dependence structure by representing the spatial surface in terms of a set of chosen spatial basis functions. For instance, motivated by neuroimaging studies, Reiss and Ogden (2010) develop functional principal components regression (FPCR) which relies on the principal components decomposition for the dimension reduction and uses the penalized spline basis expansions to model the coefficient function for the two-dimensional image predictor. Reiss et al. (2015) propose an FPCR in the wavelet domain by replacing the roughness penalty with the wavelet thresholding method for regression with image predictors, thus the model can be more sensitive to spikes or discontinuities, which is commonly encountered in neuroimaging data. Wang et al. (2014) describe another wavelet-based method as well.

Instead, the SIR models that we focus on are high-dimensional linear regression models and have the advantage of interpretability. Due to the high-dimensional nature and strong spatial correlation of the imaging predictors, shrinkage or penalization methods must be employed. For instance, in the non-Bayesian setting, Wang et al. (2017) introduces a class of generalised SIR models via total variation penalty. Alternatively, Bayesian approaches implicitly induce shrinkage through the choice of a prior distribution and have been shown to dominate frequentist methods, particularly in low-information settings (Celeux et al., 2012). In the Bayesian setting, two classes of priors maybe used: shrinkage priors or spike-and-slab priors. For example in the case of Bayesian SIR based on shrinkage priors, Kang et al. (2018) develop a novel model that imposes shrinkage by modelling the coefficients through a soft-thresholding transformation of latent Gaussian processes (STGP). Most of the literature in Bayesian SIR employs spike-and-slab priors based on Markov random fields (MRFs).

MRFs are commonly used in Bayesian image analysis and spatial statistics to characterise the spatial correlation structure in the data based on a predefined neighbourhood structure and a smoothing parameter. MRFs are spatial processes associated with grid-like structures, where the conditional probability at each node of the grid is defined in terms of the neighbourhood relations, as per the Markov property (Besag, 1986; Li, 2009). This property ensures a simple expression for conditional probabilities as each term depends only on the values in the neighbourhood, making computing the normalising constant of each full conditional straightforward. As a result of the Markov property, this approach is particularly well suited to scalable computations. The Ising model is an example of a binary spatial MRF which is commonly utilised as a prior distribution in imaging data. For example, Smith et al. (2003) and Smith and Fahrmeir (2007) use an Ising prior for the detection of the active region in functional magnetic resonance imaging (fMRI). Huang et al. (2013) develop a hierarchical Bayesian SIR framework which employs an Ising prior distribution as a sparsity-inducing prior to induce both sparsity and spatial smoothness on the coefficient image. Goldsmith et al. (2014) proposes the Ising-Gaussian Markov random field (Ising-GMRF) model adopting two spatial process priors, that is a combination of an Ising and Gaussian Markov random field (GMRF) prior distributions to account for the sparsity and spatial dependence. The Ising-Dirichlet process (Ising-DP) model by Li et al. (2015) uses an Ising prior distribution to spatially smooth latent binary indicator variables but the model does not consider explicitly the spatial variations in the nonzero model coefficients.

More recently, Lee et al. (2021) propose a different approach, a tree-based approach to work in a high-dimensional linear regression context and use a graph to establish relationships between the regression coefficients. In order to deal with the parameters that are assumed to be structured sparse and smooth, they introduce a Bayesian tree-based low-rank Horseshoe (T-LoHo) model, which consists of a spanning tree partition model to model contiguous partitions of graphs and a low-rank multivariate Horseshoe prior to impose sparse homogeneity assumption. They simplify the challenging combinatorial graph partition problem by representing partitions as connected components resulting from edge cuts from a graph's spanning tree.

In the following, we review in detail four SIR models which have different proposals for spike-and-slab or shrinkage priors for the coefficients: the Ising (Huang et al., 2013), Ising-GMRF (Goldsmith et al., 2014), Ising-DP (Li et al., 2015) and STGP (Kang et al., 2018) models. We note that these will form the main competitors in the experiments of Chapter 5.

2.1.1 Ising model

The Ising model (Huang et al., 2013) includes a binary latent indicator image, $\boldsymbol{\gamma}$, to indicate which locations in the coefficient image have an impact on the scalar outcome, where $\gamma(\boldsymbol{s}_j) = 1$ denotes the presence of β_j, and $\gamma(\boldsymbol{s}_j) = 0$ denotes the absence of β_j. The indicator variables are spatially smoothed by employing an Ising prior, a binary MRF, for $\boldsymbol{\gamma}$. This enables a more credible selection of the variables in the regression model, notably in the event of poor information by borrowing strength from neighbouring units. By employing the Ising prior, it triggers the formation of clusters of like-valued neighbouring binary variables (Smith and Fahrmeir, 2007), which leads to spatial smoothing of $\boldsymbol{\gamma}$.

Formally, the Ising prior for $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)^T$ $\big($with $\gamma_j := \gamma(\boldsymbol{s}_j)\big)$ is denoted by:

$$p(\boldsymbol{\gamma}) = c_{(\boldsymbol{a},\boldsymbol{b})} \exp \left\{ \boldsymbol{a}^T \boldsymbol{\gamma} + \sum_j \left(\sum_{k \in \Delta_j} b_{jk} \mathbb{1}_{\gamma_j = \gamma_k} \right) \right\}, \tag{2.2}$$

where $c_{(\boldsymbol{a},\boldsymbol{b})}$ indicates the normalisation constant, $(\boldsymbol{a}, \boldsymbol{b})$ are the parameters of the Ising distribution, Δ_j is the collection of neighbouring locations of location $\boldsymbol{s}_j$, and $\mathbb{1}_{\gamma_j=\gamma_k}$ equals 1 if γ_j and γ_k are the same and 0 otherwise. We concentrate on the first-order neighbours, which will consist of four for most nodes, but less for the nodes at edges and corners. The overall sparsity and neighbour interactions are manipulated by the parameters $\boldsymbol{a}$ and $\boldsymbol{b}$, respectively. While in full generality the parameters of the Ising distribution, $\boldsymbol{a}$ and $\boldsymbol{b}$, may vary over the entire image locations, we follow Huang et al. (2013) and set these parameters to be constants (a, b). Lower a enforces more sparsity whereas higher b encourages more spatial smoothness.

The full Ising model (Huang et al., 2013) is shown below:

$$\begin{aligned} y_i &\sim \mathrm{N}(\boldsymbol{w}_i^T \boldsymbol{\mu} + \boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma_\epsilon^2), \\ \beta_j \mid \gamma_j &\sim \begin{cases} \delta_0 & \text{if } \gamma_j{=}0 \\ \mathrm{N}(0, \sigma_\beta^2) & \text{if } \gamma_j{=}1, \end{cases} \\ \boldsymbol{\gamma} &\sim \mathrm{Ising}(a, b), \end{aligned}$$

where δ_0 denotes a point-mass at zero and σ_β is the variance for $\boldsymbol{\beta}$. In Huang et al. (2013), the parameters $(a, b, \sigma_\beta^2, \sigma_\epsilon^2)$ are model tuning parameters. Huang et al. (2013) propose to use the residual sum of squares for the five-fold cross-validation procedure to obtain the optimal model tuning parameters with extremely short MCMC chains, i.e. 250 or 500 iterations. As previously mentioned a and b determine the level of sparsity and neighbourhood interactions. It is important to note that the selection of Δ_j is also crucial. The choice of the parameter σ_β^2 critically influences the posterior mean and variance of $\boldsymbol{\beta}$. The parameter σ_ϵ^2 has a significant impact on the model predictions and inferred coefficient image and activation probabilities. If the value is too low, the effect of each unit is overemphasized, resulting in overfitting and little sparsity. If the value is too big, the influence of each unit on predicting the response is underestimated, causing regression coefficients to be uniformly zero.

2.1.2 Ising-Gaussian Markov random field (Ising-GMRF) model

In most SIR applications, while a large number of locations will exhibit no significant relationship with the response, there will also be a high proportion of units with non-negligible associations. In this case, the non-zero coefficients should also be spatially smoothed. The Ising model in Section 2.1.1 is extended to the Ising-GMRF model (Goldsmith et al., 2014) by including a prior on the non-zero coefficients that encourage spatial smoothness. The spatial dependence between units is given by way of the GMRF prior for $\boldsymbol{\beta}$ (Rue, 2005). The GMRF prior can result in smoothness, and it is conjugate to the linear regression likelihood.

Mathematically, the GMRF prior for the $\boldsymbol{\beta}$ is:

$$\beta_j | \boldsymbol{\beta}_{-j} \sim \mathrm{N}\left(\bar{\beta}_{\Delta_j}, \frac{\sigma_\beta^2}{N_j}\right), \tag{2.3}$$

where $\boldsymbol{\beta}_{-j}$ denotes the coefficient image with the jth location removed and $\bar{\beta}_{\Delta_j} = 1/N_j \sum_{k\in\Delta_j} \beta_k$ with N_j is the number of elements in Δ_j. The neighborhood structure for the GMRF prior is defined in the same way as for the Ising prior. The Ising-GMRF model is a composite of an Ising prior (Equation (2.2)) for $\boldsymbol{\gamma}$ and a GMRF prior (Equation (2.3)) for the regression coefficients, $\boldsymbol{\beta}$. The Ising prior imposes a comparatively small quantity of non-zero regression coefficients while ensuring that they are spatially grouped. Meanwhile, the GMRF prior aims to ensure that the non-zero coefficients change seamlessly in the space. The goal is to enforce that the signal in $\boldsymbol{\beta}$ is both spatially sparse and smooth in the nonzero units. The full Ising-GMRF model (Goldsmith et al., 2014) is:

$$\begin{aligned} y_i &\sim \mathrm{N}(\boldsymbol{w}_i^T\boldsymbol{\mu} + \boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma_\epsilon^2), \\ \beta_j \mid \boldsymbol{\beta}_{-j}, \gamma_j &\sim \begin{cases} \delta_0 & \text{if } \gamma_j{=}0 \\ \mathrm{N}\left(\bar{\beta}_{\Delta_j}, \frac{\sigma_\beta^2}{N_j}\right) & \text{if } \gamma_j{=}1, \end{cases} \\ \boldsymbol{\gamma} &\sim \mathrm{Ising}(a, b). \end{aligned}$$

Again, the parameters $(a, b, \sigma_\beta^2, \sigma_\epsilon^2)$ are model tuning parameters and crucially affect the model's performance. When employing the model in the context of neuroimaging data, the model's assumptions can be interpreted as presuming that most of the image locations are not significantly meaningful in estimating and predicting the scalar response, while for those in the relevant areas of the brain, the adjacent units (pixels or voxels) must be homogeneous concerning their effects.

2.1.3 Ising-Dirichlet process (Ising-DP) model

In a similar vein to the Ising model (Section 2.1.1) and Ising-GMRF model (Section 2.1.2), the Ising-DP model (Li et al., 2015) employs the Ising component of the prior (Equation (2.2)) for $\boldsymbol{\gamma}$ to smooth the binary selection indicators.

The probability distribution of the non-zero coefficients is unknown and denoted as F. To avoid any possibly restrictive parametric assumptions on the unknown distribution F, a BNP prior is imposed on F, namely, a Dirichlet process (DP) prior, denoted by $F \sim \mathrm{DP}(\alpha, F_0)$, with concentration parameter α and base measure F_0 (Ferguson, 1973, 1974). The DP prior produces discrete realisations, with probability one, and thus induces clustering of the regression coefficients. This pools the great number of non-zero coefficients to a tiny set of values, greatly reducing the effective dimension of the feature space. Moreover, this also helps to improve the strength of the signal, as the unique, cluster-specific coefficient values reflect the cumulative effect of all units within each cluster. It should be noted that the number of active clusters is controlled by the concentration parameter α.

The full Ising-DP model (Li et al., 2015) is:

$$\begin{aligned} y_i &\sim \mathrm{N}(\boldsymbol{w}_i^T\boldsymbol{\mu} + \boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma_\epsilon^2), \\ \beta_j \mid \gamma_j, F &\sim \begin{cases} \delta_0 & \text{if } \gamma_j{=}0 \\ F & \text{if } \gamma_j{=}1, \end{cases} \\ F &\sim \mathrm{DP}(\alpha, F_0), \\ \boldsymbol{\gamma} &\sim \mathrm{Ising}(a, b). \end{aligned}$$

To avoid and minimize sensitivity to the grid search performed in Ising and Ising-GMRF models, Li et al. (2015) assign a hyperprior to the noise variance. Specifically, an improper prior is placed on the unknown noise variance, resulting in an inverse-gamma full conditional distribution for σ_ϵ^2, i.e. $\sigma_\epsilon^2 \sim \mathrm{IG}(n/2, u_{\sigma_\epsilon})$, where $u_{\sigma_\epsilon} = \sum_i y_i - \boldsymbol{w}_i^T\boldsymbol{\mu} - \boldsymbol{x}_i^T\boldsymbol{\beta}$. Following the stick-breaking representation of Sethuraman (1994), it is possible to reformulate the random probability measure F as a weighted sum of infinitely many point masses:

$$\begin{aligned} F(\cdot) &= \sum_{h=1}^{\infty} \omega_h \delta_{\beta_h^*}(\cdot), \\ \beta_h^* &\overset{i.i.d}{\sim} F_0, \\ \omega_h &= \omega_h' \prod_{k<h} (1 - \omega_k'), \\ \omega_h' &\overset{i.i.d}{\sim} \mathrm{Beta}(1, \alpha). \end{aligned} \tag{2.4}$$

Through this representation (Equation (2.4)), we can rewrite the spike-and-slab prior for each β_j as:

$$\beta_j|(\gamma_j, \boldsymbol{\omega}, \boldsymbol{\beta}^*) \sim (1-\gamma_j)\delta_0 + \gamma_j \sum_{h=1}^{\infty} \omega_h \delta_{\beta_h^*}(\cdot),$$

where $\boldsymbol{\omega} = (w_1, \ldots, \omega_h, \ldots)$ and $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_h^*, \ldots)^T$. In practice, a truncated stick-breaking approximation to the DP is considered (Ishwaran and James, 2001), where $F \simeq \sum_{h=1}^{H} \omega_h \delta_{\beta_h^*}$, with $H < \infty$ representing a conservative maximum limit on the number of clusters and $\omega_H' = 1$, such that ω_H is equal to the remaining length of the stick. Additionally, the base measure is $F_0 = \mathrm{N}(0, v^2)$, and (a, b, α, v^2) represent model tuning parameters. For the Ising prior, Li et al. (2015) sets a and b according to their proposed method which yields constraints for a and b in order to remedy the phase transition issue of the Ising prior whereas for the DP prior, Li et al. (2015) chooses $H = 20$ and $v = 10$ so that F_0 is flat over a large domain.

2.1.4 Soft-thresholded Gaussian process (STGP) model

An alternative model is proposed in Kang et al. (2018), where the image coefficient is modelled through a soft-thresholded transformation of latent Gaussian processes, which is referred to as soft-thresholded Gaussian processes (STGP). This prior ensures that the image coefficient is sparse, spatially smooth and continuous. The objective is to establish a smooth transition from the zero to non-zero effects of the adjacent regions. The sites with non-zero coefficients group spatially, and in regions of non-zero coefficients, those values change seamlessly.

The full STGP model (Kang et al., 2018) is shown below:

$$\begin{aligned} y_i &\sim \mathrm{N}(\boldsymbol{w}_i^T\boldsymbol{\mu} + \boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma_\epsilon^2), \\ \boldsymbol{\beta} &\sim \mathrm{STGP}(\varkappa, \mathcal{K}), \end{aligned}$$

where the STGP prior is defined as:

$$\begin{aligned} \beta_j &= g_\varkappa(\tilde{\beta}_j), \\ g_\varkappa(\tilde{\beta}_j) &= \begin{cases} 0 & |\tilde{\beta}_j| \leq \varkappa \\ \mathrm{sgn}(\tilde{\beta}_j)\left(|\tilde{\beta}_j| - \varkappa\right) & |\tilde{\beta}_j| > \varkappa, \end{cases} \\ \tilde{\boldsymbol{\beta}} &\sim \mathrm{GP}(0, \mathcal{K}). \end{aligned}$$

The notation $\tilde{\boldsymbol{\beta}} \sim \mathrm{GP}(0, \mathcal{K})$ with $\tilde{\beta}_j := \tilde{\beta}(\boldsymbol{s}_j)$ denotes the Gaussian process prior on the latent dense image $\tilde{\boldsymbol{\beta}}$, with zero mean function and stationary covariance function $\mathcal{K}$. The function $g_\varkappa$ is the soft-thresholding function with parameter $\varkappa$, where $\mathrm{sgn}(\tilde{\beta}_j)$ takes value 1 if $\tilde{\beta}_j$ is positive and $\mathrm{sgn}(\tilde{\beta}_j)$ takes value -1 if $\tilde{\beta}_j$ is non-positive. The parameter $\varkappa$ governs the prior extent of sparsity.

Higdon et al. (1999) describe the kernel convolution technique for generating both stationary and non-stationary spatial processes. As detailed in Higdon et al. (1999), any stationary Gaussian process $\mathcal{V}(\boldsymbol{s})$ can be defined by the convolution of a kernel function, $K(\cdot)$, that is $\mathcal{V}(\boldsymbol{s}) = \int K(\boldsymbol{s} - \boldsymbol{t})\mathcal{W}(\boldsymbol{t})\, d\boldsymbol{t}$, where $\mathcal{W}$ is a Gaussian white-noise. The covariance function $\mathcal{K}(\cdot)$ is then related to the kernel function $K(\cdot)$ as follows:

$$\mathrm{cov}(\boldsymbol{s}, \boldsymbol{s} + \boldsymbol{h}) = \mathcal{K}(\boldsymbol{h}) = \int K(\boldsymbol{s} - \boldsymbol{t})K(\boldsymbol{s} + \boldsymbol{h} - \boldsymbol{t})\, d\boldsymbol{t}.$$

The kernel representation is then approximated by restricting the process (i.e. the latent Gaussian process, $\tilde{\boldsymbol{\beta}}$) to be a finite grid of locations $\{\boldsymbol{t}_l : l = 1, \cdots, L\}$, which is given by

$$\tilde{\beta}_j = \sum_{l=1}^{L} \mathcal{L}(\boldsymbol{s}_j - \boldsymbol{t}_l)a_l, \tag{2.5}$$

where $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_L \in \mathbb{R}^d$ (for any integer $d \geq 1$ and note that our case is $d = 2$) are a grid of spatial knots with $L \leq p$ and the knots are assumed to be located within a rectangular spatial domain. The $a_l \sim \mathrm{N}(0, \sigma_a^2)$ is the kernel coefficient associated with knot l and $\mathcal{L}$ is a local kernel function. Tapered Gaussian kernels with bandwidth σ_l are used:

$$\mathcal{L}(\boldsymbol{s}_j - \boldsymbol{t}_l) = \exp\left(-\frac{\|\boldsymbol{s}_j - \boldsymbol{t}_l\|_2^2}{2\sigma_l^2}\right) \mathbb{1}_{\|\boldsymbol{s}-\boldsymbol{t}_l\|_2 < 3\sigma_l},$$

where $\|\cdot\|_2$ denotes $\mathbb{L}^2$ norm. Additionally, a conditionally autoregressive (CAR) prior (Gelfand et al., 2010) is imposed on the a_l:

$$a_l | \boldsymbol{a}_{-l} \sim \mathrm{N}\left(\frac{\vartheta}{n_l}\sum_{k\sim l} a_k, \frac{\sigma_a^2}{n_l}\right), \tag{2.6}$$

where $\boldsymbol{a}_{-l}$ denotes the kernel coefficients with the lth element removed, $l \sim k$ indicates knots $\boldsymbol{t}_l$ and $\boldsymbol{t}_k$ are neighbouring on the array, n_l is the sum of knots adjoining to knot l and (ϑ, σ_a^2) are the parameters for the CAR prior. The parameter ϑ, defined on the interval $(0, 1)$, governs the spatial relationships between regions of an image, whereas the σ_a^2 governs the range of the non-zero coefficients.

Combining Equation (2.5) and Equation (2.6), the prior distribution for $\tilde{\boldsymbol{\beta}}$ is defined as:

$$\tilde{\boldsymbol{\beta}} \sim \mathrm{N}\left(0, \sigma_a^2 \boldsymbol{K}(\boldsymbol{M} - \vartheta\boldsymbol{A})^{-1}\boldsymbol{K}^T\right),$$

with the matrices $\boldsymbol{M} = \mathrm{diag}(n_1, \ldots n_L)$, adjacency matrix $\boldsymbol{A}$ with $\boldsymbol{A}_{(k,l)} = 1$ if $k \sim l$ and zero otherwise, kernel matrix $\boldsymbol{K} \in \mathbb{R}^{p\times L}$ with $\boldsymbol{K}_{(j,l)} = \boldsymbol{K}(\boldsymbol{s}_j - \boldsymbol{t}_l)$. In this case, the model tuning parameters include $(\sigma_\epsilon^2, \sigma_a^2, \vartheta, \varkappa)$, controlling the noise variation, range of the nonzero coefficients, spatial dependency, and sparsity, respectively. Kang et al. (2018) use hyperpriors on those model tuning parameters, which are $\sigma_\epsilon^2 \sim \mathrm{IG}(0.1, 0.1), \sigma_a^2 \sim \mathrm{HN}(0, 1), \vartheta \sim \mathrm{Beta}(10, 1)$ and $\varkappa \sim \mathrm{Unif}(\varkappa_l, \varkappa_u)$, where IG(), HN(), Beta() and Unif() denote inverse-gamma, half normal, beta and uniform distributions respectively. Additionally, one must select the number of knot points L and bandwidth σ_l.

2.1.5 Summary

We have provided a review of SIR regression and covered four SIR models in detail: the Ising, Ising-GMRF, Ising-DP and STGP models here. The first three employ spatial spike-and-slab priors for the coefficient image and use an Ising prior to incorporate the spatial information in the spike component and sparsity structure. However, the Ising model does not consider spatial information in the slab component and nonzero coefficients. The Ising-GMRF model is proposed to circumvent the problem by integrating the spatial information in the nonzero units through a GMRF. Both the Ising model and the Ising-GMRF model have a relatively restrictive assumption on the probability distribution of the non-zero coefficients. On the other hand, the Ising-DP model does not impose any restrictive parametric assumptions on the probability distribution of the non-zero coefficients. The Ising-DP model works by combining an Ising prior to incorporate the spatial information in the sparsity structure with a DP prior to group the non-zero coefficients. Still, the spatial information is only incorporated in the sparsity structure and not in the BNP clustering model, which could result in regions that are dispersed throughout the image. Lastly, the STGP model works differently by using soft-thresholded Gaussian processes to induce sparsity and achieves both spatially smooth and continuous in the coefficient image. These four models form the main competitors for our experiments in Chapter 5, and more results and discussions are provided therein.

2.2 Dependent random partition models

Bayesian nonparametrics (BNP) (Ghosal and Van der Vaart, 2017) is an expanding field, characterised by flexible models that are able to recover a wide range of data-generating mechanisms. In fact, BNP models arise from natural assumptions. For example, exchangeability assumes that a sequence of observable random variables is invariant to a permutation or reordering of the indices, and the celebrated de Finetti's Representation Theorem states that a sequence of random variables is exchangeable if and only if

$$\begin{aligned} y_i \mid F &\overset{i.i.d}{\sim} F, \\ F &\sim Q. \end{aligned} \tag{2.7}$$

In this case, Q is called the de Finetti measure and represents the prior over the unknown random probability measure, F. The Dirichlet process (DP) (Ferguson, 1973, 1974) is a cornerstone in BNP and is a popular choice for the prior Q. However, from the stick-breaking representation of DP (Sethuraman, 1994), it is clear that draws are discrete with probability one. Under the stick-breaking representation, it is possible to reformulate the random probability measure F as a weighted sum of infinitely many point masses:

$$\begin{aligned} &F(\cdot) = \sum_{m=1}^{\infty} \omega_m \delta_{\beta_m^*}(\cdot), \\ &\omega_m = \omega_m' \prod_{k<m} (1 - \omega_k'), \qquad \sum_{m=1}^{\infty} \omega_m = 1, \\ &\omega_m' \overset{i.i.d}{\sim} \text{Beta}(1, \alpha), \end{aligned} \tag{2.8}$$

where the β_m^*'s are drawn independently from F_0, which is the prior expectation F and is also called the base measure, and α is called the concentration parameter, controlling the variability around F_0. In this thesis, we assume base measure F_0 is non-atomic.

To overcome the discrete nature, DP mixture models are commonly employed which assume the hierarchical model (Lo, 1984):

$$\begin{aligned} y_i \mid \beta_i &\overset{ind.}{\sim} \mathrm{pr}(y_i \mid \beta_i), \\ \beta_i \mid F &\overset{i.i.d}{\sim} F, \\ F &\sim Q. \end{aligned} \tag{2.9}$$

The discrete nature of F implies the existence of ties among $\beta_1, \beta_2 \ldots$ with positive probability. For any finite n, we can reparameterise $\beta_1, \ldots, \beta_n$ in terms of the unique values $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_M^*)^T$ with $M \leq n$ and a partition $\pi_n = \{C_1, \ldots, C_M\}$, a set denoting a partition of n units into M nonempty, mutually exclusive, and exhaustive clusters $C_1, \ldots, C_M$ such that $\cup_{C \in \pi_n} C = \{1, \ldots n\}$. Each unit has its own cluster label $z_i \in \{1, \ldots, M\}$, i.e $z_i = m$ if and only if $i \in C_m$. The model (2.9) can be marginalized by integrating out F and shown to be of the form:

$$\begin{aligned} y_i \mid z_i = m, \beta_m^* &\overset{ind.}{\sim} \mathrm{pr}(y_i \mid \beta_m^*), \\ \beta_m^* \mid F_0 &\overset{i.i.d}{\sim} F_0, \\ \pi_n &\sim \mathrm{pr}(\pi_n). \end{aligned} \tag{2.10}$$

In the case of the DP, the random partition model is given by

$$\mathrm{pr}(\pi_p) = \frac{\alpha^M}{\alpha^{(p)}} \prod_{m=1}^{M} \Gamma(\mid C_m \mid), \tag{2.11}$$

where the notation $x^{(a)} = x(x+1)\ldots(x+a-1)$ denotes the rising factorial. However, other priors beyond the DP (Lijoi and Prünster, 2009) may be employed leading to different random partition models. For instance, the DP is related to the Ewens distribution whereas the Pitman Yor process (PY) is associated with the Ewens-Pitman distribution. If we choose Q as DP in Equation (2.9), the partition distribution $\mathrm{pr}(\pi_p)$ is the Ewens distribution whereas if we choose Q as PY in Equation (2.9), the partition distribution $\mathrm{pr}(\pi_p)$ becomes the Ewens-Pitman distribution, defined as

$$\mathrm{pr}(\pi_p) = (\delta + \alpha)_{(\delta)}^{(M-1)} \prod_{m=1}^{M} \frac{\Gamma(\mid C_m \mid - \delta)}{\Gamma(1-\delta)}, \tag{2.12}$$

where $\delta \in (0,1)$ denotes the discount parameter and $\alpha > -\delta$ denotes the concentration parameter as for DP. The notation $x_{(b)}^{(a)} = x(x+b)\ldots(x+(a-1)b)$ represents the Pochhammer symbol with increment b.

The development of BNP models that incorporate general covariate information is a growing research area in the literature to accommodate the growing complexity of data. Many proposals in the direction focus on the stick-breaking representation in Equation (2.8) and extending the weights or atoms to depend on covariates, examples include the dependent Dirichlet processes (DDP) (MacEachern, 1999), order-based dependent Dirichlet processes (Griffin and Steel, 2006) and weighted mixtures of Dirichlet processes (Dunson et al., 2007)), to name a few (see also Quintana et al. (2022) for a review of other dependent Dirichlet processes). Instead, in the following, we focus on proposals that extend the random partition models in Equation (2.10). We relax the exchangeability assumption in the prior for the partitions and mainly focus on and review the case when the covariate represents spatial coordinates. Specifically, our review includes the product partition model (PPM) with covariates (PPMx) (Müller et al., 2011), spatial PPM (sPPM) (Page and Quintana, 2016), distance-dependent Chinese restaurant process (CRP) (ddCRP) (Blei and Frazier, 2011; Ghosh et al., 2011), spatial ddCRP (Ghosh et al., 2011) and restricted CRP (rCRP) (Wehrhahn et al., 2020). In Chapter 3, we provide a thorough discussion and overview of two other classes of spatial random partition models, namely the Ewens-Pitman attraction (EPA) distribution and Potts-Gibbs random partition (Potts-Gibbs) models which will form one of the major building blocks of the novel models proposed in this thesis.

2.2.1 Product partition model with covariates (PPMx)

The formulation of models starting from a partition distribution has been a fruitful approach, exemplified by the development of product partition models (PPM). The PPM by Barry and Hartigan (1993), construct the random partition π_p by utilising cohesion functions that depend on only one cluster at a time. The probability of the partition is given by the product of cluster-dependent functions as shown below:

$$\mathrm{pr}(\pi_p) = \prod_{m=1}^{M} \mathcal{C}(C_m),$$

for some cohesion functions $\mathcal{C}(C_m)$ that measure how tightly the units in cluster C_m are clustered. A popular option for the cohesion function is $\mathcal{C}(C_m) \propto (\mid C_m \mid -1)!$, which equivalently results in the random partition model implied by the DP prior. It is widely known that the corresponding posterior distribution of $\mathrm{pr}(\pi_p)$ is again in the same form.

To allocate the units into clusters that are more homogeneous in covariates, denoted by $\boldsymbol{w}_i$ for $i, \cdots, n$, Müller et al. (2011) introduce an additional factor in the traditional PPM to achieve the desired clustering which is known as the covariate-dependent PPM (PPMx). The PPMx is of the form:

$$\mathrm{pr}(\pi_p) \propto \prod_{m=1}^{M} \mathcal{F}(\boldsymbol{w}_m^*)\mathcal{C}(C_m),$$

where $\boldsymbol{w}_m^* = (\boldsymbol{w}_j,\ j \in C_m)$ denotes covariates in cluster C_m and $\mathcal{F}(\cdot)$ a non-negative similarity function. Draws from the PPMx result in partitions with clusters that have more similar covariates values.

2.2.2 Spatial product partition model (sPPM)

From a spatial modelling perspective, Page and Quintana (2016) propose the spatial PPM (sPPM), which extends PPMs and places a prior on the partitions according to the spatial locations of units by making the cohesion function depend on the spatial location:

$$\mathrm{pr}(\pi_p) \propto \prod_{m=1}^{M} \mathcal{C}(C_m, \boldsymbol{s}_m^*),$$

where $\boldsymbol{s}_m^* = (\boldsymbol{s}_j,\ j \in C_m)$ denotes the locations in cluster C_m. The cohesion function of the sPPM provides the flexibility to incorporate and consider spatial information in the prior on partitions. Page and Quintana (2016) provide advice on the cohesion function, resulting in various types of spatial structures. For instance, the cohesion function

$$\mathcal{C}(C, \boldsymbol{s}_m^*) = \begin{cases} \alpha \times \Gamma(\mid C \mid) & \text{if } C \text{ is spatially connected} \\ 0 & \text{otherwise,} \end{cases}$$

may be used to encourage a small number of large clusters, with α controlling the number of clusters. However, a cohesion function defined in this way only assigns prior mass to partitions that are spatially connected, which is intuitive but computationally challenging to implement for a large number of locations. Instead, the authors suggest four reasonable cohesion functions that can be used. These cohesion functions partition the locations into disjoint dependence neighborhoods based on the defined spatial setting.

Regarding both the PPMx and sPPM, it is pointed out by Page and Quintana (2018) that as the number of covariates grows (but not necessarily the number of observations), their influence on clustering tends to overwhelm information from the response and has been shown to have a significant impact on the resulting clustering structure. It often leads to a high posterior probability of getting partitions with either a large number of singleton clusters or one large cluster. As expected, this negatively impacts inference and predictions.

2.2.3 Distance-dependent Chinese restaurant process (ddCRP) and spatial ddCRP

The popular Chinese restaurant process (CRP) is a distribution over infinite partitions of positive integers. The assignment of customers to tables is an intuitive metaphor to illustrate the CRP, which defines a random partition: customers enter a restaurant sequentially, and the customer (unit) picks an existing table (cluster) with probability proportional to the number of customers sitting there, or a new table with probability proportional to the concentration parameter, α. Let $z_j \in \{1, \cdots, M\}$ be the cluster label of unit j and $\boldsymbol{z}_{1:j-1}$ represents the vector of cluster labels of units $1, \ldots, j-1$ which were previously assigned, then, mathematically,

$$\mathrm{pr}(z_j = m|\boldsymbol{z}_{1:j-1}, \alpha) = \begin{cases} \frac{|C_{m,1:j-1}|}{\alpha+j-1} & m \leq M_{1:j-1} \\ \frac{\alpha}{\alpha+j-1} & m = M_{1:j-1}+1, \end{cases}$$

where $|C_{m,1:j-1}|$ denotes the number of units in cluster m before unit j (i.e. based only units $1, \cdots, j-1$) and $M_{1:j-1}$ denotes the number of unique clusters before unit j. The CRP defines a sequence of predictive distributions for the allocation of units into groups.

The CRP yields an exchangeable partition distribution, i.e. the distribution of cluster structure is invariant of the order of allocation of units, and is called the Ewens distribution. The pmf for the Ewens distribution, $\mathrm{pr}(\pi_p)$ exists and in fact is equivalent to the random partition model induced by the DP in Equation (2.11). A potential downside of using the CRP is the so-called "rich-get-richer" property, that is large clusters tend to get larger, while small clusters stay small.

Blei and Frazier (2011) generalise the conventional CRP to provide a flexible framework for clustering by taking into account the temporal, spatial and other structured dependencies between the units. They propose the distance-dependent CRP (ddCRP) with a modification on the CRP to incorporate pairwise distances based on covariate information via a distance-dependent decay function, $\mathcal{F}$. For each individual, $I_j \in \{1, \cdots, p\}$ denotes the index of the individual that customer j decides to sit with. Let $\boldsymbol{D}$ be the distance matrix with d_{jk} denoting the distance measurement between unit j and k and the conditionals are given as follows:

$$\mathrm{pr}(I_j = k \mid \boldsymbol{D}, \alpha) \propto \begin{cases} \mathcal{F}(d_{jk}) & j \neq k \\ \alpha & j = k. \end{cases}$$

Now the metaphor describes how each customer (unit) chooses with whom they prefer to sit, with probability proportional to how close they are. The ddCRP corresponds to the customer to table assignments ($z_j \in \{1, \cdots, M\}$) which is implicitly induced from the customer to customer assignments ($I_j \in \{1, \cdots, p\}$). However, certain properties of the CRP are lost after taking into account the distance information. This process relaxes the exchangeable property of conventional CRP. Furthermore, we can no longer express its pmf over partitions explicitly. The pmf can only be computed implicitly by summing up all possible assignments that map to a particular partition, consequently, when the ddCRP is used, MCMC algorithms like Metropolis-Hastings cannot be implemented to estimate the posterior distribution. Nevertheless, it is pointed out by Blei and Frazier (2011) that efficient approximate inference with ddCRP is possible with Gibbs sampling.

In the spatial setting, the ddCRP was restricted to enforce homogeneous regions through appropriately defined distances and used to study image segmentation for computer vision (Ghosh et al., 2011) and to model geometric variability in spinal images (Seiler et al., 2013). The spatial ddCRP incorporates the spatial distance between units in the clustering process; thus, it is more biased towards creating spatially contiguous clusters.

2.2.4 Restricted Chinese restaurant process (rCRP)

The restricted CRP, a spatially restricted prior distribution for random partitions has been recently introduced by Wehrhahn et al. (2020) with the purpose of identifying disease clusters in areal units that are most at risk compared with neighbouring areas. The key feature of the restricted CRP is enforcing every cluster to be fully connected. It is pointed out by Page and Quintana (2016) that computational challenges arise in PPMs with a restricted cohesion function to allocate non-zero probability only to those cluster configurations which results in spatially connected clusters. The restricted CRP resolves these difficulties. The process constrains clusters to be made of the adjacent areal unit by using an adjacency matrix $\boldsymbol{A}$, in which each component $a_{jk} = 1$ if regions j and k share a common boundary, otherwise 0. The corresponding full conditionals are given as follows:

$$\operatorname{pr}(z_j = m \mid z_{1:j-1}, \alpha, \boldsymbol{A}) \propto \begin{cases} |C_{m,1:j-1}| & m \leq M_{1:j-1} \text{ and } \mathcal{Q}(z_{1:j}, \boldsymbol{A}) = 1 \\ \alpha & m = M_{1:j-1} + 1 \text{ and } \mathcal{Q}(z_{1:j}, \boldsymbol{A}) = 1 \\ 0 & \mathcal{Q}(z_{1:j}, \boldsymbol{A}) = 0, \end{cases}$$

where $\mathcal{Q}(\cdot)$ is a function which is equal to 1 whenever $z_{1:j}$ is an admissible cluster configuration under $\boldsymbol{A}$.

2.2.5 Summary

We have discussed a few variations of product partition models (PPM) and Chinese restaurant processes (CRP) which take into account relevant covariate information in the model. The covariate-dependent PPM (PPMx) incorporates another similarity measure to quantify how alike the units are. The spatial PPM (sPPM) works by putting the location of the units into the cohesion function to handle and enforce spatially contiguous regions in the random partition. For the CRP, we consider distance-dependent CRP (ddCRP), spatial ddCRP, and restricted CRP (rCRP). The ddCRP generalises the traditional CRP with a distance-dependent decay function to handle the dependence structure in the data. The spatial ddCRP modifies ddCRP slightly to focus the attention more on the information about spatial distance. Lastly, the development of rCRP is motivated by ensuring partitions consist of clusters that are spatially connected. While these proposals all provide interesting constructions to incorporate spatial dependence in the random partition model, in the remaining chapters we instead focus on two alternative classes of random partition models. First, we consider the Ewens-Pitman attraction distribution, motivated by its analytical expressions, allowing for example posterior inference of key hyperparameters. Second, we focus on the class of the Potts-Gibbs random partition models motivated by their scalability.

Chapter 3

Random Image Partition Models

In this chapter, we describe in detail the Ewens-Pitman attraction (EPA, Section 3.2) distribution (Dahl et al., 2017) and the Potts-Gibbs random partition (Potts-Gibbs, Section 3.3) models and lay out the reasons why we choose them instead of the dependent random partition models reviewed in Section 2.2. For the Potts-Gibbs models, we cover the Potts-Dirichlet process random partition (Potts-DP, Section 3.3.4) model, Potts-Pitman Yor process random partition (Potts-PY, Section 3.3.5) model and Potts-mixture of finite mixtures random partition (Potts-MFM, Section 3.3.6) model. We study the main properties of the implied partition structure by each random image partition model chosen, particularly the number of clusters, the size of the clusters and the number of connected neighbours. The predictive distribution of the Potts-MFM model is formulated, and the prior expected number of clusters for the Potts-MFM model is also derived. The proofs are provided. Moreover, we present in-depth comparisons between each random image partition model via the prior simulations.

Note: *We are planning to submit a review paper on Bayesian spatial clustering providing comparisons among different Bayesian spatial clustering methods (Potts-DP, Potts-PY, Potts-MFM, EPA distribution, sPPM and ddCRP) to Statistical Science.*

3.1 Introduction

In clustering applications, the assumption of exchangeability may not be suitable in some settings such as spatial data and time series data. It is appealing to combine the information underlying data such as the proximity of units to influence the partition structure, which leads to a feature-dependent partition. Random spatial partition models have been widely applied to application domains such as image segmentation (Ghosh et al., 2011), disease mapping (Denison and Holmes, 2001; Wehrhahn et al., 2020), traffic modelling (Durand et al., 2021), spatial crime modelling (Balocchi and Jensen, 2019); and others.

Exchangeability is indeed no longer an appropriate assumption in spatial clustering since covariate information is included in the spatial data or images, which we want to use to enhance model performance and interpretability. For our proposed SIR framework, we also assume certain structural assumptions on the spatial partition. In particular, we incorporate spatial random image partitions within a SIR framework to address limitations of existing SIR models which we fully detail in Chapter 4.

The remainder of this chapter is organised as follows. First, in Section 3.2, we present the EPA distribution and discuss its main properties. In Section 3.3, a detailed description of the general Potts-Gibbs models is given along with full details on three models of interest within the Potts-Gibbs framework, namely, the Potts-DP, Potts-PY and Potts-MFM models. Novel properties are presented for the Potts-MFM model including the predictive distribution and the prior expected number of clusters. In Section 3.4, we provide a discussion on why we focus on the EPA distribution and the Potts-Gibbs models instead of the dependent random partition models reviewed in Section 2.2. In Section 3.5, we present prior simulations to study and compare the results under the prediction rule from the EPA distribution and the Potts-Gibbs models. We finish in Section 3.6 with an image segmentation study to show the clustering performance of the EPA distribution and the Potts-Gibbs models. Conclusions are provided in Section 3.7.

Before delving into the random image partition models in this chapter, we first layout some general notation used throughout this thesis. The notation $x^{(a)}_{(b)} = x(x+b)\dots(x+(a-1)b)$ represents the Pochhammer symbol with increment b, where $x^{(a)}_{(0)} = x^a$ and $x^{(a)}_{(1)} := x^{(a)}$. The notation $x^{(a)} = x(x+1)\dots(x+a-1)$ denotes the rising factorial while the notation $x_{(a)} = x(x-1)\dots(x-a+1)$ denotes falling factorial. Note that the expressions $x^{(a)}$ and $x_{(a)}$ may be rewritten in terms of gamma function, i.e. $x^{(a)} = \Gamma(x+a)/\Gamma(x)$ and $x_{(a)} = \Gamma(x+1)/\Gamma(x-a+1)$, where $\Gamma(x)$ stands for the gamma function. The notation $a_p \gtrsim b_p$ indicates lim sup $a_p/b_p \geq 1$. The notation $j \sim k$ means that j and k are neighbours. Let $S_m = \sum_{j\sim k} \mathbb{1}_{z_j=z_k=m}$ be the number of connected neighbour pairs in cluster m, $S_{m,j} = \sum_{k:j\sim k} \mathbb{1}_{z_k=m}$ be the number of neighbours of j in cluster m and $S = \sum_{m=1}^{M} S_m$ be the total number of connected neighbour pairs.

3.2 Ewens-Pitman attraction (EPA) distribution

Definition 2.1: Ewens-Pitman attraction distribution

The pmf for a partition π_p associated with the Ewens-Pitman attraction (EPA) distribution is constructed sequentially from the product of conditional probabilities:

$$\mathrm{pr}(\pi_p|\alpha,\delta,\mathcal{S},\Psi) = \prod_{j=1}^{p} \mathrm{pr}_j\left\{\alpha,\delta,\mathcal{S},\pi(\psi_1,\ldots,\psi_{j-1})\right\}, \tag{3.1}$$

with $\mathrm{pr}_j\{\alpha,\delta,\mathcal{S},\pi\left(\psi_1,\ldots,\psi_{j-1}\right)\} = 1$ for $j = 1$ and is otherwise defined as

$$\begin{aligned}&\mathrm{pr}_j\{\alpha,\delta,\mathcal{S},\pi(\psi_1,\ldots,\psi_{j-1})\}\\ &= \mathrm{pr}\{\psi_j \in C_{m,j}|\alpha,\delta,\mathcal{S},\pi(\psi_1,\ldots,\psi_{j-1})\}\\ &= \begin{cases} \frac{j-1-\delta M_{j-1}}{\alpha+j-1}\cdot\frac{\sum_{\psi_k\in C_{m,j}}\mathcal{S}(\psi_j,\psi_k)}{\sum_{k=1}^{j-1}\mathcal{S}(\psi_j,\psi_k)} & \text{for } C_{m,j}\in\pi\left(\psi_1,\ldots,\psi_{j-1}\right)\\ \frac{\alpha+\delta M_{j-1}}{\alpha+j-1} & \text{for } C_{m,j} \text{ being a new cluster,}\end{cases}\end{aligned} \tag{3.2}$$

where $\Psi = (\psi_1,\ldots,\psi_p)$ denotes the permutation to indicate the sequence in which the p units are allocated, $\mathcal{S}(\cdot)$ represents the similarity function with $\mathcal{S}(\psi_j,\psi_k)$ denoting the similarity of units ψ_j and ψ_k, $\pi(\psi_1,\ldots,\psi_{j-1})$ is the partition of $\{\psi_1,\ldots,\psi_{j-1}\}$ up to $j-1$ steps with M_{j-1} number of unique clusters up to $j-1$ steps and $C_{m,j}$ indicates the subset of indices contained in the mth cluster after j steps. Without loss of generality, we drop the subscript j from $C_{m,j}$ and M_j when clear from the context. The α and δ represent the concentration and discount parameters respectively.

Dahl et al. (2017) presents another type of random partition model, the Ewens-Pitman attraction (EPA) distribution which incorporates covariate information via the pairwise distance. The EPA distribution sequentially assigns the units to subsets according to the attractiveness between the units, which are based on the pairwise distances. One important feature of the EPA distribution is the existence of the closed-form pmf (Definition 2.1).

The EPA distribution is parameterized by a concentration parameter α, a discount parameter δ, and a similarity function $\mathcal{S}(\cdot)$. Both α and δ play a pivotal role in inferring the distribution of the number of clusters and the cluster sizes. The EPA distribution incorporates covariate information via pairwise distances and the similarity function. The similarity function $\mathcal{S}(\cdot)$ is defined based on a decay function $\mathcal{F}(\cdot)$ and a distance matrix $\boldsymbol{D}$ with elements d_{jk} denoting the distance measure between observations j and k. The decay function $\mathcal{F}(\cdot)$ is non-increasing and the similarity function $\mathcal{S}(\cdot)$ is specifically defined as:

$$\mathcal{S}(j,k) = \mathcal{F}(d_{jk}).$$

In the case of spatial covariate information, the pairwise distances d_{jk} represent the distance between the spatial coordinates of $\boldsymbol{s}_j$ and $\boldsymbol{s}_k$. The decay function, $\mathcal{F}(\cdot)$, controls how distances influence the distribution over the clustering structure. One of the possible ways to measure the homogeneity of j and k is an exponential decay function, $\mathcal{F}(d_{jk}) = \exp\left(-\tau d_{jk}\right)$, with τ acting as a temperature parameter of the function, e.g. for the Euclidean pairwise distances:

$$\mathcal{F}(d_{jk}) = \exp\left\{-\tau\sqrt{(s_{j1}-s_{k1})^2+(s_{j2}-s_{k2})^2}\right\}. \tag{3.3}$$

The function produces larger values for units that are similar. Dahl et al. (2017) use the same τ for the Euclidean distance, however, we can potentially use different τ in each spatial direction, i.e. τ_1 and τ_2 to encourage apriori more flexible elliptically-shaped spatial clusters. In addition to location-specific information, there may be other information that we can include in the pairwise distance.

The number of clusters

The EPA distribution allocates units based on their attraction to existing clusters, where the attraction to a particular cluster is a function of the pairwise similarities between the current unit and the units in that cluster. There is an existence of the closed-form distribution of the number of clusters (and their moments). It is invariant to the similarity information, which allows for standard MCMC algorithms to be easily applied for posterior inference on the partition π_p and any parameters that influence the partition, e.g. concentration parameter α and discount parameter δ.

Dahl et al. (2017) proved that the distribution of the number of clusters is unchanged from the usual Ewens and Ewens-Pitman distributions, and one's intuition about the concentration parameter α and discount parameter δ from these familiar distributions carry over. As highlighted in previous Section 2.2, there is a connection between Ewens and DP (Equation (2.11)), as well as Ewens-Pitman and PY (Equation (2.12)).

Proposition 2.1: Number of clusters (Buntine and Hutter, 2010)

A random partition induced by a Pitman–Yor process (PY) which is characterised by a concentration parameter α and discount parameter δ, exhibits the expected prior number of clusters, $\mathbb{E}[M]$ for a sample size of p:

$$\begin{aligned}\mathbb{E}[M] &= \frac{\alpha}{\delta}\frac{(\alpha+\delta)^{(p)}}{\alpha^{(p)}} - \frac{\alpha}{\delta}\\ &\simeq \frac{\alpha}{\delta}\left(1+\frac{p}{\alpha}\right)^{\delta}\exp\left\{\frac{\delta p}{2\alpha(\alpha+p)}\right\} - \frac{\alpha}{\delta}, \qquad for p, \alpha \gg \delta.\end{aligned} \tag{3.4}$$

And the prior variance of the number of clusters, $\mathrm{Var}[M]$ for a sample size of p, when $\delta > 0$ can be written as:

$$\begin{aligned}\text{Var}[M] &= \frac{\alpha(\delta+\alpha)}{\delta^2}\frac{(\alpha+2\delta)^{(p)}}{\alpha^{(p)}} - \frac{\alpha}{\delta}\frac{(\alpha+\delta)^{(p)}}{\alpha^{(p)}} - \left\{\frac{\alpha}{\delta}\frac{(\alpha+\delta)^{(p)}}{\alpha^{(p)}}\right\}^2 \\ &\simeq \frac{\alpha}{\delta}\left(1+\frac{p}{\alpha}\right)^{2\delta}\exp\left\{\frac{\delta p}{\alpha(\alpha+p)}\right\}, \qquad forp, \alpha \gg \delta.\end{aligned} \tag{3.5}$$

Note that the approximations in Equation (3.4) and Equation (3.5), represented by the symbol $\simeq$, are asymptotic expressions. They hold when the sample size, p, and the concentration parameter, α, are much larger than the discount parameter, δ.

The Dirichlet process (DP) is a special case where discount $\delta = 0$:

$$\begin{aligned}\mathbb{E}[M] &= \alpha\left\{\psi_0(\alpha+p) - \psi_0(\alpha)\right\} \\ &\simeq \alpha\log\left(1+\frac{p}{\alpha}\right), \qquad \text{for } p, \alpha \gg 0, \\ \text{Var}[M] &= \alpha\left\{\psi_0(\alpha+p) - \psi_0(\alpha)\right\} + \alpha^2\left\{\psi_1(\alpha+p) - \psi_1(\alpha)\right\} \\ &\simeq \alpha\log\left(1+\frac{p}{\alpha}\right), \qquad \text{for } p, \alpha \gg 0,\end{aligned} \tag{3.6}$$

where $\psi_0(\cdot)$ represents the digamma function and $\psi_1(\cdot)$ represents the first derivative of the digamma function, also known as the polygamma function of order one.

For the PY, as seen in Equation (3.4)–(3.5) in Proposition 2.1, if $\delta > 0$ and $p \gg \alpha \gg \delta$, then the prior standard deviation is approximately $\mathbb{E}[M]/\sqrt{\alpha/\delta}$, which is smaller than the expected number of clusters, $\mathbb{E}[M]$. For the DP case, the prior standard deviation is approximately the square root of the expected number of clusters, $\sqrt{(}\mathbb{E}[M])$. In both cases, the expected number of clusters, $\mathbb{E}[M]$ is roughly linear in concentration parameter α. The information is useful to set bounds for the hyperparameters, which can be used to set the hyperprior or for a grid search to find optimised values of each hyperparameter.

3.3 Potts-Gibbs models

The class of Potts-Gibbs random partition (Potts-Gibbs) models provides another framework to include spatial information within the random partition model. As the name suggests, the Potts-Gibbs class combines two components: the Gibbs-type random partition model and the Potts model. The theoretical properties of the Potts-Gibbs models are examined, including both components.

	DP	PY	MFM
ϕ	(α)	(α, δ)	(ψ, γ)
$V_p(M)$	$\frac{\Gamma(\alpha)\alpha^M}{\Gamma(\alpha+p)}$	$\frac{\Gamma(\alpha+1)\prod_{m=1}^{M-1}(\alpha+m\delta)}{\Gamma(\alpha+p)}$	$\sum_{l=1}^{\infty} \frac{\Gamma(\gamma l)l!}{\Gamma(\gamma l+p)(l-m)!} p_L(l \mid \psi)$
$W_{\lvert C_m \rvert}(\phi)$	$\Gamma(\lvert C_m \rvert)$	$\frac{\Gamma(\lvert C_m \rvert-\delta)}{\Gamma(1-\delta)}$	$\frac{\Gamma(\lvert C_m \rvert+\gamma)}{\Gamma(\gamma)}$

Table 3.1: Formulas of $V_p(M)$, $W_{|C_m|}(\phi)$ and parameters ϕ for DP, PY and MFM.

3.3.1 Gibbs-type random partition models

Gibbs-type priors were first introduced in Pitman (2003) and then extensively studied in Gnedin and Pitman (2006), including their relation to Bayesian nonparametrics. Further references also include Cerquetti (2008); Lijoi and Prünster (2009); Pitman (2006). Gibbs-type random partitions define a general family of exchangeable random partitions. An exchangeable random partition π_p of the first p positive integers is said to be of Gibbs form if the exchangeable partition probability function (EPPF) of π_p can be expressed in the product form:

$$\mathrm{pr}(\pi_p) = \mathrm{pr}(|C_1|, \ldots, |C_M|) = V_p(M) \prod_{m=1}^{M} W_{|C_m|}(\phi), \tag{3.7}$$

for all $1 \leq M \leq p$, and all compositions $|C_1|, \ldots, |C_M|$ of p, where we use the general notation ϕ to denote the parameters of the Gibbs-type partition models. Gnedin and Pitman (2006) show that to define an exchangeable partition the set of non-negative weighing $\{W_{|C|}(\phi) : 1 \leq |C| \leq p\}$ must have the form

$$W_{|C|}(\phi) = \frac{\Gamma(|C| - \delta)}{\Gamma(1 - \delta)},$$

where the parameter δ satisfies $\delta \in [-\infty, 1)$. Moreover, the set of weights $\{V_p(M) : p \geq 1, 1 \leq M \leq p\}$ must satisfy the recursive relation

$$V_p(M) = (p - \delta M)V_{p+1}(M) + V_{p+1}(M + 1),$$

with $V_1(1) = 1$.

The parameters ϕ and weights $V_p(M)$ are the main components for defining the Gibbs-type random partitions and the associated Gibbs-type priors. Hence, the choice of ϕ and weights $V_p(M)$ are important, especially ϕ since it determines the clustering structure as well as the asymptotic behaviour of the Gibbs-type model.

We focus our study on three cases within the Gibbs-type family:

1. Dirichlet process (DP) with concentration parameter $\alpha > 0$;
2. Pitman Yor process (PY) with discount parameter $\delta \in [0, 1)$ and concentration parameter $\alpha > -\delta$; and
3. mixture of finite mixtures (MFM) with parameter $\gamma > 0$ (larger values encouraging more equally sized clusters) and a distribution $p_L(\cdot|\psi)$ with parameter ψ related to the prior on the number of clusters, where L denotes the number of components (Miller and Harrison, 2018).

Table 3.1 summarises the $V_p(M)$ and $W_{|C_m|}(\phi)$ for DP, PY and MFM.

3.3.2 Markov random field

Potts models belong to the general class of Markov random field (MRF) models, as such, we begin with a brief review of MRFs. For a more in-depth treatment, the reader may refer to Besag (1974); Geman and Geman (1984); Li (2009); Winkler (2003). The MRF is a spatial process related to grid-like structures, which is frequently used in spatial statistics and in image segmentation applications to incorporate the spatial interactions between adjacent neighbours. A random field is called an MRF with respect to the predefined neighbourhood structure $\{j \sim k\}$ when

$$\text{pr}(z_j|z_k, j \neq k) = \text{pr}(z_j|z_k, j \sim k).$$

Each node of the grid only interacts with its neighbours. These nodes are also referred to as sites. For a regular grid with size $r \times c$ (r rows and c columns), the nearest first-order neighbours are at locations $(r+1, c), (r-1, c), (r, c+1), (r, c-1)$, if applicable. For the rest of the article, we will assume that the nearest first-order neighbours are considered as the neighborhood system.

The model is referred to as the Ising model when the z_j's are discrete with only $M = 2$ categories, i.e. the realised values of $z_j \in \{-1, 1\}$ (Ising, 1924). The Potts model is a generalisation of the Ising model, where the number of categories $M \geq 2$ (Potts and Domb, 1952). It allows for spatial correlation between neighbouring labels in the form of an MRF. A M-state Potts model can be defined in term of site-wise term $\mathcal{A}(z_1, \cdots z_p)$ and interaction term $\mathcal{B}(z_1, \cdots z_p)$:

$$\begin{aligned} \text{pr}(z_1, \cdots, z_p) &\propto \mathcal{A}(z_1, \cdots z_p)\mathcal{B}(z_1, \cdots z_p), \quad \text{with} \\ \mathcal{A}(z_1, \cdots z_p) &:= \frac{1}{c_{\mathcal{A}}} \exp\left(\sum_{j=1}^{p} h_{z_j}\right), \\ \mathcal{B}(z_1, \cdots z_p) &:= \frac{1}{c_{\mathcal{B}}} \exp\left(\sum_{j \sim k} \upsilon_{jk} \mathbb{1}_{z_j = z_k}\right), \end{aligned} \tag{3.8}$$

where $\mathbb{1}_{z_j=z_k}$ equals 1 if j and k share the same cluster label and 0 otherwise. Both $c_{\mathcal{A}}$ and $c_{\mathcal{B}}$ represent the normalising constant. The $\boldsymbol{h} = (h_1, \cdots, h_M)$ is an additional external field parameter, where each h_m is a scalar. The interaction term $\mathcal{B}(z_1, \cdots z_p)$ models spatial correlation. Often, the external field parameter is constant, i.e.
$h_1 = \cdots = h_M$ so that only the interaction term remains. The $\upsilon_{jk} > 0$ is called the smoothing parameter where larger υ_{jk} encourages more spatial smoothing. Often one simplifies and uses a single smoothing parameter υ, with $\upsilon_{jk} = \upsilon$ for all j, k. If z_j is distinct from all neighbours, $\exp(\sum_{j\sim k} \upsilon_{jk} \mathbb{1}_{z_j=z_k}) = 1$, whereas $\exp(\sum_{j\sim k} \upsilon_{jk} \mathbb{1}_{z_j=z_k}) > 1$ if at least one neighbour is assigned to the same category.

For a square grid ($r = c$), the total number of connected neighbours pairs, $S \leq 2(r^2 - r)$ while for a rectangular grid, $S \leq 2r \times c - r - c$. When υ_{jk} equals to 0, the z_j's are independent and distributed on $\{0, \cdots, M\}$ according to site-wise term. For a sufficiently large υ, the asymptotic value of the expectation of S approaches the total number of edges, while the variance is close to zero. This is because all units/sites tend to be attracted to a single category when υ increases above a certain threshold. This is known as the phase transition of the Potts model.

The predictive probability of the Potts model for $m \leq M$ (within constant site-wise term):

$$\mathrm{pr}_{\mathrm{Potts}}(z_j = m | \boldsymbol{z}_{1:j-1}, \upsilon) \propto \exp\left(\sum_{j\sim k} \upsilon \mathbb{1}_{z_j=z_k=m}\right).$$

During the phase transition, the critical value of υ marks the transitions of the Potts model from a disordered ($\upsilon < \upsilon_{crit}$) to an ordered ($\upsilon > \upsilon_{crit}$) state. For a regular 2D grid, Potts and Domb (1952) shows that the critical value can be exactly computed using the following formula:

$$\upsilon_{crit} = \log(1 + \sqrt{M}).$$

This is the point at which a phase transition occurs for a grid with r rows and an infinite number of columns. However, as the size of the system, represented by p, increases, the error caused by a finite boundary decrease, as a result of the finite-dimensional scaling property of the Potts model.

Detecting the phase transition of the Potts model in a graph with a dimension higher than one can be quite difficult and challenging due to the intractability of the normalising constant. This is because the Potts model tends to undergo an abrupt and drastic change in phase transition as coupling υ increases. There are certain combinations of hyperparameters that lead to the allocation of all units to a single category. Therefore, a proper specification of the hyperparameters for the Potts model is very much dependent upon finding the bounds of the phase transition.

3.3.3 Markov random field constrained Gibbs-type priors

Here we describe the Potts-Gibbs models with further details provided for three specific cases: the Potts-Dirichlet process random partition (Potts-DP) model, the Potts-Pitman Yor process random partition (Potts-PY) model and Potts-mixture of finite mixtures random partition (Potts-MFM) model. The Potts-Gibbs models combine BNP random partition models, which avoid the need to prespecify the number of clusters, allowing it to be determined and grow with the data, with a Potts-like spatial smoothness component (Potts and Domb, 1952). Spatial random partition models in this direction are a growing research area, including MRFs with the PPM (Pan et al., 2020), with DP (Da Xu et al., 2016; Orbanz and Buhmann, 2008), with PY (Lü et al., 2020) and with MFM (Hu et al., 2022; Zhao et al., 2020). Precisely, within the BNP framework, we focus on the class of Gibbs-type random partitions (Cerquetti, 2008; Gnedin and Pitman, 2006; Lijoi and Prünster, 2009; Pitman, 2006), motivated by their comprise between tractable predictive rules and richness of the predictive structure, including important cases, such as the DP (Ferguson, 1973), PY (Perman et al., 1992; Pitman, 1996), and MFM (Miller and Harrison, 2018).

Definition 3.1: Potts-Gibbs random partition models

The Potts-Gibbs random partition (Potts-Gibbs) models are defined as:

$$\begin{aligned}\mathrm{pr}(\pi_p) &\propto \mathcal{B}(z_1,\cdots z_p)\mathrm{pr}(|C_1|,\ldots,|C_M|),\\ \mathcal{B}(z_1,\cdots z_p) &\propto \text{Potts model}(\upsilon),\\ \mathrm{pr}(|C_1|,\ldots,|C_M|) &\propto \text{Gibbs-type random partition models}(\phi),\end{aligned}$$

where the first term is defined by the interaction term $\mathcal{B}(z_1,\cdots z_p)$ from the Potts model in Equation (3.8) to capture spatial interaction among vertices and the second term $\mathrm{pr}(|C_1|,\ldots,|C_M|)$ is a Gibbs-type random partition model with ϕ denoting the parameters of the Gibbs-type random partition model. The normalising constant is defined as

$$c_{(\upsilon,\phi)} = \sum_{\pi_p}\mathcal{B}(z_1,\cdots z_p)\mathrm{pr}(|C_1|,\ldots,|C_M|).$$

The Potts-Gibbs models can be summarised as:

$$\mathrm{pr}(\pi_p) \propto \exp\underbrace{\left(\sum_{j\sim k}\upsilon\mathbb{1}_{z_j=z_k}\right)}_{\text{Potts model}}\underbrace{\left(V_p(M)\prod_{m=1}^{M}W_{|C_m|}(\phi)\right)}_{\text{Gibbs-type random partition models}}.$$

Spatial smoothness is introduced through the Potts model as indicated by the equation. Under the smoothness constraints on cluster assignments, the Potts term prefers to allocate two adjacent units to the same cluster. However, note that the constrained model exhibits a key property that Potts model constraints only change the finite components of the Potts-Gibbs models (Proposition 3.1). The finite components represent the clusters that the model has created. In other words, unless the base measure F_0 has discrete atoms, a draw from the base measure F_0 always defines a new cluster, and the corresponding unit will not be affected by the smoothness constraint, the Potts model. In the following, we assume the spatial locations lie on a rectangular grid with first-order neighbours and all neighbour pairs have a common coupling parameter $\upsilon_{jk} = \upsilon$. The higher the value of υ, the model encourages more spatially smooth partitioning.

Proposition 3.1: Predictive distribution

The predictive distribution of the Potts-Gibbs random partition model is:

$$\mathrm{pr}_{\text{Potts-Gibbs}}(z_{j+1} = m | \boldsymbol{z}_{1:j}, \phi, \upsilon) = \begin{cases} \frac{V_{j+1}(M_j)}{V_j(M_j) + V_{j+1}(M_j)\eta_{j+1}} \lambda_{j+1,m} & m \leq M_j \\ \frac{V_{j+1}(M_j+1)}{V_j(M_j) + V_{j+1}(M_j)\eta_{j+1}} & m = M_j + 1, \end{cases}$$

where $\lambda_{j+1,m} = (|C_{m,j}| - \phi)\exp(\upsilon S_{m,j+1})$ with and $\eta_{j+1} = \sum_{m:S_{m,j+1}>0}(|C_{m,j}| - \phi)\{\exp(\upsilon S_{m,j+1}) - 1\}$.

3.3.4 Potts-Dirichlet process (Potts-DP) model

Definition 3.2: Potts-Dirichlet process random partition model

Potts-Dirichlet process random partition (Potts-DP) model is defined as the random partitions π_p induced by a combination of the Potts model and Dirichlet process (DP):

$$\mathrm{pr}_{\text{Potts-DP}}(\pi_p) \propto \exp\left(\sum_{j\sim k} \upsilon \mathbb{1}_{z_j = z_k}\right) \frac{\alpha^M}{\alpha^{(p)}} \prod_{m=1}^{M} \Gamma(|C_m|), \tag{3.9}$$

with

$$V_p(M) = \frac{\Gamma(\alpha)\alpha^M}{\Gamma(\alpha + p)} = \frac{\alpha^M}{\alpha^{(p)}},$$

where α is the concentration parameter for DP.

Equation (3.9) arises as a product of three factors: the first one is the Potts interaction term depends only on S, the total number of connected neighbours, the second depends only on (p, M) and the third one depends on the frequencies $(|C_1|, \ldots, |C_M|)$ via the product $\prod_{m=1}^{M} \Gamma(|C_m|)$.

Polya urn scheme/Restaurant process

Proposition 3.2: Predictive distribution

The predictive distribution of the Potts-DP model is:

$$\mathrm{pr}_{\text{Potts-DP}}(z_{j+1} = m | \boldsymbol{z}_{1:j}, \alpha, \upsilon) \propto \begin{cases} \frac{|C_{m,1:j}|}{\alpha+j+\eta_{j+1}} \exp(\upsilon S_{m,j+1}) & m \leq M_j \\ \frac{\alpha}{\alpha+j+\eta_{j+1}} & m = M_j + 1, \end{cases} \tag{3.10}$$

where the parameter α acts as a prior weight on the formation of a new cluster.

A fundamental feature of partitions generated by the DP is a "rich-getting-richer" property resulting in partitions containing a few large clusters and many small clusters. The prediction rule from Equation (3.10) suggests that the use of a DP prior will lead to partitions that are dominated by a few large clusters since larger clusters will tend to attract new observations during the sequential creation of a partition.

From Equation (3.10), we see that a new cluster is generated with probability proportional to the concentration parameter α. In other words, a large value of α increases the expected number of clusters. There is another parameter, the smoothing parameter, υ from the Potts component which a large value of υ encourages spatial smoothness. We would expect a larger value of υ from the Potts component for the Potts-DP model to compete with the DP component and essentially eliminate the small clusters. Thus, partitions tend to be composed of the large connected component and many small/singleton clusters, and care needs to be taken to avoid and not exacerbate the phase transition of the Potts model.

The number of clusters

As $p \to \infty$, the expected number of unique clusters M for a DP model is

$$\mathbb{E}[M] \simeq \alpha \log(p). \tag{3.11}$$

Moreover, Korwar and Hollander (1973) show that the number of cluster for size p, $M_p \simeq \alpha \log(p)$, almost surely as $p \to \infty$.

Proposition 3.3: Number of clusters (Lü et al., 2020)

Assume that the graph has a maximal degree D. For a Potts-DP model, the lower bound on the expected number of unique clusters M is:

$$\mathbb{E}[M] \gtrsim \frac{\alpha}{\exp(D\upsilon)} \log(p). \tag{3.12}$$

It is worth mentioning that the MRF component of a Potts-DP model will only cause the reduction in the prior expected number of clusters, thus Equation (3.11) provides an asymptotic upper bound on the expected number of clusters.

Cluster sizes

The asymptotic behaviour of the expected number of clusters of a given size under the DP prior alone when p tends to infinity is defined as below:

$$\lim_{p\to\infty} \mathbb{E}[N_{|C_m|}] = \frac{\alpha}{|C_m|},$$

where $N_{|C_m|}$ denotes the number of clusters of size $|C_m|$(Wallach et al., 2010).

This well-known result (Arratia et al., 2003) means that when $p \to \infty$, regardless of the value of α, the expected number of clusters of size $|C_m|$ is inversely proportional to $|C_m|$. In other words, in expectation, there will be a small number of large clusters and vice versa. We can write the equation on a log-log scale as for sufficiently large p, it yields:

$$\log\left\{\mathbb{E}[N_{|C_m|}]\right\} \simeq c - \log(|C_m|),$$

where c is a constant.

With the addition of the Potts model, two opposing effects emerge in the Potts-DP model. They are the concentration parameter α from the DP component and the smoothing parameter υ from the Potts component. These two opposing forces compete with each other and influence the prior distribution of the cluster sizes. The greater value of υ results in producing partitions encompassing larger clusters with the largest sizes, and fewer clusters of medium size. One possible explanation is that for a sufficiently large υ, beyond a specific size, almost surely most of the units are attracted to the same cluster and ultimately they get to a size approaching the maximum size limit.

3.3.5 Potts-Pitman Yor process (Potts-PY) model

From a purely conceptual point of view, having the sampling probability of a new unit not depend on the number of distinct units present in the current partition seems too restrictive. Some argue that it would be preferable to also explicitly include in the probability the number of distinct units present in the partition currently under consideration as it helps to generalise the heterogeneity in the partition. Thereby, in certain applications, one might prefer the Potts-PY model, which we describe in more detail here.

Definition 3.3: Potts-Pitman Yor process random partition model

The probability for a partition π_p having the Potts-Pitman Yor process random partition (Potts-PY) model is a combination of the Potts model and Pitman Yor process (PY) as shown below:

$$\mathrm{pr}_{\text{Potts-PY}}(\pi_p) \propto \exp\left(\sum_{j\sim k} \upsilon \mathbb{1}_{z_j=z_k}\right) (\delta+\alpha)_{(\delta)}^{(M-1)} \prod_{m=1}^{M} \frac{\Gamma(|C_m|-\delta)}{\Gamma(1-\delta)}, \tag{3.13}$$

with discount parameter $\delta \in (0,1)$, concentration parameter $\alpha > -\delta$, and

$$V_p(M) = \frac{\Gamma(\alpha+1)\prod_{m=1}^{M-1}(\alpha+m\delta)}{\Gamma(\alpha+p)} = \frac{\prod_{m=1}^{M-1}(\alpha+m\delta)}{(\alpha+1)^{(p-1)}} \propto (\delta+\alpha)_{(\delta)}^{(M-1)}.$$

When $\delta = 0$, Equation (3.13) reduces to Equation (3.9) therefore leading to the Potts-DP model. If $\upsilon = 0$, we obtain the partitions induced by the PY.

Polya urn scheme/Restaurant process

Proposition 3.4: Predictive distribution

The predictive distribution of the Potts-PY model:

$$\mathrm{pr}_{\text{Potts-PY}}(z_{j+1} = m|\boldsymbol{z}_{1:j}, \alpha, \delta, \upsilon) \propto \begin{cases} \frac{|C_{m,1:j}|-\delta}{\alpha+j+\eta_{j+1}} \exp(\upsilon S_{m,j+1}) & m \leq M_j \\ \frac{\alpha+\delta M_j}{\alpha+j+\eta_{j+1}} & m = M_j+1. \end{cases} \tag{3.14}$$

The frequencies of each cluster, as well as the number of distinct clusters present in the partition, determine the allocation in Equation (3.14). The discount parameter δ plays a significant role in controlling the combined effect of the reinforcement mechanism and the rate at which new clusters are generated. Among the observed clusters, δ mitigates the "rich-get-richer" properties. The δ can be used to adjust how strongly the probability of acquiring a new unit depends on M since the probability of acquiring a new unit increases monotonically in M.

The number of clusters

Pitman (2006) showed that as $p \to \infty$, the expected number of unique clusters for PY is:

$$\mathbb{E}[M] \simeq \frac{\Gamma(\alpha+1)}{\delta\Gamma(\alpha+\delta)} p^{\delta}. \tag{3.15}$$

Pitman (2003) shows that the number of cluster for size p, $M_p \simeq S_{\delta,\alpha} p^{\delta}$ as $p \to \infty$ where $S_{\delta,\alpha}$ is a positive random variable with density depending on δ and α.

A power law governs the expected number of clusters of a given size and the number of clusters increases at a faster rate compared with the DP.

Proposition 3.5: Number of clusters (Lü et al., 2020)

Assume that the graph has a maximal degree D. For the Potts-PY model, the expected number of unique clusters is:

$$\mathbb{E}[M] \gtrsim c p^{\delta \exp(-D\upsilon)}, \tag{3.16}$$

where c is for some positive constants.

Similar to the Potts-DP model, the Potts component can only reduce the prior expected number of clusters and thus Equation (3.15) provides an asymptotic upper bound on the prior expected number of clusters under the Potts-PY model.

Cluster sizes

Pitman's result can also be used to derive the expected number of clusters of size $|C_m|$ in a partition:

$$\mathbb{E}[N_{|C_m|}] \simeq \frac{\Gamma(1+\alpha)\prod_{l=1}^{|C_m|-1}(l-\delta)}{\Gamma(\alpha+\delta)|C_m|!} p^{\delta}.$$

This equation is best illustrated in a log-log scale as it yields, for p large enough:

$$\log\left(\mathbb{E}[N_{|C_m|}]\right) \simeq c - (1+\delta)\log(|C_M|),$$

where c is a constant. Hence in a log-log scale plot, the cluster size distribution for large p in the DP case has a slope equal to -1, while a PY features a steeper slope of $-(1+\delta)$.

For the Potts-PY model, the effect of υ on the size of clusters is similar to the Potts-DP model.

3.3.6 Potts-mixture of finite mixtures (Potts-MFM) model

Another interesting random partition model within the class of Gibbs-type priors that we want to consider apart from the DP or PY is that arising from MFM (Miller and Harrison, 2018). In the Potts-MFM model, the MFM random partition model is combined with the Potts model. The motivation for choosing to include the MFM is twofold. First, we will have more control over the prior on number of clusters, which in turn can also be used for studying the critical value of the coupling υ. Second, given the number of clusters, through the appropriate specification of the hyperparameters, most of the prior mass is on clusters of similar size. This will help mitigate the phase transition of the Potts model and a strong preference for one large connected cluster.

Mixture of finite mixtures (MFM) random partition model

While the CRP has a very attractive feature of simultaneously estimating the number of clusters and the cluster configuration, limitations exist. Specifically, Miller and Harrison (2018) proved that the CRP allocates large probabilities to clusters with relatively smaller sizes, which leads to producing extraneous clusters in the posterior leading to inconsistent estimation of the number of clusters, when the true number is finite, even when the sample size grows to infinity. A modification of the CRP called a mixture of finite mixtures (MFM) model is proposed to circumvent this issue (Miller and Harrison, 2018).

The MFM random partition model is:

$$\mathrm{pr}(|C_1|, \ldots, |C_M|) = \sum_{l=1}^{\infty} \frac{l_{(M)}}{(\gamma l)^{(p)}} p_L(l|\psi) \prod_{m=1}^{M} \frac{\Gamma(|C_m| + \gamma)}{\Gamma(\gamma)}, \tag{3.17}$$

where $\gamma > 0$ is a parameter of the MFM with larger values encouraging more homogeneous cluster sizes;

$$\begin{aligned} V_p(M) &= \sum_{l=1}^{\infty} \frac{\Gamma(\gamma l) l!}{\Gamma(\gamma l + p)(l - M)!} p_L(l|\psi) \\ &= \sum_{l=1}^{\infty} \frac{l_{(M)}}{(\gamma l)^{(p)}} p_L(l|\psi), \end{aligned}$$

with $p_L(\cdot|\psi)$ a pmf on $\{0, 1, \ldots\}$ reflecting prior belief on the number of clusters with parameter ψ related to the prior on the number of clusters. Examples of $p_L(\cdot|\psi)$ studied include (1) $\mathrm{Pois}(L-1|\lambda)$ where $\psi = \lambda > 0$ and $\mathbb{E}[L] = \lambda + 1$; (2) $\mathrm{Geom}(L|p)$, where $\psi = p \in (0, 1)$ $\mathbb{E}[L] = 1/p$; and (3) $\mathrm{Unif}(1, \ldots, L_{\max})$ where $\psi = L_{\max}$. Note that Equation (3.17) is a member of the family of Gibbs partition distributions. Normalisation $\tilde{V}_p(M) = \gamma^M V_p(M)$ is used to represent Equation (3.17) as the standard form in Equation (3.7):

$$\mathrm{pr}(|C_1|, \ldots, |C_M|) = \tilde{V}_p(M) \prod_{m=1}^{M} \frac{\Gamma(|C_m| + \gamma)}{\Gamma(1 + \gamma)}.$$

And the recursive relation for the MFM is

$$V_p(M) = (p + \gamma M) V_{p+1}(M) + \gamma V_{p+1}(M + 1).$$

Alternatively, in standard form with normalisation $\tilde{V}_p(M) = \gamma^M V_p(M)$, the recursive relation is the same format as used to define Gibbs-type priors:

$$\tilde{V}_p(M) = (p + \gamma M)\tilde{V}_{p+1}(M) + \tilde{V}_{p+1}(M+1).$$

The restaurant process associated with the MFM prior is:

$$\begin{aligned}
\mathrm{pr}_{\text{MFM}}(z_{j+1} = m | \boldsymbol{z}_{1:j}, \gamma, \psi) &= \begin{cases} \frac{V_{j+1}(M_j)}{V_{j+1}(M_j)}(\gamma + |C_{m,j}|) & m \leq M_j \\ \frac{V_{j+1}(M_j+1)}{V_{j+1}(M_j)}\gamma & m = M_j + 1 \end{cases} \\
&\propto \begin{cases} \gamma + |C_{m,j}| & m \leq M_j \\ \frac{V_{j+1}(M_j+1)}{V_{j+1}(M_j)}\gamma & m = M_j + 1. \end{cases}
\end{aligned}$$

This is similar to the CRP of the DP but notes that $\gamma > 0$ helps to mitigate the "rich-getting-richer" property. Compared to the CRP, the introduction of new clusters is slowed down by the factor $\frac{V_{j+1}(M_j+1)}{V_{j+1}(M_j)}$, which allows a model-based pruning of the tiny extraneous clusters.

Potts-mixture of finite mixtures (Potts-MFM) random partition model

Definition 3.4: Potts-mixture of finite mixtures random partition model

The Potts-mixture of finite mixtures random partition (Potts-MFM) model:

$$\begin{aligned}
\mathrm{pr}_{\text{Potts-MFM}}(\pi_p) &\propto \exp\left(\sum_{j \sim k, j<k} \upsilon \mathbb{1}_{z_j = z_k}\right) V_p(M) \prod_{m=1}^{M} \frac{\Gamma(|C_m| + \gamma)}{\Gamma(\gamma)} \\
&\propto V_p(M) \prod_{m=1}^{M} \exp\left(\upsilon S_m\right) \frac{\Gamma(|C_m| + \gamma)}{\Gamma(\gamma)} \\
&\propto V_p(M) \exp\left(\upsilon S\right) \prod_{m=1}^{M} \frac{\Gamma(|C_m| + \gamma)}{\Gamma(\gamma)}.
\end{aligned} \tag{3.18}$$

The normalising constant is:

$$\begin{aligned}
c_{\text{Potts-MFM}} &= \sum_{\pi_p} V_p(M) \exp\left(\upsilon S\right) \prod_{m=1}^{M} \frac{\Gamma(|C_m| + \gamma)}{\Gamma(\gamma)} \\
&= \sum_{M=1}^{p} V_p(M) \exp\left(\upsilon S\right) \sum_{(C_1, \ldots, C_M)} \prod_{m=1}^{M} \frac{\Gamma(|C_m| + \gamma)}{\Gamma(\gamma)}.
\end{aligned}$$

Note that Potts-MFM model was recently proposed in Pan et al. (2020) (Definition 3.4).

While they study posterior consistency for the true partition under spatial panel data, i.e. the number of locations/items to a cluster is fixed but the number of observations/time points for each location tends to infinity, they do not study prior properties, such as the prior expected number of clusters. In fact, they fix $\gamma = 1$ and $p_L(l|\psi = 10) = 10^{l-1}\exp(10)/(l-1)!$, i.e. Pois(10) truncated to the positive integers and select $\upsilon \in \{0, 0.1, \dots, 1\}$ that maximizes a prior-based approximation of the marginal likelihood. In the following, we provide novel contributions to enrich the understanding of the Potts-MFM model and the role of the hyperparameters.

Polya urn scheme/Restaurant process

Proposition 3.6: Predictive distribution

The predictive distribution of the Potts-MFM model is:

$$\mathrm{pr}_{\text{Potts-MFM}}(z_{j+1} = m|\boldsymbol{z}_{1:j}, \gamma, \psi, \upsilon) = \begin{cases} \frac{V_{j+1}(M_j)}{V_j(M_j)+V_{j+1}(M_j)\eta_{j+1}}\lambda_{j+1,m} & m \leq M_j \\ \frac{V_{j+1}(M_j+1)}{V_j(M_j)+V_{j+1}(M_j)\eta_{j+1}}\gamma & m = M_j + 1, \end{cases}$$

where $\lambda_{j+1,m} = (\gamma + |C_{m,j}|)\exp(\upsilon S_{m,j+1})$ and $\eta_{j+1} = \sum_{m:S_{m,j+1}>0}(\gamma + |C_{m,j}|)$ $\{\exp(\upsilon S_{m,j+1}) - 1\}$.

Proof of Proposition 3.6. This is a simple extension of Lü et al. (2020). First notice that:

$$\begin{aligned}\mathrm{pr}_{\text{Potts-MFM}}(z_{j+1} = m|\boldsymbol{z}_{1:j}, \gamma, \psi, \upsilon) &\propto V_{j+1}(M_{j+1})\exp\left\{\upsilon S(\boldsymbol{z}_{1:(j+1)})\right\}\prod_{m=1}^{M_{j+1}}\gamma^{(|C_{m,j+1}|)} \\ &\propto \frac{V_{j+1}(M_{j+1})\exp\left\{\upsilon S(\boldsymbol{z}_{1:(j+1)})\right\}\prod_{m=1}^{M_{j+1}}\gamma^{(|C_{m,j+1}|)}}{V_j(M_j)\exp\left\{\upsilon S(\boldsymbol{z}_{1:j})\right\}\prod_{m=1}^{M_j}\gamma^{(|C_{m,j}|)}} \\ &\propto \begin{cases} V_{j+1}(M_j)(\gamma + |C_{m,j}|)\exp(\upsilon S_{m,j+1}) & m \leq M_j \\ V_{j+1}(M_j+1)\gamma & m = M_j + 1. \end{cases}\end{aligned}$$

Thus, we have that:

$$\mathrm{pr}_{\text{Potts-MFM}}(z_{j+1} = m|\boldsymbol{z}_{1:j}, \gamma, \psi, \upsilon) = \begin{cases} \frac{V_{j+1}(M_j)}{V_{j+1}(M_j)\sum_{m=1}^{M_j}\lambda_{j+1,m}+V_{j+1}(M_j+1)\gamma}\lambda_{j+1,m} & m \leq M_j \\ \frac{V_{j+1}(M_j+1)}{V_{j+1}(M_j)\sum_{m=1}^{M_j}\lambda_{j+1,m}+V_{j+1}(M_j+1)\gamma}\gamma & m = M_j + 1. \end{cases}$$

The proof is complete by noting that

$$
\begin{aligned}
&V_{j+1}(M_j)\sum_{m=1}^{M_j}\lambda_{j+1,m} + V_{j+1}(M_j+1)\gamma \\
&= V_{j+1}(M_j)\left[\sum_{m=1}^{M_j}(\gamma+|C_{m,j}|) + \sum_{m:S_{m,j+1}>0}(\gamma+|C_{m,j}|)\left\{\exp(\upsilon S_{m,j+1})-1\right\}\right] + V_{j+1}(M_j+1)\gamma \\
&= V_{j+1}(M_j)\left\{(\gamma M_j + j) + \eta_{j+1}\right\} + V_{j+1}(M_j+1)\gamma \\
&= V_{j+1}(M_j)\eta_{j+1} + \left\{V_{j+1}(M_j)(\gamma M_j + j) + V_{j+1}(M_j+1)\gamma\right\} \\
&= V_{j+1}(M_j)\eta_{j+1} + V_j(M_j).
\end{aligned}
$$

□

Notice that the MRF has the effect of reducing the probability of a new cluster, as

$$\frac{V_{j+1}(M_j+1)}{V_j(M_j)} \geq \frac{V_{j+1}(M_j+1)}{V_j(M_j) + V_{j+1}(M_j)\eta_{j+1}}.$$

Note we can also write:

$$\mathrm{pr}_{\text{Potts-MFM}}(z_{j+1}=m|\boldsymbol{z}_{1:j},\gamma,\psi,\upsilon) \propto \begin{cases}(\gamma+|C_{m,j}|)\exp(\upsilon S_{m,j+1}) & m \leq M_j \\ \frac{V_{j+1}(M_j+1)}{V_{j+1}(M_j)}\gamma & m = M_j+1.\end{cases} \tag{3.19}$$

The number of clusters

We study how the Potts component affects the random partition model. First, we consider the number of clusters.

Proposition 3.7: Number of clusters

Assume that the graph has a maximal degree D. Then, the prior expected number of clusters for the Potts-MFM model has the following bounds:

$$\mathbb{E}_{p_L}[L]\exp(-D\upsilon) \lesssim \mathbb{E}[M_p] \lesssim \mathbb{E}_{p_L}[L], \tag{3.20}$$

where $\mathbb{E}_{p_L}[L]$ is the expectation of L with respect to $p_L(\cdot)$.

Proof of Proposition 3.7. Since $M_p = \sum_{j=0}^{p-1} D_j$, where $D_j = \mathbb{1}(z_{j+1} = M_j + 1)$ given

$\boldsymbol{z}_{1:j}$.

$$\begin{aligned}\mathbb{E}[M_p] &= \sum_{j=0}^{p-1}\mathbb{E}[D_j] = \sum_{j=0}^{p-1}\mathbb{E}\left[\frac{V_{j+1}(M_j+1)\gamma}{V_j(M_j)+V_{j+1}(M_j)\eta_{j+1}}\right] \\ &= \sum_{j=0}^{p-1}\mathbb{E}\left[\frac{V_{j+1}(M_j+1)\gamma}{V_j(M_j)}\left(\frac{1}{1+\frac{V_{j+1}(M_j)}{V_j(M_j)}\eta_{j+1}}\right)\right],\end{aligned}$$

where the last equations follows from Proposition 3.6. Note that the first term $V_{j+1}(M_j+1)\gamma/V_j(M_j)$ represents the probability of allocation to a new cluster under the standard MFM. Focusing on the second term, we first note that:

$$\frac{V_{j+1}(M_j)}{V_j(M_j)} = \frac{1-\frac{V_{j+1}(M_j+1)\gamma}{V_j(M_j)}}{j+\gamma M_j},$$

following the recursive relation of the coefficients $V_{p+1}(M_p+1)\gamma = V_p(M_p) - (p+\gamma M_p)V_{p+1}(M_p)$. Furthermore:

$$\begin{aligned}\eta_{j+1} &= \sum_{m:S_{m,j+1}>0}(\gamma+|C_{m,j}|)\{\exp(\upsilon S_{m,j+1})-1\} \\ &\leq [\exp\{\upsilon\min(D,j)\}-1]\sum_{m:S_{m,j+1}>0}(\gamma+|C_{m,j}|).\end{aligned}$$

Thus, we have:

$$\begin{aligned}\frac{V_{j+1}(M_j)}{V_j(M_j)}\eta_{j+1} &\leq [\exp\{\upsilon\min(D,j)\}-1]\left(1-\frac{V_{j+1}(M_j+1)\gamma}{V_j(M_j)}\right)\frac{\sum_{m:S_{m,j+1}>0}(\gamma+|C_{m,j}|)}{j+\gamma M_j} \\ &\leq [\exp\{\upsilon\min(D,j)\}-1],\end{aligned}$$

and

$$\frac{1}{1+\frac{V_{j+1}(M_j)}{V_j(M_j)}\eta_{j+1}} \geq \exp\{-\upsilon\min(D,j)\}.$$

Therefore,

$$\begin{aligned}\mathbb{E}[M_p] &= \sum_{j=0}^{p-1}\mathbb{E}\left[\frac{V_{j+1}(M_j+1)\gamma}{V_j(M_j)}\left(\frac{1}{1+\frac{V_{j+1}(M_j)}{V_j(M_j)}\eta_{j+1}}\right)\right] \\ &\geq \sum_{j=0}^{p-1}\mathbb{E}\left[\frac{V_{j+1}(M_j+1)\gamma}{V_j(M_j)}\right]\exp\{-\upsilon\min(D,j)\} \\ &\gtrsim \mathbb{E}_{p_L}[L]\exp(-D\upsilon),\end{aligned}$$

where the last line follows from (Miller and Harrison, 2018, Theorem 5.2), i.e. asymptotically the distribution of the number of clusters behaves like the number of components. Similarly, the upper bound follows from

$$\frac{V_{j+1}(M_j+1)}{V_j(M_j)} \geq \frac{V_{j+1}(M_j+1)}{V_j(M_j)+V_{j+1}(M_j)\eta_{j+1}}.$$

□

For a given value of the coupling parameter υ, we can choose the prior p_L accordingly based on these bounds to reflect our prior information on the number of clusters.

Cluster sizes

The distribution of the cluster sizes under MFM is:

$$\mathrm{pr}_{\mathrm{MFM}}(|C_1|,\ldots,|C_M|) \propto \prod_{m=1}^{M} |C_m|^{\gamma-1}.$$

From the equation, we see that smaller probabilities are assigned to highly imbalanced cluster sizes. The parameter γ controls the relative size of the resulting clusters.

The distribution of the cluster sizes under Potts-MFM model is:

$$\mathrm{pr}_{\text{Potts-MFM}}(|C_1|=N_1,\ldots,|C_M|=N_M) \propto \frac{V_p(M)}{M!}\prod_{m=1}^{M}\frac{\Gamma(\gamma+N_m)}{\Gamma(\gamma)}\sum_{\pi_p:|C_m|=N_m}\prod_{m=1}^{M}\exp(\upsilon S_m).$$

Given M, a larger value of γ should encourage more equal-size clusters, however, this needs to balance with the Potts part, which will encourage more connected neighbours. In Section 3.5, we will examine the prior on S and the prior on the cluster sizes for the Potts-MFM model.

The computation of the coefficients $V_p(M)$

See Section 3.2 of Miller and Harrison (2018) for details on the computation of $V_p(M)$; they note that convergence of $V_p(M)$ is at least as rapid as $\sum_{l=1}^{\infty} p_L(l|\psi)$, thus if we choose the prior on the number of components to not have a heavy tail, we can approximate $V_p(M)$ well numerically with a finite number of terms. In particular, Theorem 5.1 states:

$$V_p(M) \simeq \frac{M_{(M)}}{(\gamma M)^{(p)}} p_L(M|\psi) \simeq \frac{M!}{p!}\frac{\Gamma(\gamma M)}{p^{\gamma M-1}} p_L(M|\psi),$$

as $p \to \infty$.

3.4 Comparison to other partition distributions

We now compare the EPA distribution and the Potts-Gibbs models with the dependent random partition models reviewed in Section 2.2. Each dependent random partition differs on the underlying mechanism to generate partitions. The ddCRP by Blei and Frazier (2011) and Ghosh et al. (2011) defines the random partition indirectly through a probability distribution over graphs and incorporates the spatial information in the model using the distance matrix and decay function. The PPMx by Müller et al. (2011) uses both the cohesion function and similarity function to define the random partition directly rather than through some underlying discrete random probability measure.

Similarly to the ddCRP, the EPA distribution incorporates covariate information via the pairwise distance but also explicitly defines a probability distribution over partitions by sequentially allocating the units into clusters in a partition. It serves as a flexible prior, allowing us to incorporate a spatial constraint. In contrast to PPMx, the EPA distribution can accommodate a broader class of information to influence partitioning as one can always define pairwise similarity information from unit-specific covariates, but not all pairwise similarity information can be encoded as a similarity function $\mathcal{G}(\cdot)$ of unit-specific covariates, as required. For the Potts-Gibbs models, the MRF component which is responsible for the spatial dependence is imposed internally in the Gibbs-type priors, either DP, PY or MFM.

As mentioned in Chapter 2.2, the ddCRP and PPMx do not have an explicit pmf, and marginal properties are lost, such as the prior on the number of clusters. The EPA-distributed approach addresses these issues, easing prior specifications and MCMC developments. It has both an explicit formula for the distribution of the number of clusters and a known, tractable pmf. The Potts-Gibbs models also involve an intractable normalising constant, thus there is no closed-form solution for sampling from the posterior distribution directly. However, the Potts-Gibbs models have their advantages. The Potts-Gibbs models are above the EPA distribution in terms of computational efficiency. The computational cost of the Potts-Gibbs cluster assignment is $\mathcal{O}(pM)$ whereas the computational cost of the EPA sequential cluster assignment is $\mathcal{O}(p^2M)$, which can result in significant computational savings, for example in imaging applications where the number of pixels is large.

3.5 Prior simulation comparisons in finite samples

In the following, we study and compare the prior for the EPA distribution, Potts-DP, Potts-PY and Potts-MFM models. We empirically show how the combination of Potts and BNP components influence each other and how each component affects the prior on the number of clusters, the number of connected neighbours and cluster sizes. The random partition model involves a prior over the number of clusters of size 1 to p, i.e. over the random variables. We refer to this as prior distribution over $N_1, \cdots, N_p$ as the prior over cluster sizes.

In order to draw samples from the prior model, we run an MCMC algorithms for $T = 10,000$ iterations with the first 40% draws removed a burn-in. For the Potts-MFM model, as mentioned in Section 3.3.6, we need to specify $p_L(\cdot|\psi)$ which reflects the prior belief on the number of clusters. We employ the Poisson distribution for $p_L(\cdot|\psi)$ following the set-up in (Miller and Harrison, 2018).

3.5.1 Comparison of the prior on the number of clusters

We begin by empirically studying the prior on the number of clusters under the models studied in this chapter. As shown in Figure 3.1, both the concentration α and discount parameter δ parameters influence the expected number of clusters in a partition that adopts either EPA distribution or Potts-DP and Potts-PY priors. The higher the concentration parameter α and discount parameter δ, the higher the expected total number of clusters.

As discussed in Section 3.2, the prior on number of clusters for the EPA is available analytically and coincides with that of PY (or DP when $\delta = 0$). Indeed, we observe an approximately linear relationship between α and the expected number of clusters, and a non-linear relationship within δ, related to the power law behaviour.

By incorporating the Potts model in the random partition model (either MFM, DP or PY), which acts in opposition to the concentration parameter α and discount parameter δ, we observe that the coupling υ always reduces the expected total number of clusters. For the Potts-DP model, one can see that a large value of the concentration parameter α is required to avoid phase transition. Looking at the Potts-PY model, with the discount parameter δ, we can use smaller values of the concentration parameter α to avoid phase transition. Looking at the Potts-MFM model, we see that when the coupling υ equals zero, the expected number of prior clusters is roughly equal to λ. As the coupling υ increases, we need a larger value of γ to avoid phase transition. However, with the help of a larger value of λ, we can replace it with a smaller value of γ instead. For example, when the coupling $\upsilon = 0.5$ and $\lambda = 50$, we need $\gamma \geq 10$ to get away from phase transition. When we increase the λ to 100, we can use a smaller value of $\gamma \geq 5.0$.

Figure 3.2 shows the expected total number of clusters (black solid line) for the Potts-PY and Potts-MFM models with the lower bound (red dashed line) and the upper bound (blue dashed line) estimated from Equation (3.16) or Equation (3.20) respectively for increasing values of the α or λ respectively. We observe that the expected total number of clusters is between the estimated lower bound and the upper bounds from the defined equations.

3.5.2 Comparison of prior cluster sizes

Next, we empirically study the prior on cluster sizes under the different models. For the Potts-DP and Potts-PY models, the variability in the sizes of the cluster is large, varying widely from a lot of very small clusters to a very few large clusters as displayed in Figure 3.3. These clusters exhibit the "rich-getting-richer" and the "poor-getting-poorer" property. For the Potts-PY model, the introduction of the parameter δ helps to weaken the "rich-getting-richer" property by putting more prior weight on the probability of generating new clusters, thereby reducing the probability of adding a new unit to an existing cluster. However, PY still has some limitations. We see that increasing the discount parameter δ causes more new clusters to be created. The discount parameter δ does not alter the linear dependence on previous observations for cluster allocations - "rich-getting-richer" property, it still leads to apriori partitions containing a small number of large clusters. By including the Potts model in the random partition model (either MFM, DP or PY), we find that increasing the coupling υ helps those large clusters to grow even larger and form a partition with very few small components and one giant component.

We see that the Potts-MFM model offsets those limitations. It gets rid of the "rich-getting-richer" property. MFM tends to generate clustering with similar cluster sizes. We will not have the clustering that has clusters consisting of only a few units. Increasing the coupling υ will form a giant cluster, but with appropriate values of γ and λ, we can avoid the phase transition. A larger value of γ reduces the variability and decreases the uncertainty in the expected size of each cluster. However, one should be aware that a large value of γ might cause poor mixing and slower convergence as it might be stuck easily around the local maximum. We suggest using a larger value of λ instead of γ.

For the EPA model, we observe that when discount parameter δ is increased from 0.0 (Figure 3.3 (a)) to 0.25 (Figure 3.3 (b)) or 0.35 (Figure 3.3 (c)), the curves go upwards as the distribution of the size of small or medium clusters increases, consequently reducing the possibility of producing one giant cluster. When the concentration parameter α increases, the distribution of the size of small clusters increases, while the distribution of the size of large clusters drops.

3.5.3 Comparison of prior spatial connectivity

Lastly, we compare the prior on spatial connectivity under the different models. Specifically, we study the prior on the number of connected neighbours, along with a visualisation of the pairwise probability matrix that two units are clustered together and a draw from the prior. Specifically Figure 3.4 shows the expected number of connected neighbours as a function of the model hyperparameters under the different models, and Figures 3.5 - 3.8 plot the pairwise probabilities that two units are clustered together (right) along with a sampled partition (left) for the EPA distribution, Potts-DP, Potts-PY and Potts-MFM models respectively.

From Figure 3.4, it is apparent that increasing the temperature τ and coupling υ increase the expected number of like neighbour pairs in the partition. For the EPA, the temperature τ governs the degree to which the prior distance information influences the partition distribution. As shown in Figure 3.5, when τ increases, it increases the effect of the spatial distance information on the clustering structure, encouraging nearby units to cluster together; consequently only those units that have relatively small distance will have a higher probability to be clustered together.

For the Potts-Gibbs models, the behaviour is quite different compared to the EPA. The relationship in Figure 3.4 between the expected number of connected neighbours and the coupling υ is steeper. For the Potts-DP and Potts-PY models, the phase transition is evident if the value of the coupling υ is set too high, this results in a high number of like neighbour pairs and causes almost all units to be clustered in one giant cluster. For the Potts-MFM model, when γ is small ($\gamma = 1.0$), λ does not have any effect on the expected number of neighbour pairs. When γ starts increasing, λ starts to play a different role in the expected number of neighbour pairs, and the phase transition is not as steep when increasing γ or λ and in comparison to the Potts-DP and Potts-PY models. From Figures 3.6 - 3.8 as expected we observe that increasing υ helps to improve spatial smoothing.

3.6 Image segmentation

The previous section focused on a prior comparison of the models, which is useful to gain intuition on the prior over the spatial clustering structure and the role of the hyperparameters. In the following, we compare and study the models aposteriori for image segmentation tasks. The data used is obtained from the well-known Berkeley Segmentation Data Set 500 (BSDS500) benchmark (Arbelaez et al., 2010). This include 200 images for training, 100 images for validation, and 200 images for testing. Each image is labelled manually by at least 4 annotators. We focus on the 154 images from the BSDS500 dataset considered in Chatzis (2013); Chatzis and Tsechpenakis (2010) and Lü et al. (2020).

Before analysing, we carry out necessary preprocessing steps. Specifically, for each image, we assume the feature vector $\boldsymbol{x}_j$ at each spatial location $\boldsymbol{s}_j$, for $j = 1, \cdots, p$ is d-dimensional, i.e. $\boldsymbol{x}_j \in \mathbb{R}^d$. Each image is first segmented into approximately 1000 superpixels (i.e. $p \approx 1000$) (Mori, 2005). Figure 3.9 illustrates the superpixel grid obtained using the method of Mori (2005) on one image from the BSDS500. Feature vectors are computed at the superpixel level, comprising hue saturation value (HSV) colour information (3-dimensional) along with the values of the maximum response (MR) filter banks (8-dimensional) (i.e. $d = 11$).

We consider Gaussian likelihoods for each feature vector $\boldsymbol{x}_j$:

$$\boldsymbol{x}_j | \beta_{z_j}, z_j \sim \mathrm{N}(\boldsymbol{\mu}_{z_j}, \boldsymbol{\Sigma}_{z_j}), \qquad \text{for } j = 1, \cdots, p,$$

where $\beta_{z_j} = (\boldsymbol{\mu}_{z_j}, \boldsymbol{\Sigma}_{z_j})$. And we assume a joint normal-inverse-Wishart distribution over the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ of the Gaussian likelihoods:

$$\begin{aligned}
\boldsymbol{\Sigma} &\sim \mathrm{IW}_{\nu_0}(\Psi_0), \\
\boldsymbol{\mu}|\boldsymbol{\Sigma} &\sim \mathrm{N}(\boldsymbol{m}_0, \boldsymbol{\Sigma}/\kappa_0), \\
\mathrm{pr}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \mathrm{NIW}(\boldsymbol{m}_0, \kappa_0, \Psi_0, \nu_0),
\end{aligned}$$

where $\mathrm{IW}_{\nu_0}(\Psi_0)$ denotes inverse Wishart distribution with scale matrix Ψ_0 and degrees of freedom ν_0 and $\mathrm{NIW}(\boldsymbol{m}_0, \kappa_0, \Psi_0, \nu_0)$ denotes normal-inverse-Wishart distribution with mean $\boldsymbol{m}_0$, real parameter κ_0 (> 0), scale matrix Ψ_0 and degrees of freedom ν_0. The normal-inverse-Wishart priors are set empirically based on the data as $\boldsymbol{m}_0 =$ the mean across the image, $\kappa_0 < 1$ (i.e. $\kappa_0 = 0.01$), $\Psi_0 = \mathrm{Var[data]})/M^{2/d}$, $\nu_0 = d+2 = 13$ (Fraley and Raftery, 2007).

For each image, MCMC is performed to obtain the posterior distribution over the segmentation of the image. The cluster assignment is initialised by the k-means algorithm with the number of clusters equal to 10 (i.e. $M = 10$). The MCMC algorithm proceeds as follows, for $t = 1, \cdots, T$:

1. The partition is sampled according to the full conditional over the clustering:

$$\mathrm{pr}(\pi_p|\ldots) \propto \prod_{j=1}^{p} f(\boldsymbol{x}_j|\pi_p, \boldsymbol{\mu}_{z_j}, \boldsymbol{\Sigma}_{z_j}) \cdot \mathrm{pr}(\pi_p|\Theta), \tag{3.21}$$

   where Θ represents all the parameters of the random image partition models employed, either the EPA distribution or Potts-Gibbs image partition model. And the likelihood part can be simplified as follows:

$$\begin{aligned}
\log\left\{\prod_{j=1}^{p} f(\boldsymbol{x}_j|\pi_p, \boldsymbol{\mu}_{z_j}, \boldsymbol{\Sigma}_{z_j})\right\} &= \sum_{j=1}^{p} \log\left\{f(\boldsymbol{x}_j|\pi_p, \boldsymbol{\mu}_{z_j}, \boldsymbol{\Sigma}_{z_j})\right\} \\
&= -\frac{pd}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{p}\log\left\{\det\left(\boldsymbol{\Sigma}_{z_j}\right)\right\} \\
&\quad - \frac{1}{2}\sum_{j=1}^{p}(\boldsymbol{x}_j - \boldsymbol{\mu}_{z_j})^T \boldsymbol{\Sigma}_{z_j}^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_{z_j}).
\end{aligned}$$

   In order to sample from this full conditional, which iterates through each superpixel, updating its allocation conditioned on all others (further details are provided in Chapter 4).

2. For $m = 1, \cdots, M$, the full conditional of the cluster specific parameters, $\beta_m^* =$

$(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$, is available in closed form and given by:

$$\begin{aligned} \boldsymbol{\Sigma}_m &\sim \mathrm{IW}_{\nu_m}(\Psi_m), \\ \boldsymbol{\mu}_m | \boldsymbol{\Sigma}_m &\sim \mathrm{N}(\boldsymbol{m}_m, \boldsymbol{\Sigma}_m / \kappa_m), \end{aligned} \tag{3.22}$$

with parameters:

$$\begin{aligned} \kappa_m &= \kappa_0 + n_m, \\ \boldsymbol{m}_m &= \frac{\kappa_0 \boldsymbol{m}_0 + n_m \overline{\boldsymbol{x}}_m}{\kappa_m}, \\ \nu_m &= \nu_0 + n_m, \\ \Psi_m &= \Psi_0 + \sum_{j \in C_m} (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_m)(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_m)^T + \frac{\kappa_0 n_m}{\kappa_0 + n_m} (\overline{\boldsymbol{x}}_m - m_0)(\overline{\boldsymbol{x}}_m - m_0)^T \\ &= \Psi_0 + \sum_{j \in C_m} \boldsymbol{x}_j \boldsymbol{x}_j^T - n_m \overline{\boldsymbol{x}}_m \overline{\boldsymbol{x}}_m^T + \frac{\kappa_0 n_m}{\kappa_0 + n_m} (\overline{\boldsymbol{x}}_m - m_0)(\overline{\boldsymbol{x}}_m - m_0)^T, \end{aligned}$$

where $n_m = \sum_{j=1}^{p} \mathbb{1}_{z_j = m}$ and $\overline{\boldsymbol{x}}_m = 1/n_m \sum_{j=1}^{p} \boldsymbol{x}_j \mathbb{1}_{z_j = m}$.

The MCMC algorithm provides approximate posterior draws of the segmentation of the image describing posterior uncertainty in the segmentation given the imaging data. To obtain a single point estimate of the segmentation, we report the partition $\hat{\pi}_p$ which minimise the posterior variation of information. For the evaluation, the probabilistic adjusted rand index (PARI) and probabilistic rand index (PRI) are employed. Let $\{\pi_p^{0,1}, \cdots, \pi_p^{0,g}\}$ denote the set of ground truth images and $\hat{\pi}_p$ denotes the estimated segmentation map, the PARI is defined as

$$\mathrm{PARI} = \frac{1}{g} \sum_{i=1}^{g} \mathrm{ARI}(\hat{\pi}_p, \pi_p^{0,i}),$$

and the PRI is defined as

$$\mathrm{PRI} = \frac{1}{g} \sum_{i=1}^{g} \mathrm{RI}(\hat{\pi}_p, \pi_p^{0,i}).$$

3.6.1 Results on image segmentation

In this study, we compared the performance of four different random image partition models on the BSDS500 dataset: the EPA distribution and three Potts-Gibbs models (Potts-DP, Potts-PY and Potts-MFM models). Our results, shown in Table 3.2 and Figure 3.10, indicate that all four models have similar PARI scores, ranging from 0.292 to 0.296, and PRI scores, ranging from 0.755 to 0.772. The EPA distribution achieved the highest PRI score, while the Potts-MFM model had the highest PARI score.

When examining the estimated segmentations in Figure 3.10, we see that the models are

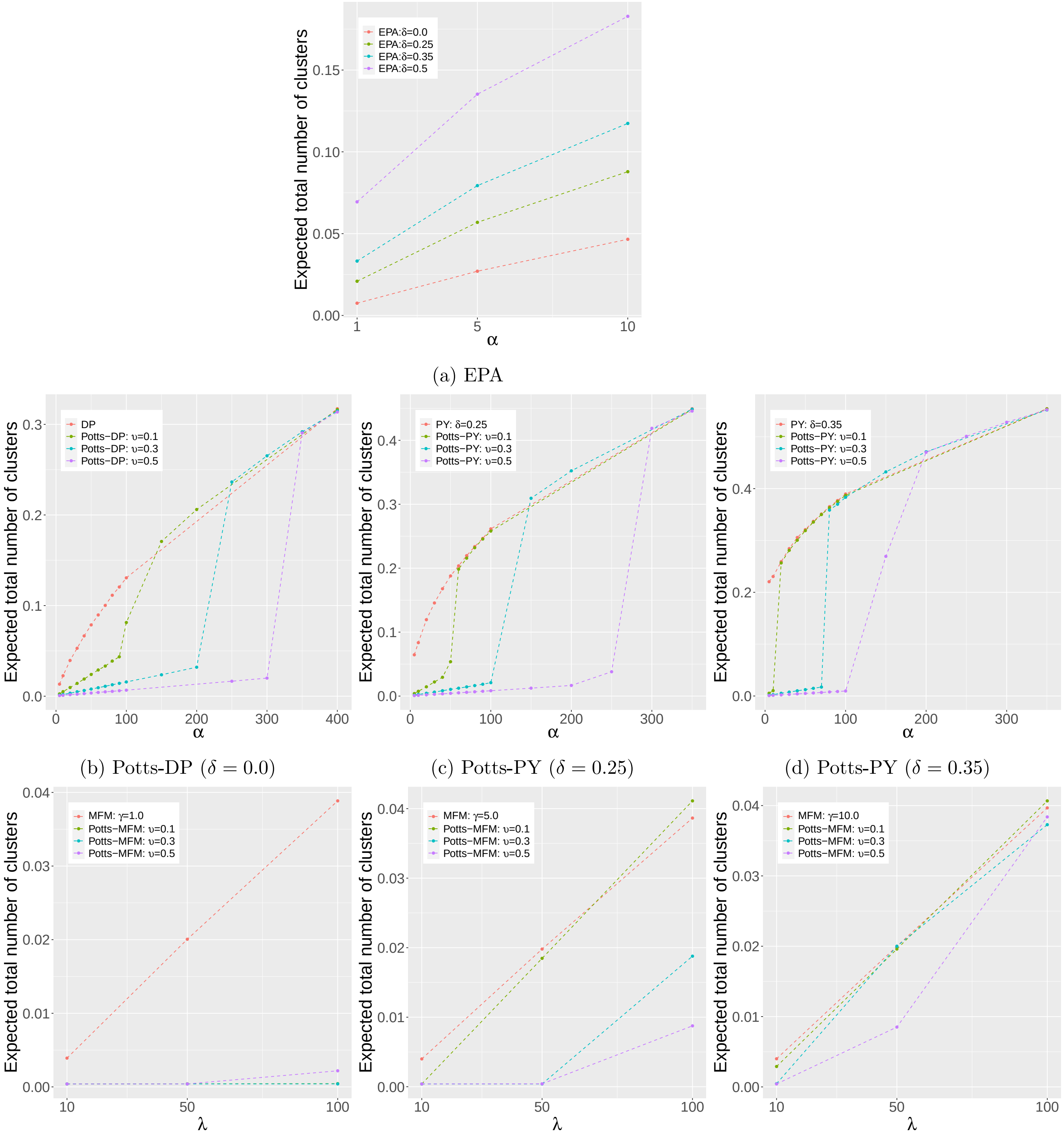
	PARI			PRI		
	Mean	Median	SD	Mean	Median	SD
EPA	0.293	0.284	0.123	0.772	0.784	0.070
Potts-DP	0.292	0.270	0.129	0.756	0.760	0.072
Potts-PY	0.292	0.273	0.132	0.755	0.762	0.073
Potts-MFM	0.296	0.295	0.123	0.767	0.781	0.072

Table 3.2: Results on the BSDS500 dataset. Summary statistics of the PARI and PRI score over the selected 154 images from BSDS500 studies by Chatzis (2013); Chatzis and Tsechpenakis (2010) and Lü et al. (2020).

generally effective at identifying the main components of the images. However, there is still room for improvement in terms of fine segmentation. By comparing the prior simulations in Figures 3.6 - 3.8 with the posterior estimates in Figure 3.10, we can appreciate small differences due to the influence of the prior. Overall, these results demonstrate the potential of using spatial random image partitions for image segmentation tasks.

3.7 Conclusions

In this chapter, we have detailed the two random image partition models, namely the Ewens-Pitman attraction (EPA) distribution and the Potts-Gibbs random partition (Potts-Gibbs) models. We have provided a thorough comparison including both theoretical properties and empirical comparisons. We focus mainly on the expected number of clusters, the size of clusters and the number of connected neighbour pairs induced from the underlying partition structure in these models. From the prior simulations conducted, we observe that the EPA distribution is well-behaved compared to the Potts-Gibbs models. Moreover, a closed-form prior is available for the random partition model, allowing full Bayesian inference for the key hyperparameters influencing the spatial clustering structure. For the Potts-Gibbs models, if we do not specify the values of the parameters wisely, we are likely to get undesirable clustering structures. The Potts-Gibbs models are sensitive to the chosen coupling v due to the phase transition that occurs. The Potts-DP and Potts-PY models are sensitive to the concentration parameter α because of the “rich-getting-rich” property. Among the Potts-Gibbs models, the phase transition of the Potts-MFM model is less extreme and easier to control. However, the main advantage of Potts-Gibbs models over the EPA is the reduced computational complexity resulting from the Markov property. In the next chapter, we construct novel scalar-on-image regression models which use the spatial random partition models studied in this chapter as the main building block.



(a) EPA

(b) Potts-DP ($\delta = 0.0$)

(c) Potts-PY ($\delta = 0.25$)

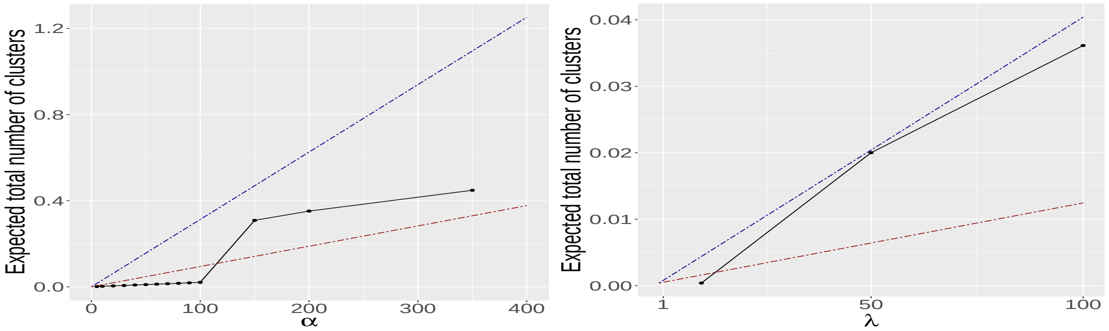
(d) Potts-PY ($\delta = 0.35$)

(e) Potts-MFM ($\gamma = 1.0$)

(f) Potts-MFM ($\gamma = 5.0$)

(g) Potts-MFM ($\gamma = 10.0$)

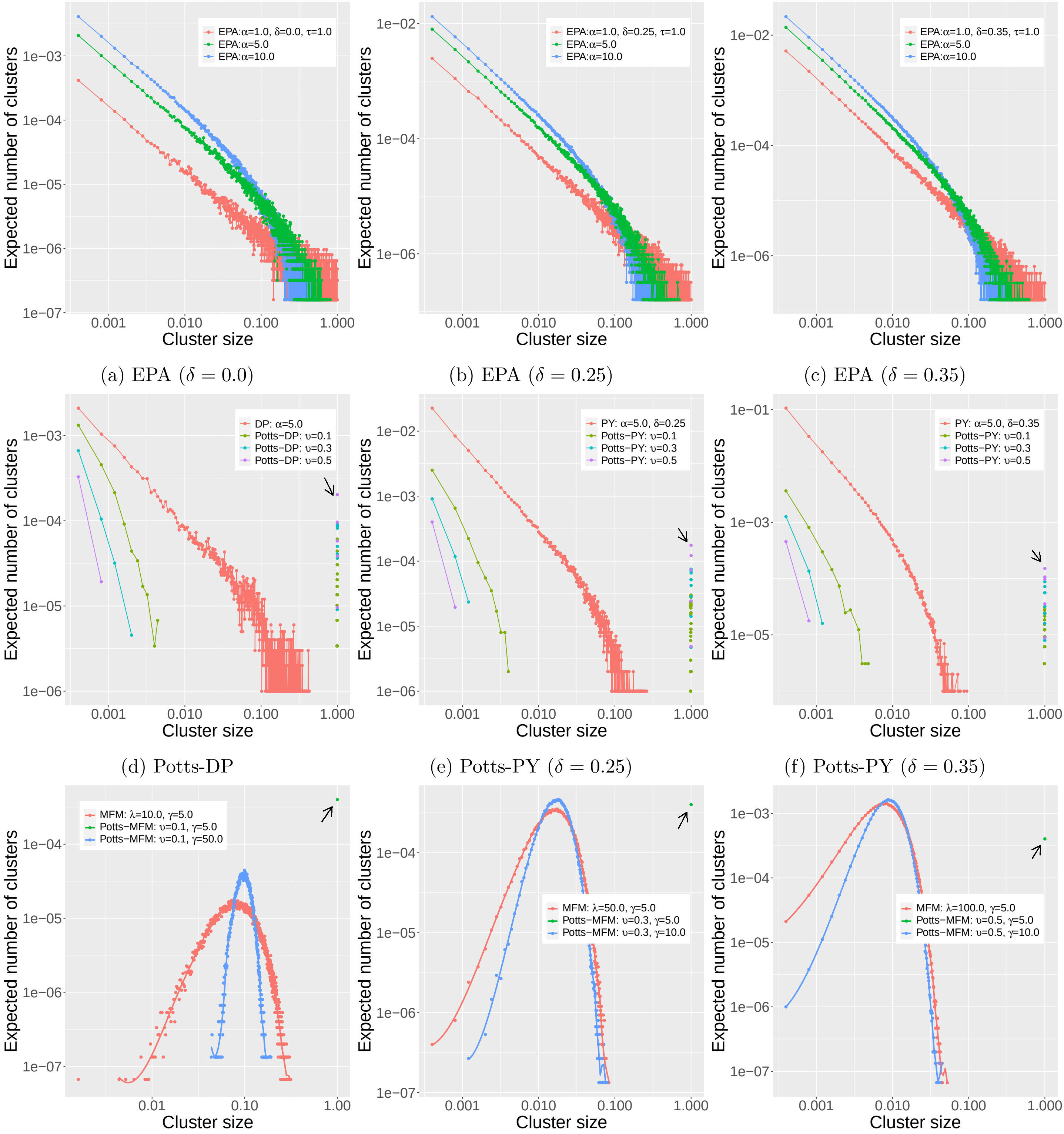
Figure 3.1: **Grid size = 50x50:** Expected number of clusters as a function of parameter α for the EPA distribution, Potts-DP and Potts-PY models or λ for the Potts-MFM model. Note the y-axis is normalised by the total number of units, p.



(a) Potts-PY ($\delta = 0.25, \upsilon = 0.3$)

(b) Potts-MFM ($\gamma = 10.0, \upsilon = 0.3$)

Figure 3.2: **Grid size = 50x50:** Expected total number of clusters (black solid line) for the **Potts-PY** and **Potts-MFM** model with the lower bound (red dashed line) and the upper bound (blue dashed line) estimated from Equation (3.16) or Equation (3.20) respectively for increasing values of the α or λ respectively. Note the y-axis is normalised by the total number of units, p.



(a) EPA ($\delta = 0.0$)

(b) EPA ($\delta = 0.25$)

(c) EPA ($\delta = 0.35$)

(d) Potts-DP

(e) Potts-PY ($\delta = 0.25$)

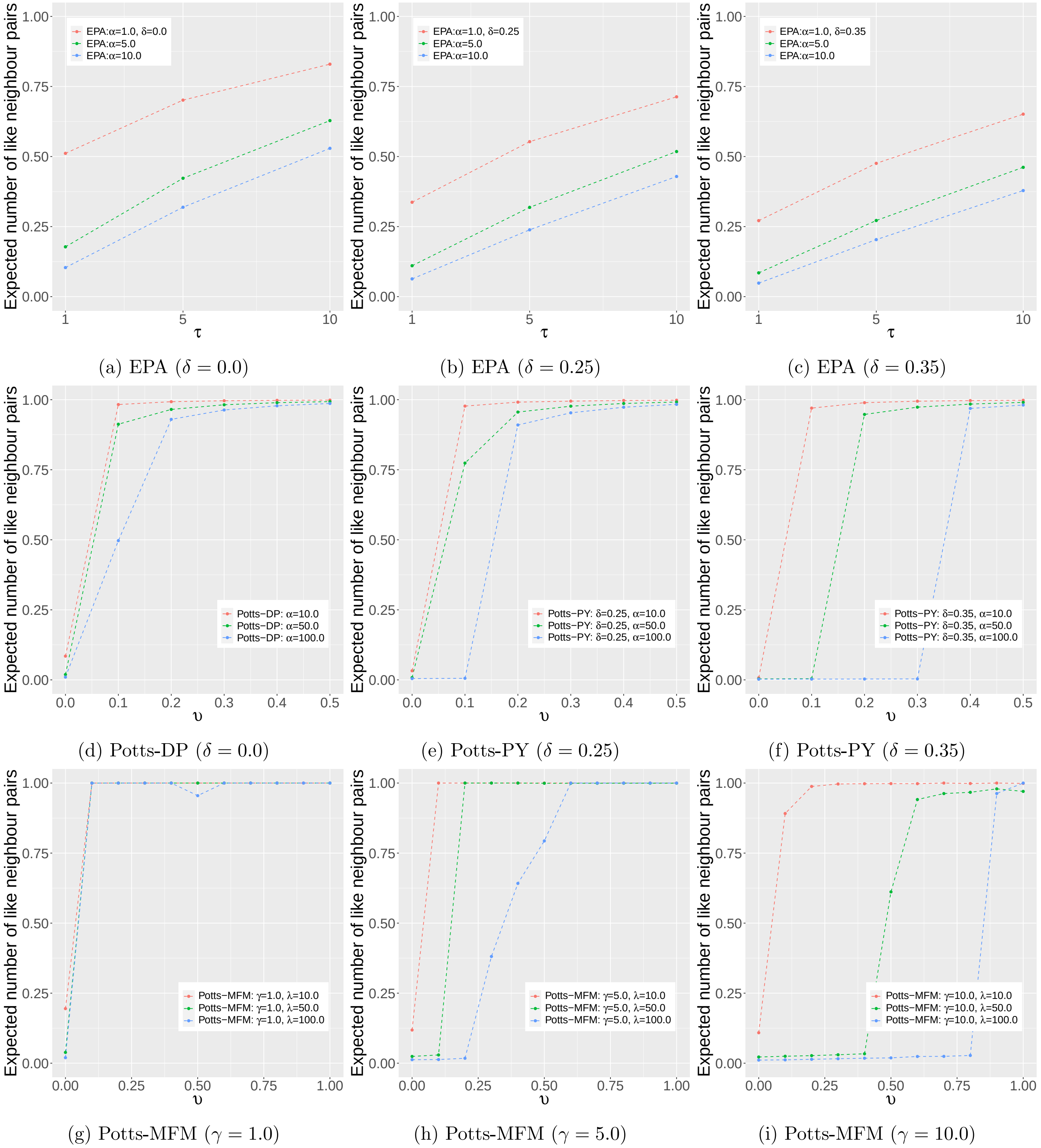
(f) Potts-PY ($\delta = 0.35$)

(g) Potts-MFM ($\lambda = 10.0$)

(h) Potts-MFM ($\lambda = 50.0$)

(i) Potts-MFM ($\lambda = 100.0$)

Figure 3.3: **Grid size = 50x50:** Expected number of clusters as a function of empirical cluster size for the EPA distribution, Potts-DP, Potts-PY models and Potts-MFM model. Note that both axes are normalised by the total number of units, p. Axes are plotted on the log scale.



(a) EPA ($\delta = 0.0$)

(b) EPA ($\delta = 0.25$)

(c) EPA ($\delta = 0.35$)

(d) Potts-DP ($\delta = 0.0$)

(e) Potts-PY ($\delta = 0.25$)

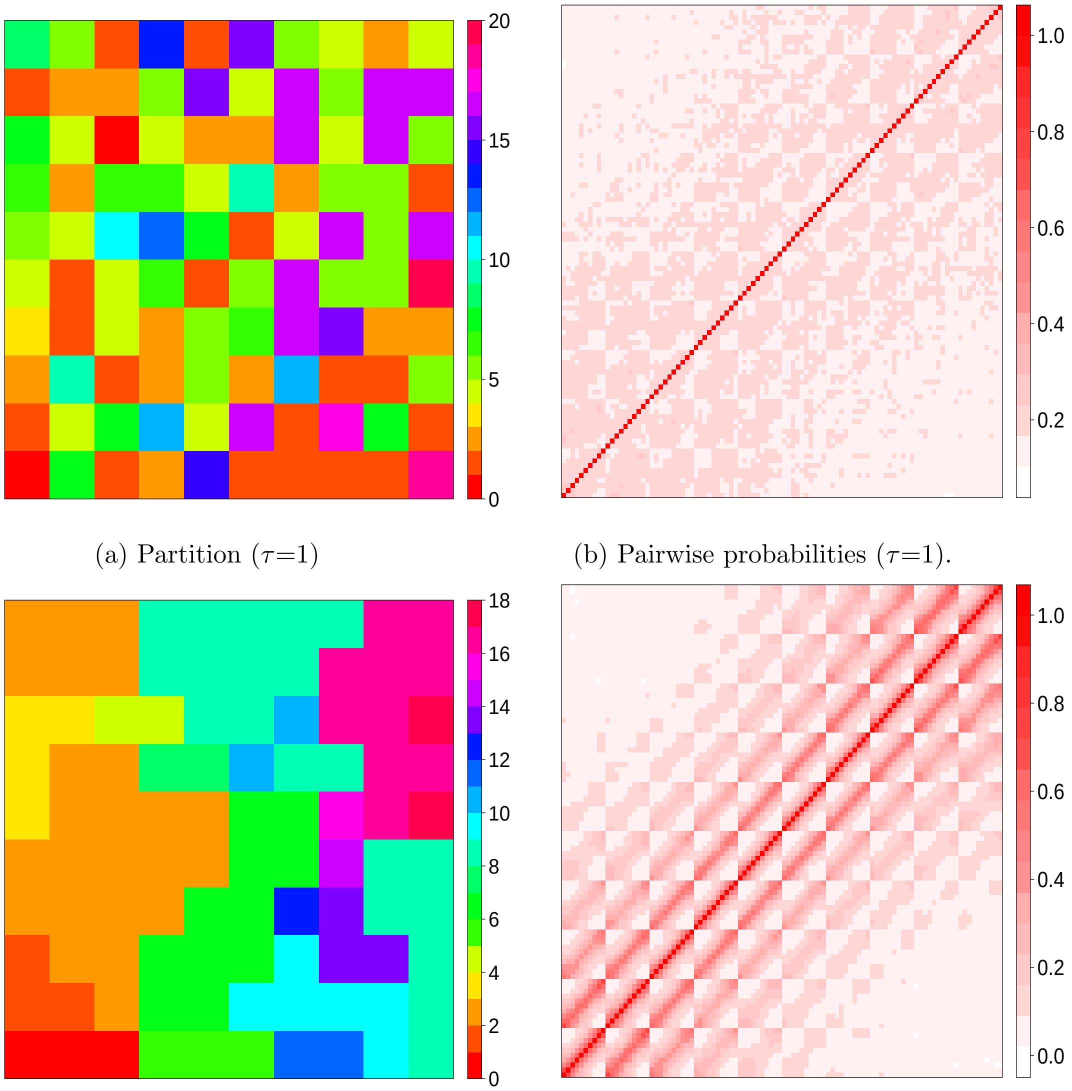
(f) Potts-PY ($\delta = 0.35$)

(g) Potts-MFM ($\gamma = 1.0$)

(h) Potts-MFM ($\gamma = 5.0$)

(i) Potts-MFM ($\gamma = 10.0$)

Figure 3.4: **Grid size = 50x50:** Expected number of neighbour pairs as a function of parameter α in proportion for the EPA distribution, Potts-DP and Potts-PY models or γ for the Potts-MFM model. Note the y-axis is normalised by the total number of neighbour pairs, S.



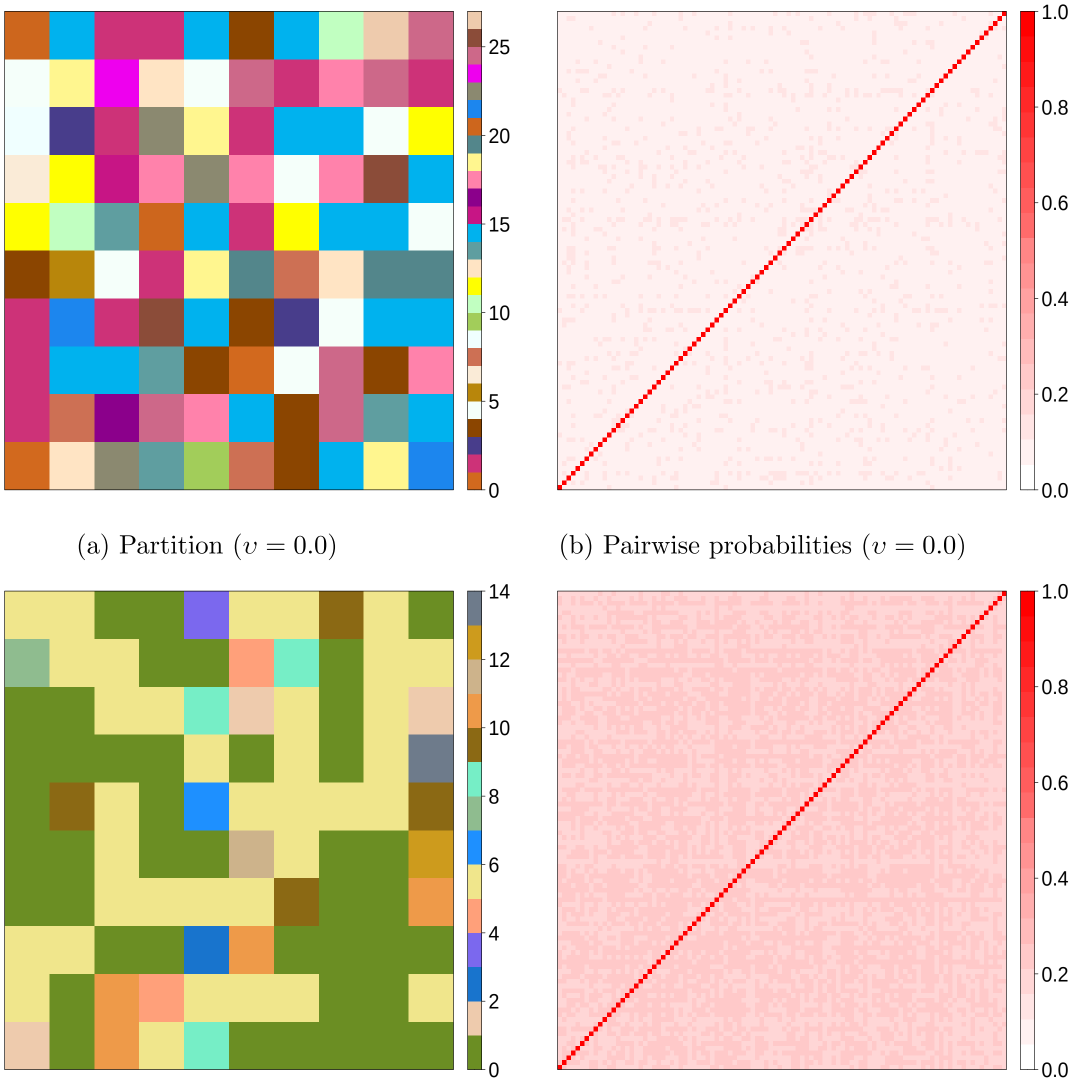
(a) Partition (τ=1)

(b) Pairwise probabilities (τ=1).

(c) Partition (τ=10)

(d) Pairwise probabilities (τ=10)

Figure 3.5: Pairwise probabilities that two units are clustered together (right) along with a sampled partition (left) for the **EPA distribution** with $\alpha = 5.0$ and $\delta = 0.0$ on a 10×10 regular grids. The top figure corresponds with $\tau = 1$, while the bottom $\tau = 10$



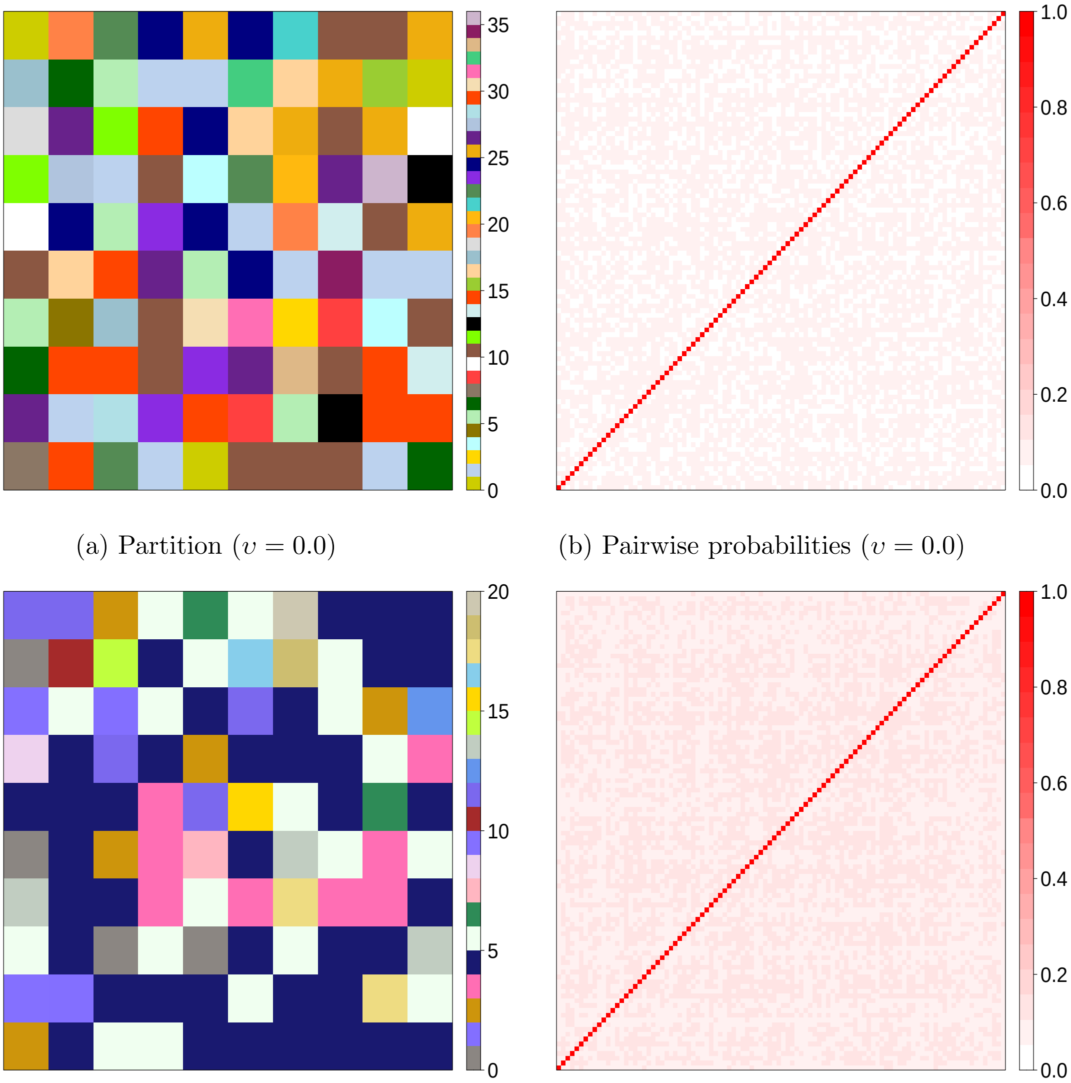
(a) Partition ($\upsilon = 0.0$)

(b) Pairwise probabilities ($\upsilon = 0.0$)

(c) Partition ($\upsilon = 0.1$)

(d) Pairwise probabilities ($\upsilon = 0.1$)

Figure 3.6: Pairwise probabilities that two units are clustered together (right) along with a sampled partition (left) for the **Potts-DP** model with $\alpha = 10.0$ and $\delta = 0.0$ on a 10×10 regular grids. The top figure corresponds with $\upsilon = 0.0$, while the bottom $\upsilon = 0.1$.



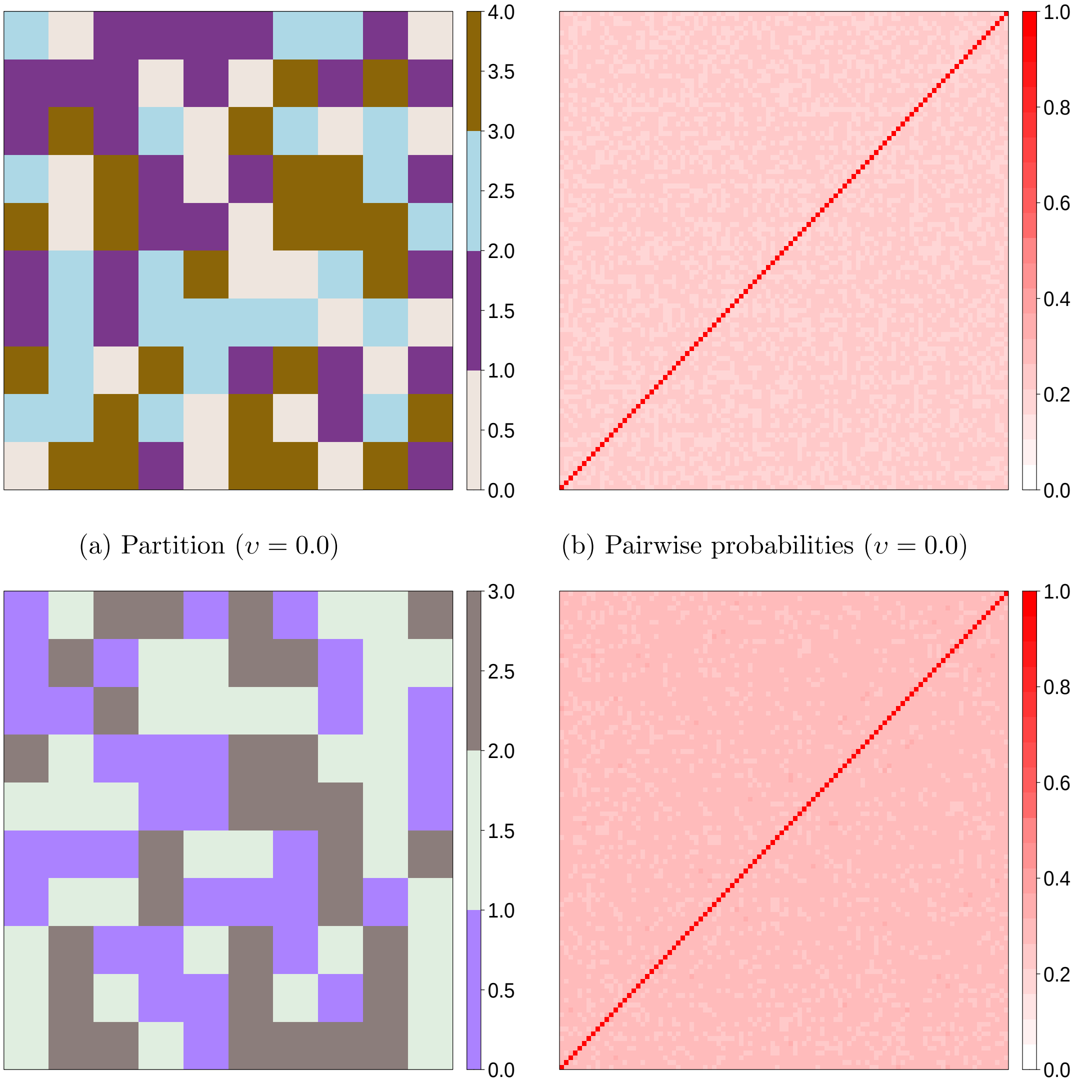
(a) Partition ($\upsilon = 0.0$)

(b) Pairwise probabilities ($\upsilon = 0.0$)

(c) Partition ($\upsilon = 0.1$)

(d) Pairwise probabilities ($\upsilon = 0.1$)

Figure 3.7: Pairwise probabilities that two units are clustered together (right) along with a sampled partition (left) for the **Potts-PY** model with $\alpha = 10.0$ and $\delta = 0.25$ on a 10×10 regular grids. The top figure corresponds with $\upsilon = 0.0$, while the bottom $\upsilon = 0.1$.



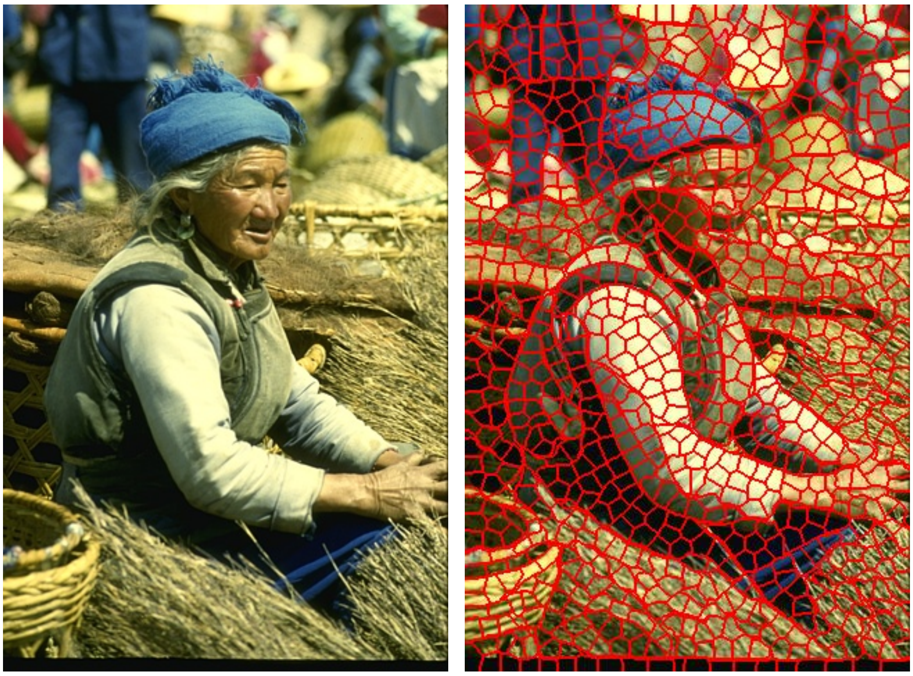
(a) Partition ($\upsilon = 0.0$)

(b) Pairwise probabilities ($\upsilon = 0.0$)

(c) Partition ($\upsilon = 0.1$)

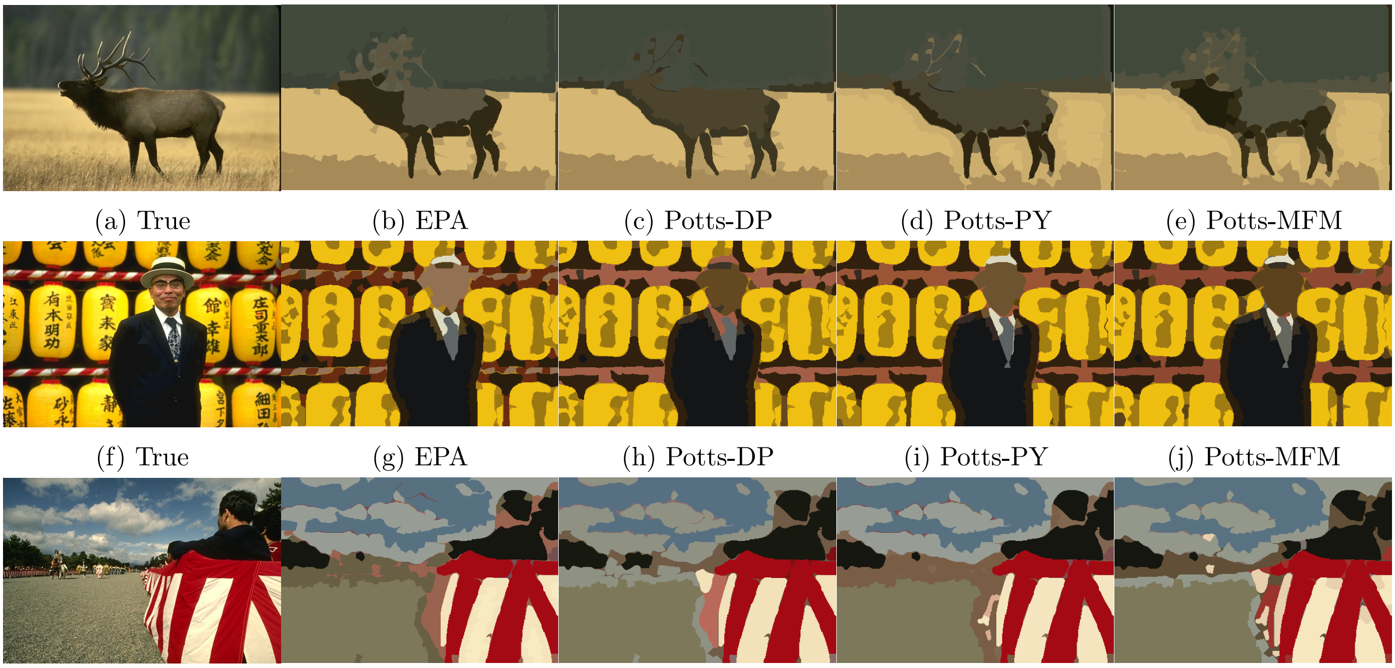
(d) Pairwise probabilities ($\upsilon = 0.1$)

Figure 3.8: Pairwise probabilities that two units are clustered together (right) along with a sampled partition (left) for the **Potts-MFM** model with $\lambda = 1.0$ and $\gamma = 10.0$ on a 10×10 regular grids. The top figure corresponds with $\upsilon = 0.0$, while the bottom $\upsilon = 0.1$.



(a) Original image (b) Image with superpixel grid

Figure 3.9: An illustration of the superpixel grid obtained using the method by Mori (2005) for one image from the BSDS500 dataset. Superpixel boundaries are depicted in red.



(a) True (b) EPA (c) Potts-DP (d) Potts-PY (e) Potts-MFM

(f) True (g) EPA (h) Potts-DP (i) Potts-PY (j) Potts-MFM

(k) True (l) EPA (m) Potts-DP (n) Potts-PY (o) Potts-MFM

Figure 3.10: Qualitative results on the BSDS500 dataset. Some visual image segmentation results obtained by the EPA distribution and Potts-Gibbs models (Potts-DP, Potts-PY and Potts-MFM models.)

Chapter 4

SIR with Random Image Partition Models

Under the SIR framework, we develop a novel class of priors combining sparsity-promoting priors on the coefficient image with spatial random partition models. We organise the description of our proposed models into two parts: the spatially-dependent partition process and the shrinkage priors. We utilise the random image partition models described in Chapter 3, specifically the Potts-Gibbs random partition models and Ewens-Pitman attraction distribution to spatially cluster the coefficients, in order to improve the signal and interpretability as well as ease computations and multicollinearity. Each cluster has a unique coefficient and we employ heavy-tailed shrinkage to penalize and identify relevant regions, specifically using t-shrinkage priors. For posterior inference, we develop a Gibbs sampler to simulate from the posterior using a generalized Swendsen-Wang (GSW) algorithm (Da Xu et al., 2016) to draw samples from the Potts-Gibbs SIR models and a Metropolis-Hastings within Gibbs for the EPA SIR model.

Note: *An article on Potts-Gibbs SIR models has been accepted and will shortly appear in the book of proceedings of the Bayesian Young Statisticians Meeting 2021 (BAYSM) in the series Springer Proceedings in Mathematics & Statistics. We are also preparing an extended version of this paper, including some empirical and theoretical results studying the properties of the priors included to be submitted to Biometrics. Furthermore, we plan to open-source the code of our proposed models.*

4.1 Introduction

Scalar-on-image regression (SIR) inherently faces non-identifiability problems (Palma et al., 2020) because of the large p small n setting and potentially strong spatial connection across the imaging predictor. A variety of different methods have been proposed by the SIR community to attenuate the issue of non-identifiability by making structural assumptions on the coefficient image, $\boldsymbol{\beta}$, for instance, imposing combinations of spatial, smoothness, sparsity or projections onto a subspace (Goldsmith et al., 2014; Huang et al., 2013; Li et al., 2015; Smith and Fahrmeir, 2007; Wang et al., 2017).

In neuroimaging applications, the effect size of each predictor on the outcome measure is typically anticipated as too small to be reliably informative about brain function, if not zero, but the cumulative effect size of predictors belonging to the same cluster can be substantial. On top of that, the image predictors are observed on a spatially structured coordinate system, thus neighbouring pixels tend to exhibit some spatial dependencies. In spatial clustering, exchangeability is indeed no longer the proper assumption as the spatial data or images contain covariate information, that we wish to leverage to improve model performance in this high-dimensional setting. It is therefore natural to assume certain structural assumptions on the spatial partition for our proposed SIR framework. In particular, we consider clustering the coefficient image to form several spatially contiguous regions to efficiently reduce the dimension of the parameter space. In this thesis, we extend the Ewens-Pitman attraction (EPA) distribution and the Potts-Gibbs random partition (Potts-Gibbs) models as the prior of the random partition distribution for our proposed models: EPA SIR and Potts-Gibbs SIR models respectively to enforce spatially dependent clustering on the coefficient image, thereby producing groups represent spatially contiguous regions and derive marginal properties in this setting.

Many modern applications involve high-dimensional datasets, in some cases, with $p \gg n$, including magnetic resonance imaging (MRI) and gene expression data. Therefore, sparsity plays a crucial role in a high-dimensional linear regression problem. To make inference in this context, it has commonly been assumed many of the covariates are small enough to be insignificant and irrelevant, intending to remove those covariates that have very little or non-existent effects on the response from the regression model or shrink them towards zero. Thus, in high-dimensional settings, it is frequently assumed that the coefficients are likely to be sparse to narrow the solution space.

Variable selection is one of the key tasks to identify the small subset of significant regression coefficients that influence a response so that a significant portion of the variation in the response can be inferred from these predictors. There have been many methods proposed from the Bayesian perspective. They commonly encourage the sparsity of the regression coefficients by choosing an appropriate prior distribution. These priors including spike-and-slab priors with point masses at zero (Castillo et al., 2015; Martin et al., 2017; Yang et al., 2016), continuous spike-and-slab priors (George and McCulloch, 1993; Ročková and George, 2018), scale-mixture shrinkage priors (Song and Liang, 2017; Van Der Pas et al., 2016) and non-local priors (Rossell and Telesca, 2017; Shin et al., 2018). In this thesis, with the theoretical support from Song and Liang (2017), we use a class of heavy-tailed priors to identify relevant regions, specifically using t-shrinkage prior which will be explained in Section 4.2.3.

The rest of this chapter is structured as follows. In Section 4.2 we provide a clear description of the design of the construction of the proposed models including the generalized linear model (GLM) in Section 4.2.1, image partition models in Section 4.2.2, shrinkage priors in Section 4.2.3 and priors for additional parameters in Section 4.2.4. It is then followed by Section 4.3 which gives suitable computational strategies for posterior inference. Section 4.4 includes various posterior summaries to describe the posterior quantities of interest. Section 4.5 provides concluding remarks.

4.2 Model

Our proposed models: the EPA SIR model and Potts-Gibbs SIR models are motivated by applications in brain imaging data: automatic identifying of brain regions to diagnose Alzheimer's disease (AD). In the statistical literature, diagnosing AD based on neuroimages can be framed as a SIR problem (Reiss et al., 2011), so-called as the responses are scalars as in a typical regression but the covariate is the entire image. In the following, we construct our proposed models under the SIR framework which we have discussed in Section 2.1. In particular, the proposed models aim to group together pixels with similar effects on the response to have a common coefficient. Moreover, they directly account for the spatial location within the cluster allocation to incorporate and provide interpretable feature extraction. We outline the proposed models with the three main components: the generalized linear model (GLM), the random image partition models (Chapter 3) and the shrinkage priors which are explained in the subsequent subsections.

4.2.1 Generalized linear model

The SIR model in Equation (2.1) can be extended for other types of responses through a generalized linear model (GLM) (McCullagh and Nelder, 2019). A GLM introduced by Nelder and Wedderburn (1972) generalises linear regression allowing the response to have a non-normal distribution. A GLM has three main components: a specified distribution belonging to the exponential family, a linear predictor and a link function. Each response y_i assumed to follow the specified distribution belonging to the exponential family with probability mass function (pmf) for discrete variables or probability density function (pdf) for continuous variables:

$$f(\boldsymbol{y};\theta) = \exp\left\{\frac{\boldsymbol{y}\mathcal{A}(\theta) - \mathcal{B}(\theta)}{\mathcal{C}(\varphi)} + \mathcal{D}(\boldsymbol{y},\varphi)\right\},$$

where θ is the parameter of the exponential family and φ is the scale parameter. The distribution is said to be in canonical form (or natural form) if $\mathcal{A}(\theta)$ is the identity function, and θ is commonly referred to as the canonical parameter (or natural parameter). The $\mathcal{B}$, $\mathcal{C}$ and $\mathcal{D}$ are known functions specific to the distribution within the exponential family. A linear predictor is used to determine the canonical parameter θ through a series of transformations. A link function $\mathcal{G}(\cdot)$ connects together the mean

$$u_i = \mathbb{E}[y_i|\boldsymbol{x}_i, \boldsymbol{w}_i],$$

and the linear component $\boldsymbol{w}_i^T\boldsymbol{\mu} + \boldsymbol{x}_i^T\boldsymbol{\beta}$. Often, the natural parameter θ is used to relate the mean u_i to the linear component $\boldsymbol{w}_i^T\boldsymbol{\mu} + \boldsymbol{x}_i^T\boldsymbol{\beta}$:

$$\mathcal{G}(u_i) = \mathcal{A}(\theta) = \theta = \boldsymbol{w}_i^T\boldsymbol{\mu} + \boldsymbol{x}_i^T\boldsymbol{\beta}.$$

In the following, we provide three examples of GLMs to account for three different forms of the outcomes $\boldsymbol{y}$ (continuous, categorical and ordinal). These will form the cases studied in Chapter 5 but the framework can of course be extended for other types of responses.

Gaussian. We model the continuous outcomes $\boldsymbol{y}$ by considering a GLM with normal distribution and identity link function, which gives us the linear regression model. For each data point, $y_1, \ldots, y_n$, we have

$$y_i|\boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\mu}, \boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}(u_i, \sigma^2), \tag{4.1}$$

where we define $u_i = \boldsymbol{w}_i^T\boldsymbol{\mu} + \boldsymbol{x}_i^T\boldsymbol{\beta}$.

Binary. We model the binary outcomes $\boldsymbol{y}$ ($y_i \in \{0, 1\}$) using a Bernoulli distribution and either logistic or probit link function where the link function maps (0, 1) to the real line. For each data point, $y_1, \ldots, y_n$, we have

$$y_i|\boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\mu}, \boldsymbol{\beta} \sim \text{Bern}\left(\mathcal{G}^{-1}(u_i)\right),$$

where $\mathrm{pr}(y_i = 1|\boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\mu}, \boldsymbol{\beta}) = \mathbb{E}[y_i|\boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\mu}, \boldsymbol{\beta}] = \mathcal{G}^{-1}(u_i)$ and $\text{Bern}(a)$ denotes a Bernoulli distribution with probability a. For the logistic link function,

$$\mathrm{pr}(y_i = 1|\boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\mu}, \boldsymbol{\beta}) = \frac{1}{1 + \exp(u_i)}.$$

For the probit link function,

$$\mathrm{pr}(y_i = 1|\boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\mu}, \boldsymbol{\beta}) = \Phi\left(u_i\right),$$

where $\Phi(\cdot)$ is s a cumulative distribution function. In this case, the model can be equivalently formulated through a latent response $\tilde{y}_i$ that is Gaussian distributed with mean u_i and unit variance. In particular, $\tilde{y}_i|u_i \sim \mathrm{N}(u_i, 1)$ and

$$y_i = \begin{cases} 0 & \text{if } \tilde{y}_i \leq 0 \\ 1 & \text{if } \tilde{y}_i > 0. \end{cases}$$

The probit model is recovered by marginalising the latent $\tilde{y}_i$.

Ordinal. We model the ordinal outcomes $\boldsymbol{y}$ taking ordered values $c = 0, \ldots, C$ through a categorical distribution and with either an ordered logistic or probit link function, that is,

$$\mathrm{pr}(y_i \leq c|\boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\mu}, \boldsymbol{\beta}, b_c) = \mathcal{G}^{-1}(b_c - u_i),$$

where $0 = b_0 < b_1 < \ldots < b_{C-1}$ represent the cutoffs and the link function maps $(0, 1)$ to the real line. For the logistic link function,

$$\mathrm{pr}(y_i \leq c | \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\mu}, \boldsymbol{\beta}, b_c) = \frac{\exp(b_c - u_i)}{1 + \exp(b_c - u_i)}.$$

For the probit link function,

$$\mathrm{pr}(y_i \leq c | \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\mu}, \boldsymbol{\beta}, b_c) = \Phi\left(b_c - u_i\right). \tag{4.2}$$

In this case, the model can be equivalently formulated through a latent response $\tilde{y}_i$ that is Gaussian distributed with mean u_i and unit variance. In particular, $\tilde{y}_i | u_i \sim \mathrm{N}(u_i, 1)$ and

$$y = \begin{cases} 0 & \text{if } \tilde{y} \leq 0 \\ c & \text{if } b_{c-1} < \tilde{y} \leq b_c \\ C & \text{if } \tilde{y} > b_{C-1}. \end{cases}$$

The ordered probit model is recovered by marginalising the latent $\tilde{y}$.

4.2.2 Image partition models

We model the high-dimensional coefficient image, $\boldsymbol{\beta}$, by spatially clustering the units into M regions and assuming common coefficients $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_M^*)^T$ within in each cluster, i.e. $\beta_j = \beta_m^*$ given the cluster label $z_j = m$. Thus, the prior on the coefficient image is decomposed into two parts: the random image partition model for spatially clustering the units and shrinkage prior for the cluster-specific coefficients $\boldsymbol{\beta}^*$. Following from Chapter 3, we consider two different random image partition models: EPA distribution and Potts-Gibbs models. The nonparametric approach avoids prespecifing the number of regions and incorporates the spatial information directly into the clustering (as opposed to the Ising-DP model which may result in scattered clusters throughout the image). The clustering procedure helps to improve signal and interpretability and ease computations and multicollinearity. By doing this, it allows automatic detection of regions, along with uncertainty; and we use the notation $\boldsymbol{x}_i^* = (x_{i1}^*, \cdots x_{im}^*)$ to denote the extracted features from the ith image for each of the regions $m = 1, \cdots, M$ defined by π_p. Moreover, it allows for sharp discontinuities in the coefficient image across regions, which may be relevant in medical applications to capture irregularities (Wang et al., 2017).

4.2.3 Shrinkage priors

Song and Liang (2017) presented the asymptotic behaviour of a general class of continuous shrinkage priors in an ordinary high-dimensional linear regression model. The theoretical discovery found that the continuous shrinkage priors can achieve nearly the same posterior contraction rate and variable selection consistency as the widely used spike-and-slab priors for recovering the model parameters and the valid subset of covariates in the model. On top of that, the shrinkage priors are computationally more efficient than the spike-and-slab priors, especially when conjugacy exists. It is worth mentioning that the developed theory is mostly dependent upon the concentration and tail properties of the density of the continuous shrinkage prior.

For each unique value of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ of each cluster, which is reparameterised as $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_M^*)^T$, we apply a class of heavy-tailed priors (with the theoretical support from Song and Liang (2017)) to identify relevant regions. Specifically, a t-shrinkage prior is used for the base measure F_0 to specify a prior for the unique value β^* of each cluster:

$$(\beta_m^*)\,|\sigma^2 \overset{i.i.d}{\sim} t_{df}(s\sigma), \quad \text{for all } m = 1, \ldots, M, \tag{4.3}$$

where $t_{df}(s)$ denotes t-distribution with degree of freedom df and scale parameter s.

For posterior inference, the t distribution (Equation (4.3)) can be rewritten as a hierarchical inverse-gamma scaled Gaussian mixture:

$$\begin{aligned} \eta_m^* &\sim \text{IG}\left(a_\eta, b_\eta\right), \\ (\beta_m^*)\,|\sigma^2, \eta_m^* &\sim N(0, \eta_m^*\sigma^2), \quad \text{for all } m = 1, \ldots, M, \end{aligned} \tag{4.4}$$

where $a_\eta > 0$ and $b_\eta > 0$, respectively are the shape and scaling parameter of the mixing distribution for each η_m^* corresponds to degrees of freedom $df = 2a_\eta$ and scale $s = \sqrt{b_\eta/a_\eta}$ for the t-shrinkage prior. This representation can also be viewed as a global-local shrinkage prior, where b_η controls the overall shrinkage towards the origin, whereas η_m^* are the local shrinkage parameters, allowing each β_m^* deviate with different levels of shrinkage. The representation of the t-shrinkage prior in Equation (4.4) provides a convenient way to implement posterior sampling.

The selection of scale parameter, s or b_η, is critical for the performance of Bayesian variable selection to avoid not over-shrinking or under-shrinking. Choosing an excessively large scale parameter weakens the shrinkage effects; thus, it might fail to shrink some η_m^* towards 0. On the other hand, choosing a scale parameter that is too small may cause many η_m to be aggressively shrunk to 0, which might accidentally wipe out the effects of important predictors. Theorem 3.1 in Song and Liang (2017) shows that one could set the scale parameters $s^2 \simeq b_\eta \simeq 1/\{n\log(\mathfrak{p})\mathfrak{p}^{-2c}\}$, for a sufficiently large value of c, where $\mathfrak{p}$ denotes the number of features, i.e. $\mathfrak{p} = p$ in the standard high-dimensional linear regression framework studied in Song and Liang (2017) and $\mathfrak{p} = M$ in our proposed SIR model. For the student-t distribution, we set the degree of freedom to be 3, i.e. $a_\eta = 1.5$. Additionally, due to the important role of b_η, we put a hyperprior on b_η, i.e. $b_\eta \sim \text{G}(a_o, b_o)$, where $\text{G}(a_o, b_o)$ denotes a gamma distribution with shape a_o and rate b_o. We specify the a_o and b_o to achieve the expected value roughly equivalent to the recommended scale parameter $s^2 \approx b_\eta \approx 1/\{n\log(M)M^{-2c}\}$, where c ranges from -0.25 to 1.10.

Thus, the prior for $\boldsymbol{\beta}^*$ is specified hierarchically as follows to penalize and identify the relevant regions:

$$\begin{aligned}\boldsymbol{\beta^*}|\boldsymbol{\eta^*},\sigma^2 &\sim \mathrm{N}(\mathbf{0}_M,\sigma^2\boldsymbol{\Sigma}_{\beta^*}),\\ \eta^*_m|b_\eta &\sim \mathrm{IG}(a_\eta,b_\eta),\\ b_\eta &\sim \mathrm{G}(a_o,b_o),\end{aligned} \tag{4.5}$$

where $\boldsymbol{\Sigma}_{\beta^*}=\mathrm{diag}(\eta^*_1,\ldots,\eta^*_M)$.

4.2.4 Additional parameters

For the parameter $\boldsymbol{\mu}$ involved in Equation (4.1), we assign the prior distribution as follows:

$$\boldsymbol{\mu}|\sigma^2 \sim \mathrm{N}(\boldsymbol{m}_\mu,\sigma^2\boldsymbol{\Sigma}_\mu), \tag{4.6}$$

where $\boldsymbol{m}_\mu=(m_{\mu_1},\ldots,m_{\mu_q})$, $\mathbf{c}_\mu=(c_{\mu_1},\ldots,c_{\mu_q})$, and $\boldsymbol{\Sigma}_\mu=\mathrm{diag}(c_{\mu_1},\ldots,c_{\mu_q})$.

For the Gaussian case, we have σ^2 and assign the prior distribution:

$$\sigma^2 \sim \mathrm{IG}\left(a_\sigma,b_\sigma\right). \tag{4.7}$$

In the case of ordinal outcomes, for identifiability we fix $b_0=0$ and consider the improper uniform prior:

$$\mathrm{pr}(b_1,\cdots,b_{c-1}) \propto \mathbb{1}_{(0<b_1<\cdots<b_{c-1}<\infty)}. \tag{4.8}$$

In the case of the EPA model, priors may also be considered for the concentration α and discount δ parameters.

4.2.5 Full model

In summary, the two proposed BNP SIR models: the EPA SIR and Potts-Gibbs SIR models can be formulated in the following hierarchical order,

$$y_i|\boldsymbol{\mu},\boldsymbol{\beta}^*,\pi_p,\varphi \sim \mathrm{GLM}(\boldsymbol{w}_i^T\boldsymbol{\mu}+\boldsymbol{x}_i^{*T}\boldsymbol{\beta}^*,\varphi), \qquad \forall i=1,\ldots,n,$$

SIR

$$\pi_p \sim \begin{cases}\mathrm{EPA}(\Theta) & \text{for EPA SIR,}\\ \text{Potts-Gibbs}(\Theta) & \text{for Potts-Gibbs SIR,}\end{cases}$$

Random image partition model

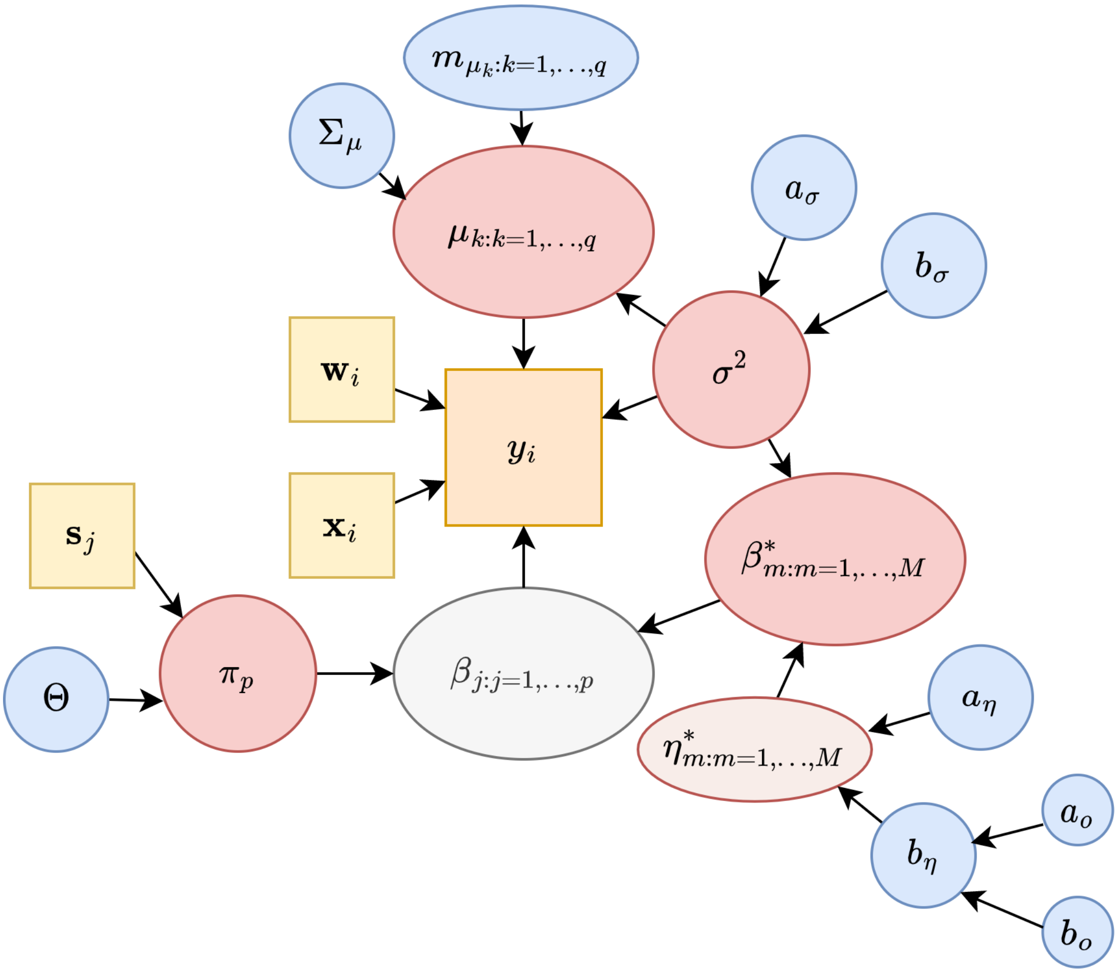


Figure 4.1: The DAG represents the proposed models' structure (Gaussian case). Squares denote data; red nodes denote parameters, blue nodes denote hyperparameters and edges denote dependencies.

$$\boxed{\begin{aligned} \boldsymbol{\beta}^*|\boldsymbol{\eta}^*, \sigma^2 &\sim \mathrm{N}(\mathbf{0}_M, \sigma^2\boldsymbol{\Sigma}_{\beta^*}), \\ \eta^*_m|b_\eta &\sim \mathrm{IG}\left(a_\eta, b_\eta\right), \qquad \forall m = 1, \dots, M, \\ b_\eta &\sim \mathrm{G}(a_o, b_o), \\ &\textbf{\textit{t}-shrinkage prior} \end{aligned}}$$
$$\boxed{\begin{aligned} \boldsymbol{\mu}|\sigma^2 &\sim \mathrm{N}(\boldsymbol{m}_\mu, \sigma^2\boldsymbol{\Sigma}_\mu), \\ \sigma^2 &\sim \mathrm{IG}(a_\sigma, b_\sigma), \qquad \text{for Gaussian case}, \\ &\textbf{Other parameters} \end{aligned}} \tag{4.9}$$

where $x^*_{im} = \sum_{j=1}^{p} x_{ij}\mathbb{1}(j \in C_m)$ represents the total value, e.g. volume in the mth

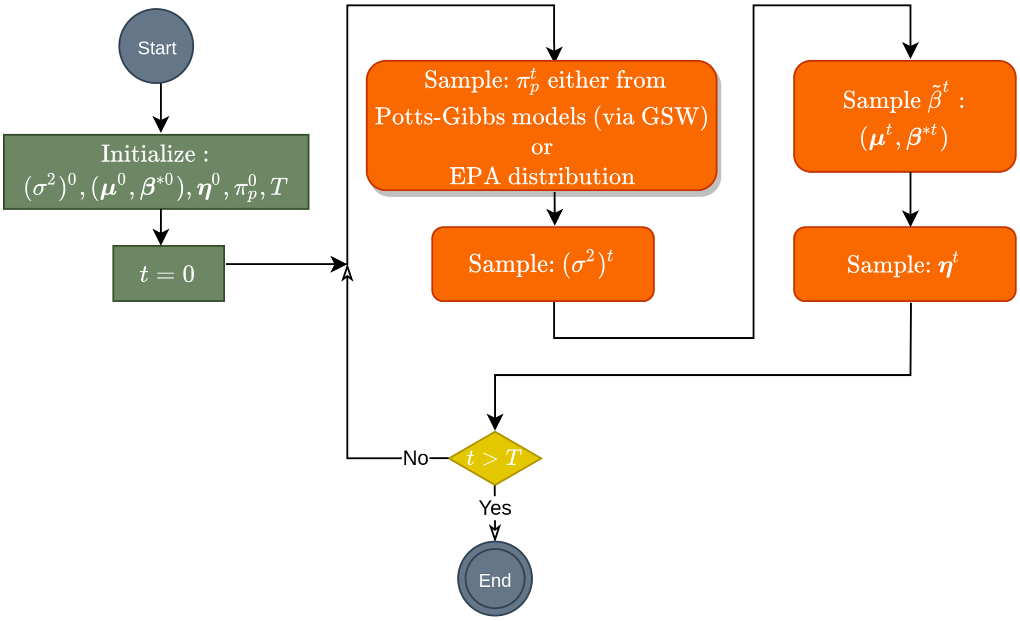

Start
Initialize : $(\sigma^2)^0, (\boldsymbol{\mu}^0, \boldsymbol{\beta}^{*0}), \boldsymbol{\eta}^0, \pi_p^0, T$
$t = 0$
Sample: π_p^t either from Potts-Gibbs models (via GSW) or EPA distribution
Sample: $(\sigma^2)^t$
Sample $\tilde{\beta}^t$: $(\boldsymbol{\mu}^t, \boldsymbol{\beta}^{*t})$
Sample: $\boldsymbol{\eta}^t$
$t > T$
No
Yes
End

Figure 4.2: The flow chart of MCMC for the proposed models.

region of the image. Note that we initially rescale the image predictor by dividing by $\sqrt{p}$; this ensures that the total effects of the image predictors are bounded away from zero (Kang et al., 2018). The Θ represents all the parameters defined by the random image partition models employed. The relationship between observed data, model parameters, and hyperparameters for the proposed models is illustrated in a directed acyclic graph (DAG) in Figure 4.1 for the Gaussian case.

4.3 Posterior inference

For posterior inference, we devise an MCMC algorithm that provides asymptotically exact samples from the posterior of interest. Figure 4.2 shows the flow chart of MCMC for the proposed models. We define $\tilde{\boldsymbol{\beta}} = (\boldsymbol{\mu}, \boldsymbol{\beta}^*)$ and $\tilde{\boldsymbol{X}}$ as the matrix of size $n \times (q+M)$ with rows $(\boldsymbol{w}_i, \boldsymbol{x}_i^*)$. The posterior inference proceeds through the following steps:

Starting MCMC loop

Step 1. Image partition, π_p. Sample the π_p given $\boldsymbol{\eta}^*$ and the data with $\tilde{\boldsymbol{\beta}}, \sigma^2$ marginalised:

$$\mathrm{pr}(\pi_p|\ldots) \propto f(\boldsymbol{y}|\pi_p, \boldsymbol{\eta}^*) \cdot \mathrm{pr}(\pi_p|\Theta),$$

where for the Gaussian case:

$$f\left(\boldsymbol{y}|\pi_p, \boldsymbol{\eta}^*\right) \propto \frac{|\Sigma_{\tilde{\beta}}^{-1}|^{1/2}}{|\Sigma_{\tilde{\beta}}^{-1} + \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}|^{1/2}} \left(b_\sigma + S^2/2\right)^{-(a_\sigma + n/2)},$$

$$S^2 = (\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{m}_{\tilde{\beta}})^T(\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{m}_{\tilde{\beta}}) - (\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{m}_{\tilde{\beta}})^T\tilde{\boldsymbol{X}}(\Sigma_{\tilde{\beta}}^{-1} + \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T(\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{m}_{\tilde{\beta}}),$$

and for the binary and ordinal case:

$$f\left(\boldsymbol{y}|\pi_p, \boldsymbol{\eta}^*\right) \propto \frac{|\Sigma_{\tilde{\beta}}^{-1}|^{1/2}}{|\Sigma_{\tilde{\beta}}^{-1} + \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{y} - \frac{1}{2}\boldsymbol{m}_{\tilde{\beta}}^T\Sigma_{\tilde{\beta}}^{-1}\boldsymbol{m}_{\tilde{\beta}} + \frac{1}{2}\hat{\boldsymbol{m}}_{\tilde{\beta}}^T\hat{\Sigma}_{\tilde{\beta}}^{-1}\hat{\boldsymbol{m}}_{\tilde{\beta}}\right).$$

The $\mathrm{pr}(\pi_p|\Theta)$ represents either the EPA or Potts-Gibbs image partition model. Full details on how the image partition is sampled are provided in Section 4.3.1.

Step 2. **Coefficients and noise variance, $\tilde{\boldsymbol{\beta}}$ and** σ^2. Sample $\tilde{\boldsymbol{\beta}}$ and σ^2 jointly given the partition π_p, $\boldsymbol{\eta}^*$ and the data from the corresponding full conditional distribution:

$$\tilde{\boldsymbol{\beta}}|\boldsymbol{\eta}^*, \sigma^2, \cdots \sim \mathrm{N}\left(\hat{\boldsymbol{m}}_{\tilde{\beta}}, \sigma^2\hat{\Sigma}_{\tilde{\beta}}\right),$$

$$\sigma^2|\cdots \sim \mathrm{IG}\left(\hat{a}_\sigma, \hat{b}_\sigma\right),$$

where $\hat{\Sigma}_{\tilde{\beta}} = \left(\Sigma_{\tilde{\beta}}^{-1} + \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\right)^{-1}$, $\hat{\boldsymbol{m}}_{\tilde{\beta}} = \hat{\Sigma}_{\tilde{\beta}}\left(\Sigma_{\tilde{\beta}}^{-1}\boldsymbol{m}_{\tilde{\beta}} + \tilde{\boldsymbol{X}}^T\boldsymbol{y}\right)$, and $\mathrm{IG}(\hat{a}_\sigma, \hat{b}_\sigma)$ denotes the inverse-gamma distribution with updated shape $\hat{a}_\sigma = a_\sigma + n/2$ and scale $\hat{b}_\sigma = b_\sigma + 1/2\boldsymbol{m}_{\tilde{\beta}}^T\Sigma_{\tilde{\beta}}^{-1}\boldsymbol{m}_{\tilde{\beta}} + 1/2\boldsymbol{y}^T\boldsymbol{y} - 1/2\hat{\boldsymbol{m}}_{\tilde{\beta}}^T\hat{\Sigma}_{\tilde{\beta}}^{-1}\hat{\boldsymbol{m}}_{\tilde{\beta}}$.

The computational complexity is linear with respect to the number of unique clusters, M, i.e. the cost is $\mathcal{O}(M)$.

Step 3. **Local shrinkage parameters**, $\boldsymbol{\eta}^*$. Sample $\boldsymbol{\eta}^*$ given $\boldsymbol{\beta}^*$, σ^2 and b_η. The corresponding full conditional distribution for each η_m^* is an inverse-gamma distribution with updated shape $\hat{a}_\eta = a_\eta + 1/2$ and scale $\hat{b}_\eta = b_\eta + (\beta_m^*)^2/(2\sigma^2)$, as follows,

$$\eta_m^*|\cdots \sim \mathrm{IG}\left(\hat{a}_\eta, \hat{b}_\eta\right),$$

for all $m = 1, \ldots, M$.

The computational complexity is linear with respect to the number of unique clusters, M, i.e. the cost is $\mathcal{O}(M)$.

Step 4. **Global shrinkage parameter**, b_η. Sample b_η from the gamma distribution with updated shape $\hat{a}_o = a_o + Ma_\eta$ and rate $\hat{b}_o = b_o + \sum_{m=1}^M 1/\eta_m^*$, as follows,

$$b_\eta|\cdots \sim \mathrm{G}\left(\hat{a}_o, \hat{b}_o\right).$$

4.3.1 The update of the partition π_p

Gibbs sampling (Geman and Geman, 1984) was originally designed for drawing updates from the Gibbs distribution. However, poor mixing can be seen in single-site Gibbs sampling, which in our case sequentially updates the allocation of a single-pixel given all others due to the posterior correlation between the unit labels. Therefore, when the correlation is high, it takes a long time to converge. The Swendsen–Wang algorithm (Swendsen and Wang, 1987) addresses this problem by constructing efficient split and merge moves that form nested clusters of neighbouring units and then update all of the labels within a nested cluster to the same value. The generalized Swendsen-Wang (GSW) generalises SW to include additional algorithmic tuning parameters (i.e. κ and τ) to improve mixing (Adrian and Zhu, 2005; Barbu and Zhu, 2007). Classical SW algorithm is when $\kappa = 1$ and $\tau = 0$.

The GSW sampler updates simultaneously the cluster allocation of groups of units and hence improves the exploration of the posterior. The algorithm relies on the introduction of auxiliary binary bond variables,

$$r_{jk} = \begin{cases} 1 & \text{if } j \text{ and } k \text{ are bonded} \\ 0 & \text{otherwise,} \end{cases}$$

for sites $1 \leq j < k \leq p$. The bond variables r_{jk} induce nested groups of sites which have the same cluster label. This defines a partition of the units into nested clusters $A_1, \ldots, A_O$, where $O \geq M$ denotes the number of nested clusters. For each $1 \leq j < k \leq p$ such that $j \sim k$, we sample the bond variables as follows,

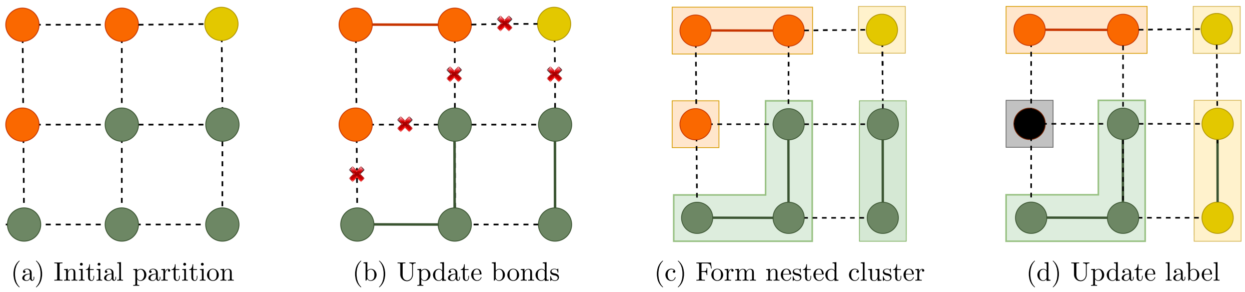
$$r_{jk} \sim \text{Ber}\left(1 - \exp(-\upsilon \zeta_{jk} \mathbb{1}_{z_j = z_k})\right), \tag{4.10}$$

where $\mathbb{1}_{z_j=z_k}$ is an indicator variable equals 1 when both units j and k are assigned the same cluster label and ζ_{jk} are the tuning parameters of the GSW sampler. Note that units j and k may be bounded, i.e. r_{jk} may be non-zero only for those units which are neighbours and are in the same cluster ($z_j = z_k$).

The design of ζ_{jk} is application specific. A good choice will inform the nested clustering step and achieve faster convergence. Importantly, it has to be symmetric, i.e. satisfy $\zeta_{jk} = \zeta_{kj}$ to ensure symmetry of the edge weights. For simplicity, we define ζ_{jk} based on the data,

$$\zeta_{jk}(\tau, \kappa) = \kappa \exp\{-\tau d(\hat{\beta}_j, \hat{\beta}_k)\},$$

where $d(\cdot, \cdot)$ is some distance measure, $\hat{\beta}_j$ is the estimated coefficient from univariate regression on the jth unit, and κ, τ are some positive tuning parameters. Notice that the algorithm reduces to single-site Gibbs when $\kappa = 0$, i.e. all nested clusters are singletons, and recovers classical SW when $\kappa = 1$ and $\tau = 0$.


(a) Initial partition

(b) Update bonds

(c) Form nested cluster

(d) Update label

Figure 4.3: Illustration of the steps of the GSW sampler. (a) An example of an initial partition with three different colours corresponding to three clusters. (b) Cross-marks are used to indicate sites that have no chance to be bonded while colour lines are used to represent sites that are bonded together according to Equation (4.10). (c) A nested cluster is created by sites that are bound together. (d) All of the labels (colours) for a given nested cluster are updated simultaneously.

Overall generalized Swendsen-Wang (GSW) sampler:

Figure 4.3 illustrates the overall GSW sampler. The overall GSW sampler proceeds through the following steps:

1. We create the nested clusters $A_1, \dots, A_O$, where O is the number of nested clusters. For each neighbour pair $j \sim k$ for $1 \le j < k \le p$, we sample the bond variables from Equation (4.10).

2. We update successively the cluster assignment of each nested cluster A_o given the cluster assignments of the other nested clusters. For $o = 1, \dots, O$, each nested cluster, A_o is removed from its current cluster in turn because a new cluster assignment for A_o will be sampled in the current iteration.

   We denote $C_1^{-A_o}, \dots, C_{M^{-A_o}}^{-A_o}$ as the clusters without nested cluster A_o with M^{-A_o} be the number of distinct clusters excluding A_o. The $\pi_p^{-A_o} = \{C_1^{-A_o}, \dots, C_{M^{-A_o}}^{-A_o}\}$ represents the partition obtained by removing all the sites $j \in A_o$ from π_p and $\pi_p^{A_o \to m}$ represents the partition obtained by adding nested cluster A_o to cluster m in partition $\pi_p^{-A_o}$. For each nested cluster A_o, it is assigned to an existing cluster $(m = 1, \dots, M^{-A_o})$ or a new cluster $(m = M^{-A_o} + 1 \dots, M^{-A_o} + h)$ according to the predictive probabilities of the chosen random image partition models and the likelihood:

$$\mathrm{pr}(A_o \in C_m^{-A_o} | \pi_p^{-A_o}, \cdots) \propto f\left(\boldsymbol{y} | \pi_p^{A_o \to m}, \boldsymbol{\eta}^*\right) \mathrm{pr}(A_o \in C_m^{-A_o} | \pi_p^{-A_o}, \Theta). \tag{4.11}$$

   Refer to Table 4.1 for the predictive probabilities of each model (second term on the right-hand side of Equation (4.11)).

	Component	Θ	Existing cluster	New cluster
EPA		α, δ, τ	$\frac{p-1-\delta M^{-j}}{\alpha+p-1} \cdot \frac{\sum_{\psi_k \in C_m} \mathcal{S}(\psi_j,\psi_k)}{\sum_{k=1}^{j-1} \mathcal{S}(\psi_j,\psi_k)}$	$\frac{\alpha+\delta M^{-j}}{\alpha+p-1}$
Potts-DP	Gibbs	α	$\frac{\Gamma(\lvert C_m^{-A_o}\rvert+\lvert A_o\rvert)}{\Gamma(\lvert C_m^{-A_o}\rvert)}$	$\alpha\Gamma(\lvert A_o\rvert)$
	Potts	υ	$\prod_{\{(j,k) \mid j \in A_o, k \in C_m^{-A_o}, r_{jk}=0\}} \exp\{\upsilon(1-\zeta_{jk})\}$	
Potts-PY	Gibbs	α, δ	$\frac{\Gamma(\lvert C_m^{-A_o}\rvert+\lvert A_o\rvert-\delta)}{\Gamma(\lvert C_m^{-A_o}\rvert-\delta)}$	$(\alpha + \delta M^{-A_o})\frac{\Gamma(\lvert A_o\rvert-\delta)}{\Gamma(1-\delta)}$
	Potts	υ	$\prod_{\{(j,k) \mid j \in A_o, k \in C_m^{-A_o}, r_{jk}=0\}} \exp\{\upsilon(1-\zeta_{jk})\}$	
Potts-MFM	Gibbs	λ, γ	$\frac{\Gamma(\lvert C_m^{-A_o}\rvert+\lvert A_o\rvert+\gamma)}{\Gamma(\lvert C_m^{-A_o}\rvert+\gamma)}$	$\frac{V_p(M^{-A_o}+1)}{V_p(M^{-A_o})} \frac{\Gamma(\lvert A_o\rvert+\gamma)}{\Gamma(\gamma)}$
	Potts	υ	$\prod_{\{(j,k) \mid j \in A_o, k \in C_m^{-A_o}, r_{jk}=0\}} \exp\{\upsilon(1-\zeta_{jk})\}$	

Table 4.1: Parameters Θ and terms of the predictive probability for assigning current cluster to either existing cluster or new cluster for EPA, Potts-DP, Potts-PY and Potts-MFM. Note that the predictive probabilities are stated up to a proportionality constant.

Sampling for non-conjugate priors:

Since we are dealing with a model with non-conjugate priors, we update the new cluster assignment by extending Gibbs sampling with the addition of auxiliary parameters, which is widely known as Neal 8 Algorithm (Neal, 2000). We define h temporary auxiliary variables $\left\{\eta^*_{M^{-A_o}+1}, \ldots, \eta^*_{M^{-A_o}+h}\right\}$ that represent possible values for the local shrinkage parameters of new clusters. If the action of removing A_o from its current cluster causes the cluster to become empty (i.e. A_o is its own outer cluster), we set its local shrinkage parameter as the first of these auxiliary parameters. That is, suppose $z_j = m$ for all $j \in A_o$, then we set $\eta^*_{m^{-A_o}+1}$ be equal to η^*_m. The others auxiliary parameter $\left\{\eta^*_{M^{-A_o}+2}, \ldots, \eta^*_{M^{-A_o}+h}\right\}$ will be sampled independently from the prior distribution. The order of other clusters is reordered to be consecutive $1, \ldots, M^{-A_o}$. Otherwise if $z_j = z_k$ for some $j \neq k$ and $k \notin A_o$, h temporary auxiliary variables $(\eta^*_{M^{-A_o}+1}, \ldots, \eta^*_{M^{-A_o}+h})$ are again drawn independently from the prior distribution.

Computational complexity:

We apply the GSW sampler to draw samples from the Potts-Gibbs models. Before updating the cluster assignments, we sample the nested clusters and compute the volume of each nested cluster for all images, with computational cost $\mathcal{O}(np)$. When updating the cluster assignments, the marginal likelihood dominates the computational cost, as it involves inversion and determinants of $(M+q) \times (M+q)$ matrices and updating the sufficient statistics for every nested cluster and every outer cluster allocation, i.e. the cost is $\mathcal{O}([[M+q]^3 + n[M+q]]OM)$.

On the other hand, for the EPA SIR model, we use the Metropolis-Hastings within Gibbs to sample a new partition, i.e. each nested cluster is a singleton. The computational cost for the marginal likelihood part is similar to the Potts-Gibbs SIR, i.e. the cost is $\mathcal{O}([[M+q]^3 + n[M+q]]pM)$. However, the EPA sequential cluster assignment poses an additional computational burden, namely, the computational cost of EPA sequential cluster assignment is $\mathcal{O}(p^2M)$. As the number of pixels p is typically much larger than the sample size n and the number of clusters M, the EPA predictive probability dominates.

4.3.2 Prior and tuning parameter specification

The prior and tuning parameter specification for posterior inference described in Section 4.3 includes:

Step 1. Set the fixed parameter values for $(a_\sigma, b_\sigma), (a_\eta, a_o, b_o)$ and $(\boldsymbol{m}_\mu, \boldsymbol{\Sigma}_\mu)$.

Step 2. Define the model architecture. Set the values for the parameters Θ for the random image partition models chosen. If EPA distribution is chosen, $\Theta = (\alpha, \delta, \tau, \Psi)$ whereas if Potts-Gibbs models are chosen, $\Theta = (\upsilon, \phi)$ depending on which Gibbs-type partition models are selected. Refer to Table 4.1 for more details of parameters for each chosen model.

Step 3. Define the tuning parameters for the GSW sampler, which is explained in Section 4.3.1. The GSW sampler is implemented to draw π_p for Potts-Gibbs SIR. Thus, if Potts-Gibbs SIR is selected, set the values of the tuning parameters for the GSW sampler, which are κ and τ. For the EPA SIR model, no action is required for this step as the Metropolis-Hastings within Gibbs are used to sample the new partition.

4.3.3 Auxiliary variable model for binary and ordinal data

There is extensive research for the analysis of binary and categorical data in the context of Bayesian ordinal probit regression. The well-known data augmentation algorithm by Albert and Chib (1993) involves augmenting with auxiliary Gaussian random variables, $\tilde{\boldsymbol{y}}$. Holmes and Held (2006) have proposed a technique to improve the performance of the auxiliary variable Gibbs sampler in probit regression simulation by updating the regression coefficients, $\tilde{\boldsymbol{\beta}}$, and the auxiliary variable, $\tilde{\boldsymbol{y}}$, jointly to reduce autocorrelation and subsequently improve the mixing. Alternatively, we could avoid data augmentation and compute the marginal likelihood and posterior for the Bayesian probit regression as Durante (2019); specifically, Durante (2019) has proved that under Gaussian priors for the regression coefficients, the posterior distribution of the Bayesian probit regression can be derived in closed-form and belongs to the class of unified skew-normal random variables. We implemented and compared the three approaches, but focus on the simple approach of Albert and Chib (1993), as the other two which have added computational complexity did not substantially improve mixing.

We consider an ordinal probit likelihood to generalise the model for an ordinal response.

Specifically, for categories $c = 0, 1, 2$:

$$\text{pr}(y_i \leq c) = \Phi\left\{\frac{b_c - (\boldsymbol{w}_i^T \boldsymbol{\mu} + \boldsymbol{x}_i^T \boldsymbol{\beta})}{\sigma}\right\}, \tag{4.12}$$

where b_c are the cutoff points for $c = 0, 1, 2$ with $b_0 = 0$ and $b_2 = \infty$. For identifiability, we typically fix $\sigma = 1$ and infer the free cutoff points.

Note that the model can be equivalently formulated through a latent Gaussian response (McCullagh, 1980):

$$\tilde{y}_i = \boldsymbol{w}_i^T \boldsymbol{\mu} + \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim \text{N}(0, \sigma^2),$$

where the latent responses are linked to the ordinal response through:

$$y_i = \begin{cases} 0 & \tilde{y}_i \leq 0 \\ 1 & 0 < \tilde{y}_i \leq b_1 \\ 2 & b_1 < \tilde{y}_i. \end{cases}$$

The ordinal probit model in Equation (4.12) is recovered after marginalisation of the latent response $\tilde{y}_i$. As in the binary case, we extend the MCMC algorithms with an additional Gibbs step to sample the latent $\tilde{y}_i$ from a truncated normal using CDF inversion (Albert and Chib, 1993). To infer the cutoff value b_1, we sample as follows:

$$b_1 | \cdots \sim \text{Unif}\left(\max\left\{\tilde{y}_i : y_i = 1\right\}, \min\left\{\tilde{y}_i : y_i = 2\right\}\right),$$

where Unif(a, b) denotes a uniform distribution with a the minimum and b the maximum values.

4.3.4 Consensus clustering

Recently Coleman et al. (2021) have proposed an ensemble approach for the Bayesian mixture models. The proposed approach addresses the problem of multimodality in the likelihood surface. Mixing problems are highly likely to happen in multimodal, high-dimension posterior distributions. The simulation studies show promising results that the proposed approach is capable of producing more stable clusterings compared to individual long chains that are prone to becoming trapped in individual modes in high dimensional problems. On top of that, the runtime is reduced. Therefore, the proposed ensemble approach can be a viable alternative when we have mixing problems or poor scalability which make the algorithm infeasible to run large numbers of iterations.

They have provided a heuristic way to determine the ensemble width, W and ensemble depth, T. First, we define the terms needed:

- Co-clustering matrix, $\boldsymbol{C}^1$ is a binary matrix with its element defined as follows:

$$c^1_{jk} = \mathbb{1}_{z_j = z_k}.$$

- Consensus matrix, $\boldsymbol{C}^2$ with its element is defined as follows:

$$c^2_{jk} = \frac{1}{W} \sum_{w=1}^{W} \mathbb{1}_{z_{wj} = z_{wk}},$$

  where $\boldsymbol{Z}$ is the cluster membership matrix with each row being the cluster membership vector $\boldsymbol{z}_w = (z_{w1}, \cdots, z_{wp})$ at chain w.

At each chain, they compute a co-clustering matrix, $\boldsymbol{C}^1$, using the final sample. After running W chains, they combine the W co-clustering matrices, $\boldsymbol{C}^1$, to form the consensus matrix, $\boldsymbol{C}^2$. And finally, they compute the mean absolute difference of the consensus matrix, $\boldsymbol{C}^2$, as an early stopping criterion. Motivated by the use of the scree plot in principal component analysis (PCA), they recommend plotting the mean absolute difference between the sequential consensus matrices, $\boldsymbol{C}^2$ for (T, W) and $(T-1, W)$ as a function of ensemble depth, T, for different values of ensemble width, W, to determine the values of the ensemble parameters.

4.4 Posterior summary and prediction

Based on the proposed models, we can produce various summaries to describe the posterior quantities of interest. The expectation for quantities of interest can be computed based on Monte Carlo approximations using the MCMC samples. We define T as the number of iterations and any variable with a "$\widehat{\ }$" accent as an estimate for the variable. For example, $\widehat{\beta}_j$ denotes the posterior mean estimate for β_j.

First, we summarise the posterior of the clustering structure by considering the posterior similarity matrix, which represents the posterior probabilities that two units belong to the same cluster, i.e.

$$\text{pr}(z_j = z_k | \text{data}) \simeq \frac{1}{T} \sum_{j=1}^{p} \mathbb{1}_{z^t_j = z^t_k}.$$

In the case of consensus clustering, the consensus matrix is similar in nature to the posterior similarity matrix and can be reported to summarize uncertainty in the clustering structure. Based on this posterior similarity matrix (or consensus matrix), we can also obtain a point estimate of the clustering by minimising the posterior expected variation of information (VI) (Wade and Ghahramani, 2018). We can also compute an estimate of the posterior expectation of β_j $(j = 1, \cdots, p)$:

$$\mathbb{E}[\beta_j|\text{data}] \simeq \widehat{\beta}_j = \frac{1}{T}\sum_{t=1}^{T} \beta_j^t.$$

We are also interested in understanding which coefficients, β_j $(j = 1, \cdots, p)$ are more likely to be included in the model. To do so, we plot the posterior inclusion map which has a value of 0 or 1. First, we define posterior probabilities of β_j for being less than 0 and greater than 0:

$$\begin{aligned}
\text{pr}(\beta_j < 0|\text{data}) &\simeq \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\beta_j^t < 0}, \\
\text{pr}(\beta_j > 0|\text{data}) &\simeq \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\beta_j^t > 0}.
\end{aligned}$$

Then we get the posterior inclusion probability (PIP) as follows:

$$\text{PIP} = \max\left\{\text{pr}\left(\beta_j < 0|\text{data}\right), \text{pr}\left(\beta_j > 0|\text{data}\right)\right\}.$$

To identify which parts of the image are relevant for predicting the response, we plot the binary posterior inclusion map by thresholding PIP at a specified significance level.

For the outcome measure, y_i $(i = 1, \cdots n)$, we separate into two cases: regression and classification. For the regression, we compute the posterior expectation:

$$\mathbb{E}[y_i|\text{data}] \simeq \widehat{y_i} = \frac{1}{T}\sum_{t=1}^{T} \tilde{\boldsymbol{x}}_i^T \tilde{\boldsymbol{\beta}}^t.$$

Note one can also compute other quantities such as the posterior density or credible intervals. For the classification problem, we compute the posterior probability of class membership for being classified as 0 (negative) or 1 (positive):

$$\begin{aligned}
\text{pr}(y_i = 1|\text{data}) \simeq \widehat{p}_i(1) &= \frac{1}{T}\sum_{t=1}^{T} \text{pr}(y_i = 1|\tilde{\boldsymbol{\beta}}^t, \tilde{\boldsymbol{x}}_i) \\
&= \frac{1}{T}\sum_{t=1}^{T} \Phi(\tilde{\boldsymbol{\beta}}^t \tilde{\boldsymbol{x}}_i), \\
\text{pr}(y_i = 0|\text{data}) \simeq \widehat{p}_i(0) &= 1 - \widehat{p}_i(1).
\end{aligned}$$

In addition, we can predict the respnse for a new individual, and perform posterior predictive checking to assess the goodness of fit for some predictive summaries of interest. Posterior predictive checking is employed as a tool for model checking to assess whether the assumed model is reasonable for the data. Let $\boldsymbol{y}^{rep}$ denote a replicated dataset sampled from the posterior predictive distribution given by:

$$\text{pr}(\boldsymbol{y}^{rep}|\text{data}) = \int \text{pr}(\boldsymbol{y}^{rep}|\boldsymbol{\mu}, \boldsymbol{\beta}^*, \pi_p, \varphi)\text{pr}(\boldsymbol{\mu}, \boldsymbol{\beta}^*, \pi_p, \varphi|\boldsymbol{y})d(\boldsymbol{\mu}, \boldsymbol{\beta}^*, \pi_p, \varphi),$$

which can be approximated by:

$$\text{pr}(\boldsymbol{y}^{rep}|\text{data}) \simeq \frac{1}{T}\sum_{t=1}^{T} \text{pr}\left(\boldsymbol{y}^{rep}|\boldsymbol{\mu}^t, \boldsymbol{\beta}^{*t}, \pi_p^t, \varphi^t\right).$$

4.5 Conclusions

In this chapter, we have developed novel SIR models that combine random image partition models with sparsity promoting priors to automatically extract regions of interest from the image predictors. We have described the building blocks of our proposed models and the proposed posterior inference schemes, along with a discussion on computational complexity. In particular, the EPA SIR model can be expensive compared to the Potts-Gibbs SIR models because of (1) the sequential calculation of the EPA predictive distribution inside the EPA SIR model resulting in a cost which is quadratic in the number of pixels; (2) the intrinsic nested clusters formed via the GSW algorithm help to reduce the computational time of sampling the new partition for the Potts-Gibbs SIR models. However, due to the high-dimensional nature of SIR, the choice of random image partition model plays an important role, as discussed in Chapter 3, the EPA and choices within the Potts-Gibbs class induce different properties and behaviour in the image partition apriori. In the next chapter, we provide a thorough comparison based on simulations and an application to predict dementia.

Chapter 5

Experiments

In this chapter, we demonstrate the potential of the proposed models, EPA SIR and Potts-Gibbs SIR models, on both simulated datasets and real datasets. We have designed four different scenarios for simulated datasets to test the proposed models. We also apply our proposed models to real data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Our proposed models are compared with competing SIR models namely the Ising model (Huang et al., 2013), Ising-GMRF model (Goldsmith et al., 2014) and Ising-DP model (Li et al., 2015). The results highlight the utility of the proposed models in extracting interpretable features, while also providing improved performance.

5.1 Introduction

This chapter explains the approaches adopted in the simulation of the datasets and outlines the experiments conducted to show the potential of our proposed models with a detailed comparison of the proposed models with the competitors. Specifically, we focus on the competing SIR models described in Section 2.1, namely the Ising model (Huang et al., 2013), Ising-GMRF model (Goldsmith et al., 2014) and Ising-DP model (Li et al., 2015). We do not include the STGP model by Kang et al. (2018) because from our preliminary experiments on the real dataset, we observe that the STGP model produces a coefficient map that is quite different from the other models which we cannot interpret well.

The remainder of this chapter is organised as follows. In Section 5.2 and 5.3, we give a description of simulated datasets and real datasets. In Section 5.4, we explain the use of the Watanabe-Akaike information criterion (WAIC) proposed by Watanabe and Opper (2010) for the selection of key hyperparameters. Also, we have given a clear indication of the hyperparameter specification for the proposed models as well as the competing models. On top of that, we include the guidance for choosing the tuning parameters of the GSW algorithm. In Section 5.5, we detail the evaluation metrics we apply to assess the quality of the clustering, the estimation of the coefficient image, $\boldsymbol{\beta}$, and the prediction for the outcome measures, $\boldsymbol{y}$, to compare the performance of the proposed models with all competing models. Furthermore, we also provide a heuristic way to specify other hyperparameters for the proposed models and competing models. In Section 5.6 and 5.7, we illustrate the proposed models with the analysis of the designed simulated data and real data from ADNI. Finally, we conclude in Section 5.8.

5.2 Data simulation

We have designed four different scenarios to illustrate the proposed models. In all cases, we consider a two-dimensional image.

For Scenarios 1 and 2, the $n = 300$ images are simulated on a two-dimensional grid size of 10×10, with spatial locations $\boldsymbol{s}_j = (s_{j1}, s_{j2}) \in \mathbb{R}^2$ for $1 \leq s_{j1}, s_{j2} \leq 10$. For simplicity's sake, we include an intercept but do not consider other covariates, $\boldsymbol{w}_i$. We concentrate on two simulation scenarios with $M = 2$ and $M = 5$. The images are drawn from a multivariate Gaussian distribution with a mean vector whose elements have a uniform distribution on the interval $(3, 4)$ and a covariance matrix constructed from a squared exponential covariance function (Rasmussen and Williams, 2005):

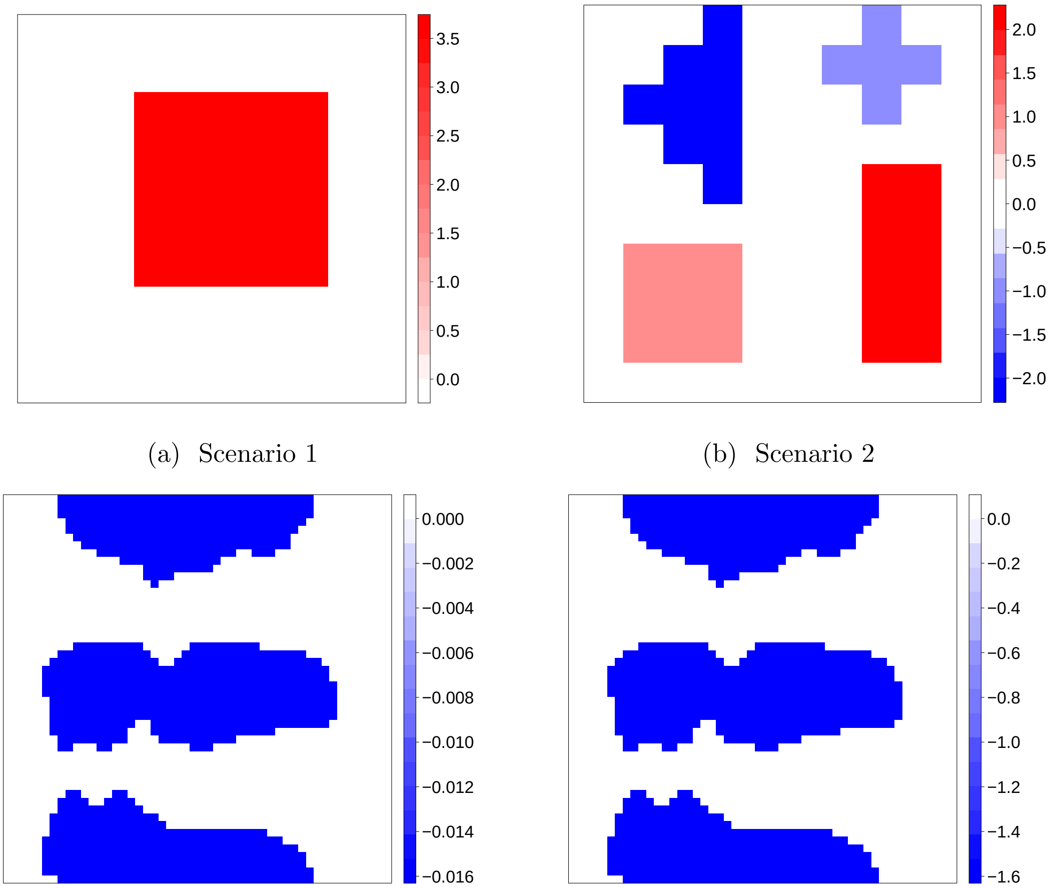
$$\begin{aligned}
\boldsymbol{x}_i &= (x_{i1}, \dots, x_{ip}) \overset{i.i.d}{\sim} \mathrm{MVN}_p(\boldsymbol{u}, \boldsymbol{\Sigma}), \\
u_j &\overset{i.i.d}{\sim} \mathrm{Unif}(3,4), \qquad \text{for } j = 1, \cdots, p, \\
\boldsymbol{\Sigma}_{jk} &= \exp\left\{ -\frac{\sum_{i=1}^{2}(s_{ji} - s_{ki})^2}{10} \right\}, \qquad \text{for } j, k = 1, \cdots, p,
\end{aligned}$$

where $\mathrm{MVN}_p(\boldsymbol{u}, \boldsymbol{\Sigma})$ stands for the p-dimensional multivariate normal distribution with $\boldsymbol{u} = (u_1, \cdots, u_p)$ as the mean vector and $\boldsymbol{\Sigma}$ as the covariance matrix.

To assess the real-world performance of our proposed models, we prepare Scenarios 3 and 4, in which the true data generating distribution is chosen to closely resemble the real image predictors which we describe in Section 5.3. The images are simulated on a grid size of 50×50 with two different clusters. Here the covariance matrix is constructed using the Matérn covariance function with smoothness parameter equal to 5/2. The images are simulated as follows:

$$\begin{aligned}
\boldsymbol{x}_i &\sim \mathrm{MVN}_p(\boldsymbol{v}, \boldsymbol{\Sigma}), \\
\boldsymbol{\Sigma}_{jk} &= \sigma_j \sigma_k \left(1 + \frac{\sqrt{5} d_{jk}}{\rho} + \frac{5 d_{jk}^2}{3\rho^2} \right) \exp\left(-\frac{\sqrt{5} d_{jk}}{\rho} \right),
\end{aligned}$$

where σ_j is estimated standard deviation from the real dataset for the jth pixel, d_{jk} is defined as $\sqrt{\sum_{i=1}^{2}(s_{ji} - s_{ki})^2}$ and ρ is the positive parameter of the covariance matrix (set to $\sqrt{p}/10$). The mean vector $\boldsymbol{v}$ is the empirical mean of the images based on the data described in Section 5.3.



(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

(d) Scenario 4

Figure 5.1: The true coefficient maps of the simulated datasets for each scenario.

In the following, we specify the true coefficient image and describe how the outcome measures, $\boldsymbol{y}$, are generated for all four scenarios. A visualisation of the true coefficient image in all cases is produced in Figure 5.1.

Scenario 1: The true coefficient map contains one cluster of similar values, which are fixed at 3.5. For each spatial location $\boldsymbol{s}_j$, we cluster as below:

$$z_j = \begin{cases} 1 & \text{if } 4 \leq s_{j1}, s_{j2} \leq 8 \\ 2 & \text{otherwise.} \end{cases}$$

The true values of β_m^* for each cluster are fixed to:

$$\beta_m^* = \begin{cases} 0 & \text{if } 4 \leq s_{j1}, s_{j2} \leq 8 \\ 3.5 & \text{otherwise,} \end{cases}$$

and the outcome measures, $\boldsymbol{y}$, are sampled as follows:

$$y_i = \sum_{m=1}^{M} v_{im}\beta_m^* + \epsilon_i, \qquad \text{with } \epsilon_i \sim \mathrm{N}(0, \sigma^2),$$

where $\boldsymbol{v}_i = \left(\sum_{j \in C_1} x_{ij}, \ldots, \sum_{j \in C_M} x_{ij}\right)$ with M equal to the number of clusters (i.e. $M = 2$ for Scenario 1 and $M = 5$ for Scenario 2).

Scenario 2: The true coefficient map contains four regions: a square, triangle, rectangle and cross. For each spatial location $\boldsymbol{s}_j$, we cluster as below:

$$z_j = \begin{cases} 1 & \text{if } \boldsymbol{s}_j \text{ belongs to the square} \\ 2 & \text{if } \boldsymbol{s}_j \text{ belongs to the triangle} \\ 3 & \text{if } \boldsymbol{s}_j \text{ belongs to the rectangle} \\ 4 & \text{if } \boldsymbol{s}_j \text{ belongs to the cross} \\ 5 & \text{otherwise.} \end{cases}$$

The true values of β_m^* for each cluster are fixed to:

$$\beta_m^* = \begin{cases} 1.0 & \text{if } m = 1 \\ -2.0 & \text{if } m = 2 \\ 2.0 & \text{if } m = 3 \\ -1.0 & \text{if } m = 4 \\ 0 & \text{otherwise.} \end{cases}$$

Thus, we have two regions with strong signals (clusters 2 and 3) and the two with weaker signals. The outcome measures, $\boldsymbol{y}$, are sampled as follows:

$$y_i = \sum_{m=1}^{M} v_{im}\beta_m^* + \epsilon_i, \qquad \text{with } \epsilon_i \sim \mathrm{N}(0, \sigma^2).$$

Scenario 3: The true coefficient map contains two regions, which is obtained via the hierarchical clustering based on spatial proximity, i.e. the image coordinates, as well as non-spatial parameter, i.e. the difference of the empirical mean of the images within the two classes (cognitively normal (CN) and Alzheimer's disease (AD)). Then, the true values of β^*_m for each cluster are fixed to:

$$
\begin{aligned}
c_1 &= \sum_{j=1}^{p} (u_{1j} - u_{0j})\, \mathbb{1}_{z_j==1}, \\
c_2 &= \sum_{j=1}^{p} (u_{1j} - u_{0j})\, \mathbb{1}_{z_j==2}, \\
\beta^*_m &= \begin{cases} \frac{c_1}{\frac{1}{n}\sum_{i=1}^{n}(v_{i1}-c_1)^2} & \text{if } m = 1 \\ \frac{c_2}{\frac{1}{n}\sum_{i=1}^{n}(v_{i2}-c_2)^2} & \text{if } m = 2, \end{cases}
\end{aligned}
$$

where c_1 and c_2 represent the total difference between the empirical mean of the images for CN and AD, for pixels which belong to the first cluster and second cluster, respectively. The mean vectors $\boldsymbol{u}_0 = (u_{01}, \cdots, u_{0p})$ and $\boldsymbol{u}_1 = (u_{11}, \cdots, u_{1p})$ are defined as the empirical mean of the images within the two classes, CN and AD, respectively. To obtain sparsity in the coefficient image, the difference, $\boldsymbol{u_1} - \boldsymbol{u_0}$, is thresholded and set to 0 if the absolute value is smaller than a specified threshold.

The continuous outcome measures, $\boldsymbol{y}$, are sampled as follows:

$$
\begin{aligned}
w_1 &= -\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{v}_i^T \boldsymbol{\beta}^*, \\
y_i &= w_1 + \sum_{m=1}^{M} v_{im}\beta^*_m + \epsilon_i, \qquad \text{with } \epsilon_i \sim \mathrm{N}(0, \sigma^2),
\end{aligned}
$$

where w_1 denotes the intercept.

Scenario 4: The true coefficient image is derived as Scenario 3, but the values are multiplied by 100. The binary outcome measures, $\boldsymbol{y}$, are sampled as follows:

$$
\begin{aligned}
w_1 &= -\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{v}_i^T \boldsymbol{\beta}^*, \\
\tilde{y}_i &= w_1 + \sum_{m=1}^{M} v_{im}\beta^*_m + \epsilon_i, \qquad \text{with } \epsilon_i \sim \mathrm{N}(0, 1), \\
y_i &= \begin{cases} 0 & \text{if } \mathbb{1}_{\tilde{y}_i<0} \\ 1 & \text{if } \mathbb{1}_{\tilde{y}_i\geq 0}. \end{cases}
\end{aligned}
$$

5.3 Real data application

Various Alzheimer's disease (AD) initiatives are taking place across the globe to collect and share a variety of AD-related data including neuroimaging, clinical, and biological

		CN	MCI	AD
Total	Total	224	392	182
Gender	Male	116	250	96
	Female	108	142	86
Age (years)	Range	62-90	55-89	55-91
		76.1 ±4.9	74.8 ±7.3	75.3 ±7.5
APOE e4 carriers	0 (e2/e3)	164	180	60
	1 (e3/e4)	55	165	86
	2 (e4/e4)	5	47	36

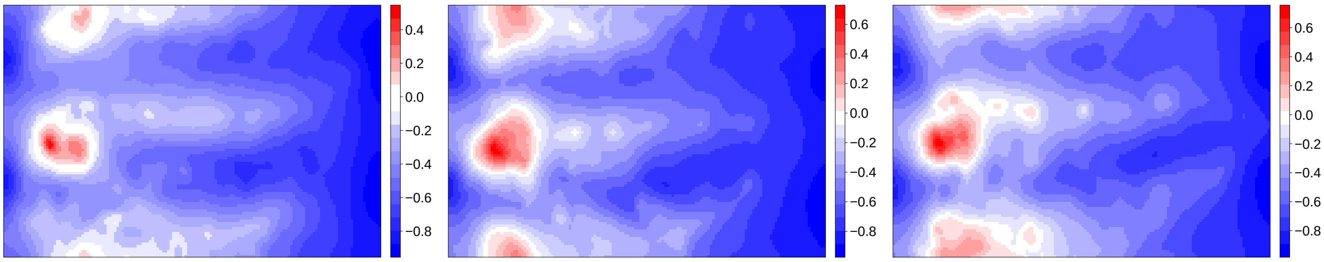
Table 5.1: Subject demographics in cognitively normal (CN), mild cognitive impairment (MCI) and Alzheimer's disease (AD).

information. To illustrate the usefulness of the proposed model, we also use real data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database: `http://adni.loni.usc.edu/about/`[1].

Our study focuses on diagnosis based on structural magnetic resonance images (sMRI). The MRI scans in ADNI have been acquired from 1.5T and 3.0T scanners Magnetom (Siemens, Erlangen). In this study, we focus on ADNI baseline data, that is the data available at the subject's first baseline visit. Data is available for a total of 798 subjects. Out of the 798 subjects, 224 (28.0 %) are cognitively normal (CN), 392 (49.1 %) have mild cognitive impairment (MCI) and 182 (22.8 %) have Alzheimer's disease (AD). Table 5.1 gives information about the demographics of the subjects.

We focus our imaging analysis on the hippocampus. The hippocampus plays a vital role in memory formation (Squire and Zola-Morgan, 1991). There is increasing evidence that it is among the primary areas seen to be more prominently affected by AD (Braak and Braak, 1991; Hyman et al., 1984). The radial distance is a measurement of hippocampus size which primarily explains morphometric changes along the surface normal direction.

[1] The ADNI was launched in 2003 by the National Institute of Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organisations, as a $ 60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitoring their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and the University of California-San Francisco. ADNI is the result of the efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the US and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research, approximately 200 cognitively normal older individuals to be followed for three years, 400 people with MCI to be followed for three years and 200 people with early AD to be followed for two years. For up-to-date information, see `www.adni-info.org`.



(a) CN subject

(b) MCI subject

(c) AD subject

Figure 5.2: The surface statistics (radial distance) of the left hippocampus.

AD patients tend to have smaller hippocampus volumes and a reduced radial distance compared to healthy older adults (Apostolova et al., 2006). We have obtained the processed data from Dr Yalin Wang (Arizona State University). Refer to Shi et al. (2013, 2014b) for more details of the image preprocessing. Thus, we have the hippocampus surface statistics of each subject with the dimension 100×150.

On top of that, we also include additional covariates to help with the analysis: age, gender and presence of the apolipoprotein E (APOE) e4 allele. For the presence of the APOE e4 allele, it can be either 0 indicates there is no copy of the APOE e4 allele, 1 indicates the person is carrying one copy of the APOE e4 allele, and 2 indicates the person is carrying two copies of APOE e4 allele. We exclude the behaviour score from the analysis because the score is used as criteria to define each subject group (CN, MCI and AD). Note that we normalise all the variables between -1 and 1 to reduce the correlation between the intercept and the variables. Refer to Table 5.1 for the info on each covariate.

For each experiment, we run the MCMC algorithm described in Section 4.3 for $T = 10,000$ (for small p) or $T = 100,000$ (for large p) iterations with the first 40% draws removed a "burn-in". Or else, if the computation is too expensive, we choose to combine many short chains ($W = 100-200$ with $T = 5,000$) instead of one long chain, motivated by Coleman et al. (2021) as described in Section 4.3.4. For ease of notation, we combine the different chains and denote the total number of iterations as T.

5.4 Hyperparameters selection

The hyperparameters Θ of the random partition model play an important role in the spatial clustering of the image (see Chapter 3). As such we propose to an empirical Bayes approach to select Θ based on Watanabe-Akaike information criterion (WAIC)

Model	Hyperparameters
EPA	α, δ, τ
Potts-DP	υ, α
Potts-PY	υ, α, δ
Potts-MFM	$\upsilon, \lambda, \gamma$
Ising	$a, b, \sigma_\beta^2, \sigma_\epsilon^2$
Ising-GMRF	$a, b, \sigma_\beta^2, \sigma_\epsilon^2$
Ising-DP	a, b, H, α, v^2

Table 5.2: Hyperparameters for each model.

which is proposed by Watanabe and Opper (2010).

We define p_{WAIC} as the sum of the posterior variance of the log predictive density for each data point y_i (LPPD$_i$). Each LPPD$_i$ can be estimated using the log-likelihood at the MCMC samples, and thereby the LPPD of the fitted model can be computed as:

$$\text{LPPD} = \sum_{i=1}^{n} \log \left\{ \frac{1}{T} \sum_{t=1}^{T} \text{pr}\left(y_i | \tilde{\boldsymbol{x}}_i, \boldsymbol{\mu}, \boldsymbol{\beta}^*, \pi_p, \varphi\right) \right\}.$$

And the p_{WAIC} can be evaluated as follows:

$$\text{V}_i = \frac{1}{T-1} \sum_{t=1}^{T} \left\{ \log \text{pr}\left(y_i | \tilde{\boldsymbol{x}}_i, \boldsymbol{\mu}, \boldsymbol{\beta}^*, \pi_p, \varphi\right) - \frac{1}{T} \sum_{t=1}^{T} \log \text{pr}\left(y_i | \tilde{\boldsymbol{x}}_i, \boldsymbol{\mu}, \boldsymbol{\beta}^*, \pi_p, \varphi\right) \right\}^2,$$

$$p_{\text{WAIC}} = \sum_{i=1}^{n} \text{V}_i,$$

where $1/T \sum_{t=1}^{T} \log \text{pr}(y_i | \tilde{\boldsymbol{x}}_i, \boldsymbol{\mu}, \boldsymbol{\beta}^*, \pi_p, \varphi)$ is the posterior expected value of the log predictive density for each data point y_i. The p_{WAIC} can be viewed as the estimated effective number of parameters in a model and acts as a measure of the complexity of the model. Then we can estimate the expected log pointwise predictive density (ELPPD) using both the computed LPPD and p_{WAIC}:

$$\text{ELPPD} = \text{LPPD} - p_{\text{WAIC}}.$$

And finally, we get the information criterion WAIC, which is equal to $-2 \times$ ELPPD. The WAIC is fully Bayesian as the WAIC uses the posterior predictive distribution. The smaller the WAIC, the better the model fits the data. Watanabe and Opper (2010) also proved that the WAIC is asymptotically equivalent to leave-one-out cross-validation (LOOCV). According to the survey by Gelman et al. (2014) and Vehtari et al. (2017), the WAIC which is fast and convenient to compute can be a good alternative to LOOCV.

For the reasons explained above, we use WAIC as a criterion to choose which combination of the hyperparameters is best for each model. Table 5.2 displays the hyperparameters for each model. A grid search is executed to locate the best combination of the hyperparameters. The grid search is carried out based on a small number of MCMC iterations, with chains of length 1,000 where the first 40% are discarded as burn-in. After identifying the optimal hyperparameters, each algorithm is run with the longer chains, as stated in Section 5.5.

5.4.1 Grid search range and other recommendations

To effectively carry out the grid search, careful specification of the range of hyperparamter values is important. In the following, we provide guidance for choosing the range of possible values.

For the random image partition models, to specify a range for the hyperparameters, we consider a prior guess of the number of clusters, $\widehat{M}$. For the EPA distribution, we then set α and δ according to $\widehat{M}$ based on Equation (3.11) if δ equals to zero, otherwise Equation (3.15), and for the τ, from the experiments we have conducted and prior simulations in Section 3.5, we found that $\tau \in (3, 5)$ are generally good values.

For the Potts-Gibbs models, we select the range of the hyperparameters following the procedure stated below:

1. **Coupling**, υ. Motivated by Equations (3.10) and (3.14), we see that the spatial part, which is the Potts component will be at most $\exp(4\upsilon)$. On the other hand, the Gibbs component, which depends on cluster size, will be at most p. As we mention in Chapter 3, there is a trade-off between the Gibbs and the Potts components. In particular, we may need to choose a fairly large value of υ in order for the spatial part to have a larger influence than the cluster size, i.e. $\exp(4\upsilon) \geq p$.

   We consider the range:
   $$\upsilon = c_\upsilon \log(p),$$
   for $c_\upsilon \in [1/8, 1/6]$, where c_υ is a constant helping to control the trade-off between the Gibbs and the Potts parts. We want to choose υ to have a high proportion of neighbours connected but also avoid the phase transition and a large value of υ that will induce one large cluster.

2. **Gibbs parameters.** Based on the $D = 4$, υ and $\widehat{M}$,

   - for the Potts-DP model, following Proposition 3.3, we set the range of the concentration α based on:

$$\begin{aligned}\mathbb{E}[M] &\gtrapprox \frac{\alpha}{\exp(Dv)}\log(p),\\ \widehat{M} &\gtrapprox \frac{\alpha}{\exp(4v)}\log(p),\\ \alpha &\lessapprox \frac{\widehat{M}\exp(4v)}{\log(p)}.\end{aligned}$$

- for the Potts-PY model, following Proposition 3.5, we set the range of the discount δ based on:

$$\begin{aligned}\mathbb{E}[M] &\gtrapprox cp^{\delta\exp(-Dv)},\\ \widehat{M} &\gtrapprox cp^{\delta\exp(-4v)},\\ \delta &\lessapprox \frac{\log\left(\widehat{M}\right)\exp(4v)}{\log(p)}.\end{aligned}$$

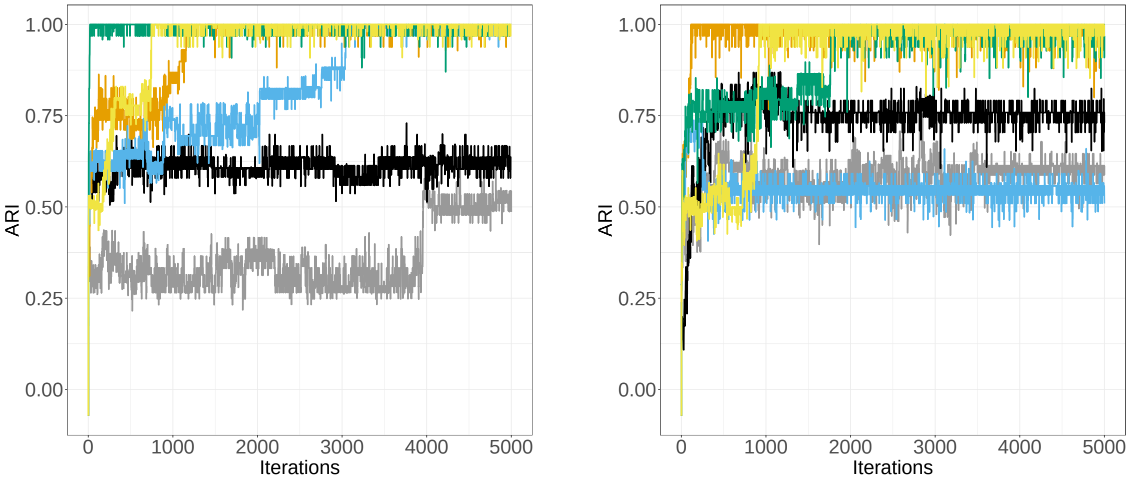
- for the Potts-MFM model, following Proposition 3.7, we set the range of λ based on:

$$\begin{aligned}\mathbb{E}[M] &\gtrapprox \frac{\lambda}{\exp(Dv)},\\ \widehat{M} &\gtrapprox \frac{\lambda}{\exp(4v)},\\ \lambda &\lessapprox \widehat{M}\exp(4v).\end{aligned}$$

For the student-t prior, we set the degree of freedom to be 3, that is $a_\eta = 1.5$, and we set a prior for the hyperparameter b_η, that is $b_\eta \sim \mathrm{G}(a_o, b_o)$. We fix the $a_\sigma = 2$ and choose b_σ according to the estimated variance from the dataset.

Competing models:

For all competing models, we also employ WAIC to select the hyperparamters among a grid of values. All three competing models employ an Ising prior. We set the range of the hyperparameters for the Ising prior as suggested by Goldsmith et al. (2014): $a \in (-4, 0)$ and $b \in (0, 2)$. The σ_β^2 from the Ising model and Ising-GMRF model and the v^2 from the Ising-DP model determine the scale of the coefficients. If the data is very noisy, it is recommended to use a low value of σ_β^2 and v^2. And σ_ϵ^2 is generally within the range $(0.01, 0.1, 1.0)$ after first standardising the outcomes $\boldsymbol{y}$ (Gaussian case). Finally, H and α from the Ising-DP model can be set roughly accordingly to a prior guess of the number of clusters in the data.



(a) $\upsilon = 0.6, \tau = 3$

(b) $\upsilon = 0.3, \tau = 5$

Figure 5.3: Trace plots of ARI for Gibbs sampling (grey), classical SW (black), GSW with κ varies from 3 (green), 5 (yellow), 10 (orange) and 20 (blue), with (a) $\upsilon = 0.6, \tau = 3$ and (b) $\upsilon = 0.3, \tau = 5$.

5.4.2 Tuning parameters of the GSW algorithm

In the following, we provide guidance for choosing the range of possible values for the tuning parameters of the GSW algorithm, κ and τ. Figure 5.3 presents the trace plots of the ARI for Gibbs sampling (grey), classical SW (black), GSW with κ varies from 3 (green), 5 (yellow), 10 (orange) and 20 (blue) for τ equal to 3 (Figure 5.3(a)) and τ equal to 5 (Figure 5.3(b)). We observe that the ARI converges faster to one when employing GSW comparing to Gibbs sampling and classical SW. Thus, GSW allows for efficient split-merge moves that help to improve mixing in this setting.

1. **Tuning parameter**, κ. Recall from Equation (4.10), the auxiliary bond variables, r_{jk}, are sampled accordingly to a Bernoulli distribution with probability equal to $1 - \exp(-\upsilon \zeta_{jk} \mathbb{1}_{z_j = z_k})$. Thus, the expected value for r_{jk} is:

$$\begin{aligned} \mathbb{E}[S] &= \text{pr}(r_{jk}) \\ &= 1 - \exp(-\upsilon \zeta_{jk} \mathbb{1}_{z_j = z_k}). \end{aligned} \tag{5.1}$$

Rearranging Equation (5.1), we obtain:

$$\begin{aligned}\mathbb{E}[S] &= 1 - \exp(-\upsilon\zeta_{jk}\mathbb{1}_{z_j=z_k}),\\ \zeta_{jk} &= -\frac{\log(1-\mathbb{E}[S])}{\upsilon} \times \mathbb{1}_{z_j=z_k},\\ \kappa\exp\{-\tau d(\widehat{\beta}_j,\widehat{\beta}_k)\} &= -\frac{\log(1-\mathbb{E}[S])}{\upsilon} \times \mathbb{1}_{z_j=z_k},\\ \kappa &\le -\frac{\log(1-\mathbb{E}[S])}{\upsilon} \times \mathbb{1}_{z_j=z_k}.\end{aligned} \tag{5.2}$$

From Equation (5.2), we see that κ should be chosen depending υ. It is probably reasonable to choose $\kappa = c_\kappa/\upsilon$, where c_κ is a constant which can be computed as $c_\kappa = -\log(1-\mathbb{E}[S])$.

2. **Tuning parameter**, τ. The tuning parameter τ is chosen based on the scale of the distance. We use the square root of the mean of the squared distances, $1/\tau = \sqrt{\sum d(\widehat{\beta}_j,\widehat{\beta}_k)^2/\mathbb{E}[S]}$.

5.5 Evaluation metrics

We compute a variety of quantities of interest and evaluation metrics to assess clustering accuracy and the robustness of predictions.

For clustering, we consider the adjusted Rand index (ARI) and variation of information (VI) between the real and estimated clustering. The ARI has a range of [0,1], with zero indicating that the two clusterings do not agree on any pair of observations, whereas one indicates that the two clusterings are the same. The VI has a value of zero when the two clusterings are equivalent and a maximum value of log(p) (Meilă, 2007).

To assess the estimation of the coefficient image, $\boldsymbol{\beta}$, we compute the mean squared error (MSE)

$$\text{MSE} = \frac{1}{p}\sum_{j=1}^{p}\left(\beta_j^0 - \widehat{\beta}_j\right)^2,$$

where β_j^0 denotes the true coefficient image for $j = 1, \cdots, p$. In addition, to describe uncertainty in the estimates, we compute credible intervals based on the highest posterior density interval for each predictor β_j , denoted by HDI_j. The empirical coverage (EC), defined as:

$$\text{EC}(\boldsymbol{\beta}) = \frac{1}{p}\sum_{j=1}^{p}\mathbb{1}_{\beta_j^0\in\text{HDI}_j},$$

assesses the uncertainty provided by the credible intervals.

To assess the prediction for the outcome measures, $\boldsymbol{y}$, two cases are considered, for regression and classification. We define x_i^{new} as new images for $i = n+1, \cdots, n+n_{\text{new}}$

with n_{new} denotes the size of new data. First, for regression, we consider the mean squared prediction error (MSPE):

$$\text{MSPE} = \frac{1}{n_{\text{new}}} \sum_{i=1}^{n_{\text{new}}} \left(y_{n+i}^{0} - \widehat{y}_{n+i}, \right)^{2},$$

where y_i^0 denotes the true outcome measure for $i = 1, \cdots, n_{\text{new}}$. We also assess uncertainty through the empirical coverage (EC):

$$\text{EC}(\boldsymbol{y}) = \frac{1}{n_{\text{new}}} \sum_{i=1}^{n_{\text{new}}} \mathbb{1}_{y_{n+i}^{0} \in \text{HDI}_i},$$

where credible intervals based on the high posterior density interval for each outcome measure y_i, denoted by HDI_i. Furthermore, we employ the negative log likelihood which is defined as:

$$\text{NLL} = \frac{n_{\text{new}}}{2} \log(2\pi) + \frac{n_{\text{new}}}{2} \log(\widehat{\sigma}^2) + \frac{1}{2\widehat{\sigma}^2} \sum_{i=1}^{n_{\text{new}}} \left\{ y_{n+i}^{0} - (\boldsymbol{w}_{n+i}^{T} \widehat{\boldsymbol{\mu}} + \boldsymbol{x}_{n+i}^{T} \widehat{\boldsymbol{\beta}}) \right\}^{2}.$$

For classification, we first define the true positives (TP), false negatives (FN), true negatives (TN) and false positives (FP):

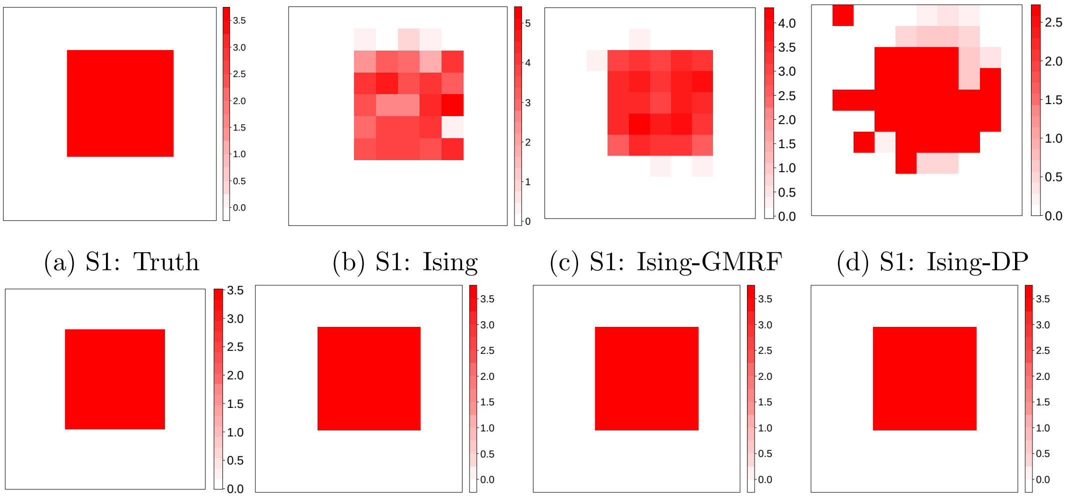
$$\begin{aligned}
\text{TP} &= \sum_{i=1}^{n_{\text{new}}} \mathbb{1}_{\left(\widehat{p}_{n+i}(1) \geq 0.5\right) \cup \left(y_{n+i}^{0} == 1\right)}, \\
\text{FN} &= \sum_{i=1}^{n_{\text{new}}} \mathbb{1}_{\left(\widehat{p}_{n+i}(0) \geq 0.5\right) \cup \left(y_{n+i}^{0} == 1\right)}, \\
\text{TN} &= \sum_{i=1}^{n_{\text{new}}} \mathbb{1}_{\left(\widehat{p}_{n+i}(0) \geq 0.5\right) \cup \left(y_{n+i}^{0} == 0\right)}, \\
\text{FP} &= \sum_{i=1}^{n_{\text{new}}} \mathbb{1}_{\left(\widehat{p}_{n+i}(1) \geq 0.5\right) \cup \left(y_{n+i}^{0} == 0\right)}.
\end{aligned}$$

These quantities are then used in the following evaluation metrics: the sensitivity, specificity, classification accuracy (CA) and cross entropy (CE), as follows:

$$\begin{aligned}\text{Sensitivity} &= \frac{\text{TP}}{\text{TP}+\text{FN}},\\ \text{Specificity} &= \frac{\text{TN}}{\text{TN}+\text{FP}},\\ \text{CA} &= \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}},\\ \text{CE} &= \sum_{i=1}^{n_{\text{new}}} -\left\{y^0_{n+i}\log\widehat{p}_{n+i}(1)+(1-y^0_{n+i})\log\widehat{p}_{n+i}(0)\right\}.\end{aligned}$$

5.6 Results on simulated datasets

In this section, the performance of the proposed models: EPA SIR and Potts-Gibbs SIR models are evaluated and compared to three competitors: Ising model (Huang et al., 2013), Ising-GMRF model (Goldsmith et al., 2014) and Ising-DP model (Li et al., 2015), using four scenarios of simulated datasets.



(a) S1: Truth (b) S1: Ising (c) S1: Ising-GMRF (d) S1: Ising-DP

(e) S1: EPA SIR (f) S1: Potts-DP SIR (g) S1: Potts-PY SIR (h) S1: Potts-MFM SIR

Figure 5.4: Figures showing the true and estimated mean coefficient maps for **Scenario 1** under each model.

Table 5.3 and 5.4 present the numerical results for Scenarios 1 to 4. As shown in Figure 5.4, the proposed models can detect perfectly the cluster structure under Scenario 1 while the Ising and Ising-GMRF models deviate slightly from the true cluster structure. The lowest ARI is observed for the Ising-DP model which is 0.589, as the model cannot learn the cluster structure well. Regarding other summary statistics, overall the proposed

			Scenario 1					
	MSE	EC($\boldsymbol{\beta}$)	MSPE	EC($\boldsymbol{y}$)	NLL	ARI	VI	M
EPA SIR	**8.05e-5**	1.0	4.189	0.923	643.517	**1.0**	2.22e-16	2.592
Potts-DP SIR	**1.18e-4**	1.0	4.188	0.926	643.625	**1.0**	2.22e-16	2.098
Potts-PY SIR	**1.29e-4**	1.0	4.190	0.926	643.800	**1.0**	2.22e-16	2.127
Potts-MFM SIR	**8.25e-4**	0.750	4.177	0.920	642.972	**1.0**	2.22e-16	2.0
Ising	0.279	0.970	6.024	0.913	722.267	0.764	0.521	2.0
Ising-GMRF	0.037	0.980	4.674	0.930	664.172	**1.0**	2.22e-16	2.0
Ising-DP	1.022	0.650	120.566	0.940	1144.521	0.589	0.929	4.725

			Scenario 2					
	MSE	EC($\boldsymbol{\beta}$)	MSPE	EC($\boldsymbol{y}$)	NLL	ARI	VI	M
EPA SIR	0.076	1.0	0.894	0.900	423.155	**0.890**	0.258	8.867
Potts-DP SIR	0.187	0.880	1.328	0.893	529.076	**0.725**	0.734	7.261
Potts-PY SIR	0.113	0.840	0.775	0.923	466.866	**0.805**	0.715	4.954
Potts-MFM SIR	0.087	0.260	1.000	0.876	534.353	**0.835**	0.554	5.195
Ising	0.140	0.790	0.930	0.966	432.696	0.0	1.560	1.996
Ising-GMRF	0.107	0.810	0.898	0.976	424.617	0.0	1.560	2.0
Ising-DP	0.621	0.500	11.251	0.966	788.934	0.262	2.238	5.0

			Scenario 3					
	MSE	EC($\boldsymbol{\beta}$)	MSPE	EC($\boldsymbol{y}$)	NLL	ARI	VI	M
EPA SIR	3.49e-6	0.994	**4.04e-4**	0.916	-609.307	**0.922**	0.268	11.614
Potts-DP SIR	1.76e-7	1.0	**3.33e-4**	0.924	-644.584	**0.993**	0.031	3.593
Potts-PY SIR	4.65e-7	1.0	**3.39e-4**	0.924	-642.541	**0.993**	0.031	2.994
Potts-MFM SIR	1.31e-6	0.995	**3.49e-4**	0.924	-637.069	**0.976**	0.105	2.0
Ising	3.63e-5	1.0	0.070	1.0	238.582	0.0	0.963	2.0
Ising-GMRF	3.52e-5	1.0	0.068	1.0	238.330	0.0	0.963	2.0
Ising-DP	6.25e-4	0.631	0.001	0.768	-280.106	0.0	0.963	9.646

Table 5.3: Mean squared error (MSE), empirical coverage for $\boldsymbol{\beta}$ (EC($\boldsymbol{\beta}$)), mean squared prediction error (MSPE), empirical coverage for $\boldsymbol{y}$ (EC($\boldsymbol{y}$)), negative log-likelihood (NLL), adjusted Rand index (ARI), variation information (VI) and the number of clusters (M) for Scenario 1-3 under each model.

models achieve better results compared to the competing models, especially the Ising-DP model.

When we increase the number of clusters to 5 for Scenario 2, the proposed models are still capable of capturing and identifying the more complex cluster structure underlying the data with a posterior ARI around 0.725 - 0.890. On the contrary, the competing models: Ising, Ising-GMRF and Ising-DP models perform worse than the proposed models. Their ARIs drop to 0.0 for the Ising and Ising-GMRF models and 0.262 for

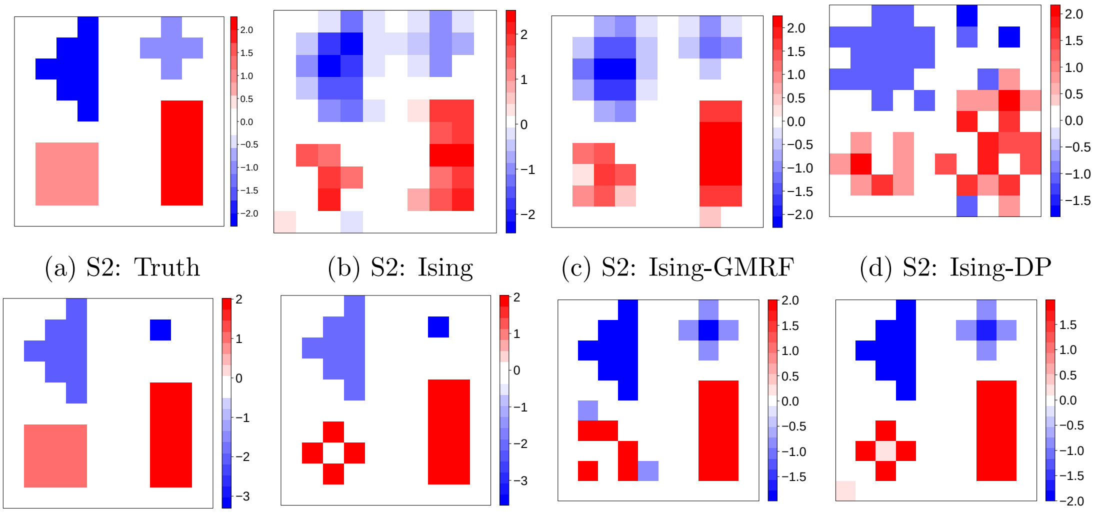
		Scenario 4			
	MSE	EC ($\boldsymbol{\beta}$)	ARI	VI	M
EPA SIR	**0.188**	0.611	**0.472**	1.498	8.528
Potts-DP SIR	**0.287**	0.816	**0.550**	1.423	7.422
Potts-PY SIR	**0.18**8	0.831	**0.601**	1.280	8.296
Potts-MFM SIR	**0.258**	0.579	**0.546**	1.343	4.596
Ising	0.875	0.612	0.0	0.963	2.0
Ising-GMRF	0.890	0.612	0.0	0.963	2.0
Ising-DP	0.875	0.588	0.0	0.963	9.975
		CA	CE	Sensitivity	Specificity
EPA SIR		97.92	84.754	0.978	0.980
Potts-DP SIR		97.76	108.7813	0.972	0.978
Potts-PY SIR		97.60	173.767	0.973	0.983
Potts-MFM SIR		97.28	183.275	0.967	0.978
Ising		87.84	453.759	0.889	0.866
Ising-GMRF		80.96	644.366	0.815	0.803
Ising-DP		94.40	400.253	0.944	0.943

Table 5.4: Mean squared error (MSE), empirical coverage (EC) for $\boldsymbol{\beta}$, adjusted Rand index (ARI), variation of information (VI), the number of clusters, M, classification accuracy (CA), cross-entropy (CE), sensitivity and specificity for Scenario 4 under each model.

the Ising-DP model. The coefficient map under the Ising-DP model looks messy. We can certainly pick up some clusters by looking at the coefficient maps, although the competing models generally fail to determine the exact boundaries of those predefined clusters.

Next, we evaluate the clustering ability of the proposed models when increasing the dimension of the data to a larger value. Figure 5.6 illustrates how the Potts-MFM SIR model performs when dealing with increasing p for the dimension of data varying from 10×10 to 90×90, exploring a range of $n/p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. When the dimension is small, the norm coefficient and norm predicted are still acceptable. When the model becomes too complex, the performance of model deteriorates rapidly. From the experiments conducted, we observe that generally, the results are still good enough when $n/p \geq 0.5$. Thus, for the large p, we group pixels into regions that have characteristics in common, known as superpixels to reach a desirable n/p ratio. For instance, if we have $n = 0.1 \times p$, then the number of superpixels that we use is roughly $0.2 \times p$ to achieve $n/p = (0.1 \times p)/(0.2 \times p) = 0.5$. Each image is pre-segmented into the desired number of superpixels using the build-in function in Matlab (Mori, 2005).

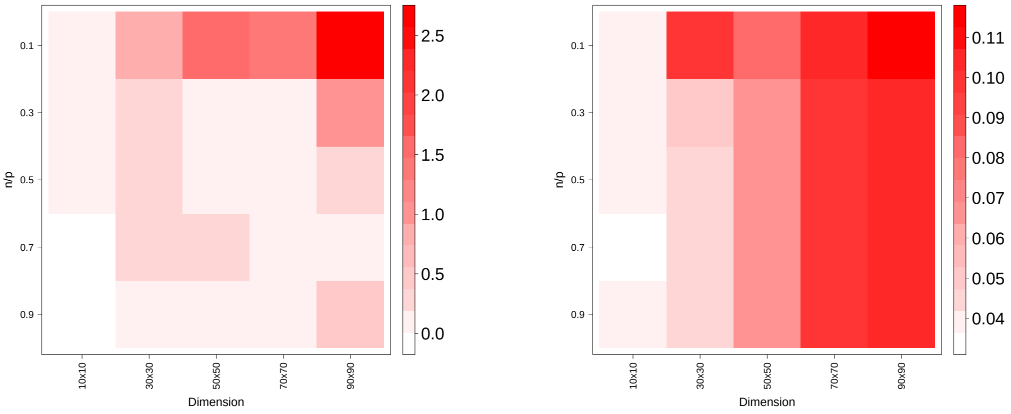
Next, we fix the dimension to 50×50 to investigate the performance of the proposed models on a larger dimension (Scenario 3). In this case, we have used the superpixels



(a) S2: Truth (b) S2: Ising (c) S2: Ising-GMRF (d) S2: Ising-DP

(e) S2: EPA SIR (f) S2: Potts-DP SIR (g) S2: Potts-PY SIR (h) S2: Potts-MFM SIR

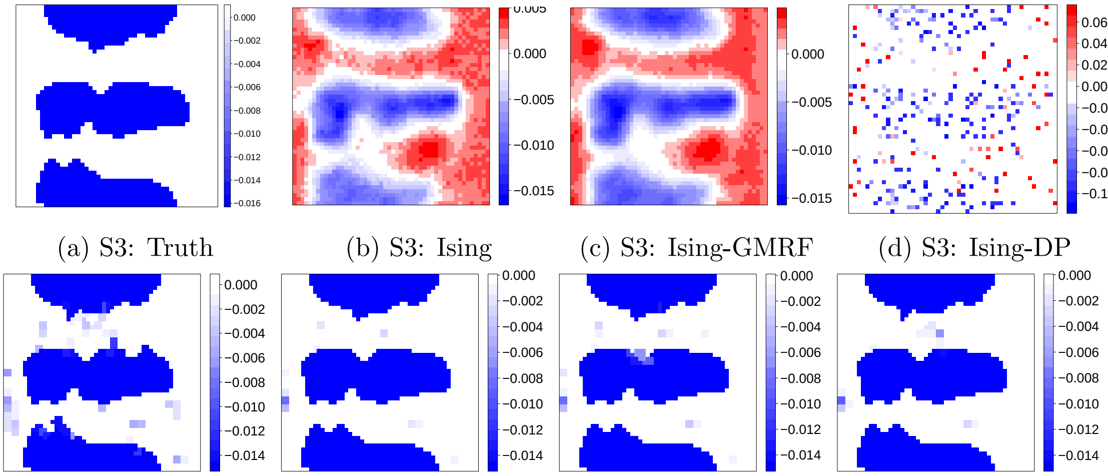
Figure 5.5: Figures showing the true and estimated mean coefficient maps for **Scenario 2** under each model.



(a) Norm coefficient (b) Norm predicted

Figure 5.6: Analysis on large p small n for **Scenario 3** (**continuous case with** $n = 0.1p$) with using **Potts-MFM SIR** model without using superpixels.

equal to $0.2 \times p$. Figure 5.7 suggests that the proposed models are still capable to learn the cluster structure in the coefficients. It can be seen that under the proposed model, most of the resulting clusters are spatially proximal. Compared to the proposed models, the ARI of the competing models drops from 0.922 - 0.993 to zero. The coefficient maps



(a) S3: Truth (b) S3: Ising (c) S3: Ising-GMRF (d) S3: Ising-DP

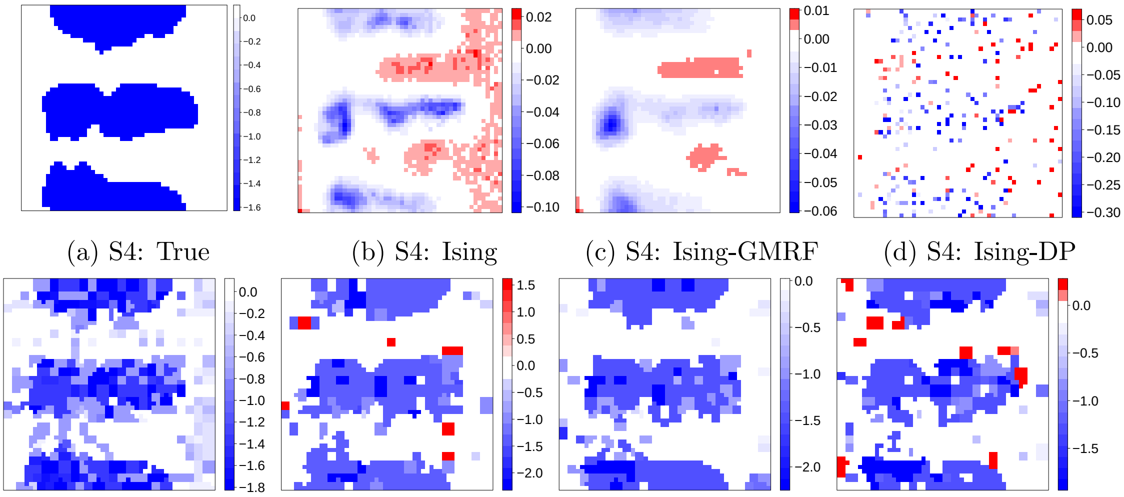
(e) S3: EPA SIR (f) S3: Potts-DP SIR (g) S3: Potts-PY SIR (h) S3: Potts-MFM SIR

Figure 5.7: Figures showing the true and estimated mean coefficient maps for **Scenario 3** (**continuous case with** $n = 0.1p$) under each model.

for the Ising and Ising-GMRF models appear to be able to detect the correct location, however, the clusters formed are quite dispersed throughout the image, causing the low values of the ARI. For the Ising-DP model, the units that are far away are grouped into one cluster. The proposed models also perform better in prediction as we observe that they achieve a smaller posterior mean of MSPE (3.33e-4 - 4.04e-4) compared to competitors (0.001 - 0.070).

When considering the binary case, the proposed models perform slightly poorly compared to the continuous case yet are still acceptable, with the ARI reduced to 0.472 to 0.601. This is not surprising since we will lose information when analysing the binary data. On the other hand, the performance of the competing models is not satisfactory, as illustrated in Figure 5.8. We observe that both Ising and Ising-GMRF produce a set of small and fragmented segments while a few scattered points are observed in the centre for the Ising-DP. The Ising-DP suffers from lower sensitivity and specificity compared with all other models. The proposed models have only a subtle difference in most of the evaluation metrics used as displayed in Table 5.4. Much greater differences are seen in the posterior mean of MSE when compared to the competing models, which suffer from higher error. Still, both the proposed models and competitors can achieve high sensitivity and specificity, generally higher than 0.900.

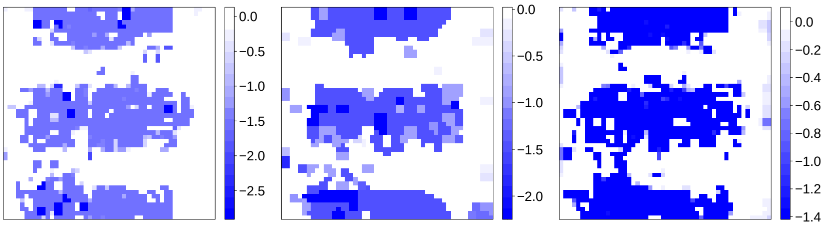
Figure 5.9 shows the estimated mean coefficient maps for Scenario 4 using a larger number of superpixels, increasing from $0.2 \times p$ to $0.5 \times p$ under the Potts-Gibbs SIR models. The proposed models are still able to learn the underlying cluster structure, but we will stick to a smaller number of superpixels as the computational cost is cheaper.



(a) S4: True (b) S4: Ising (c) S4: Ising-GMRF (d) S4: Ising-DP

(e) S4: EPA SIR (f) S4: Potts-DP SIR (g) S4: Potts-PY SIR (h) S4: Potts-MFM SIR

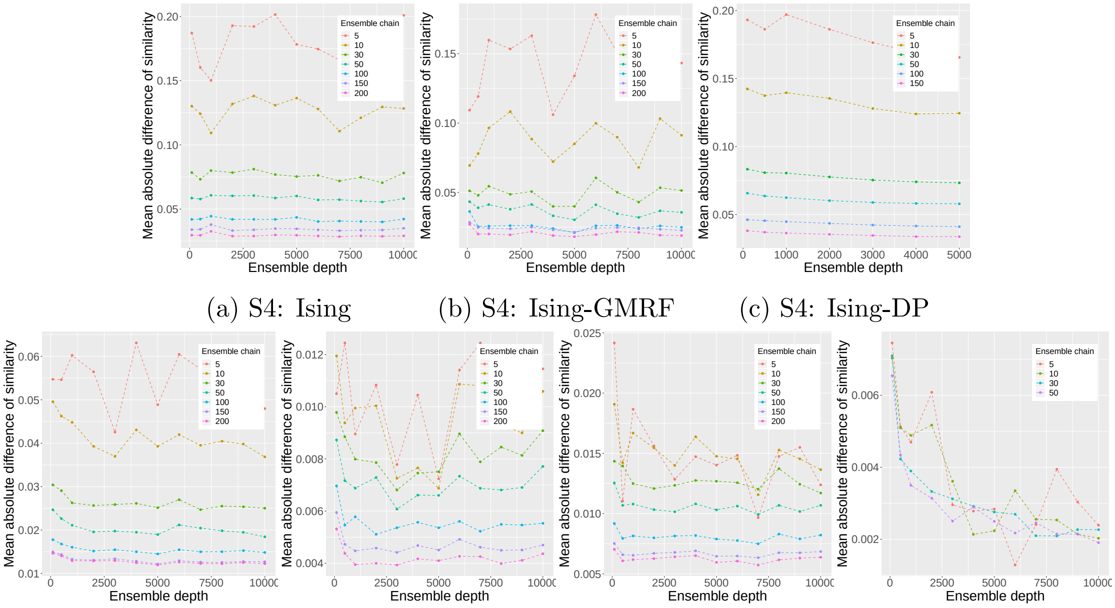
Figure 5.8: Figures showing the true and estimated mean coefficient maps for **Scenario 4** (**Binary case with** $n = 0.5p$) under each model.



(a) S4: Potts-DP SIR (b) S4: Potts-PY SIR (c) S4: Potts-MFM SIR

Figure 5.9: Figures showing estimated mean coefficient maps for **Scenario 4** (**Binary case with** $n = 0.5p$) under **Potts-DP** using a larger number of superpixels, $0.5 \times p$.

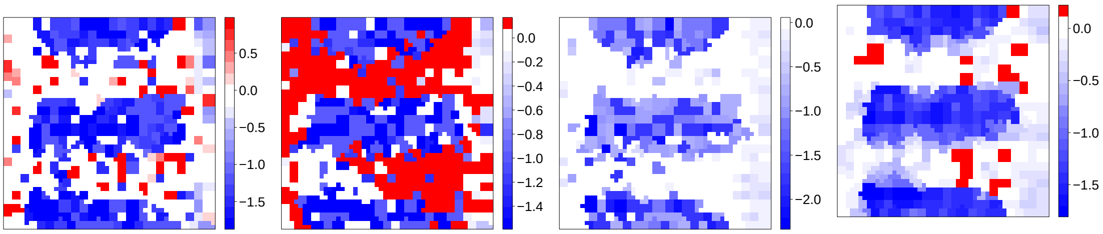
We also investigate the usage of the consensus clustering detailed in Section 4.3.4. For this illustrative purpose, we consider only the simulated dataset under Scenario 4. Figure 5.10 shows the mean absolute difference between the sequential consensus matrices for Scenario 4 under each model for an ensemble chain up to 200 and ensemble depth up to 10,000. As mentioned in Section 4.3.4, the mean absolute difference between the sequential consensus matrices is used as a criterion to obtain the optimum value of ensemble depth and ensemble chain. For the Ising, Ising-GMRF and Ising-DP models, the optimum ensemble chain is around 100 whereas for the EPA SIR, Potts-DP SIR and Potts-PY SIR models, the optimum ensemble chain is around 150 as we observe there is not much variation in the mean absolute difference between the sequential consensus



(a) S4: Ising (b) S4: Ising-GMRF (c) S4: Ising-DP

(d) S4: EPA SIR (e) S4: Potts-DP SIR (f) S4: Potts-PY SIR (g) S4: Potts-MFM SIR

Figure 5.10: Figures showing the mean absolute difference between the sequential consensus matrices for **Scenario 4** (**Binary case with** $n = 0.5p$) under each model.



(a) S4: EPA SIR (b) S4: EPA SIR (c) S4: EPA SIR (d) S4: EPA SIR

Figure 5.11: Figures showing the coefficient maps for **Scenario 4** (**Binary case with** $n = 0.5p$) under EPA SIR model obtained from running one chain with 10,000 iterations **(a)-(c)** and the average coefficient map obtained from consensus clustering **(d)**.

matrices after the 100/150 ensemble chains as shown in Figure 5.10. For the Potts-MFM SIR, we would say that it is probably unnecessary to run for consensus clustering as there is little difference in the graph across different ensemble chains. We also plot the coefficient maps under the EPA SIR model to compare the coefficient maps obtained from running for one chain with 10,000 iterations (Figure 5.11 (a) - (c)) and the average coefficient map obtained from consensus clustering (Figure 5.11 (d)). We observe that the estimated mean coefficient map is more stable when using consensus clustering especially when the model does not converge to a good optimum and remains stuck around local maxima as depicted in Figure 5.11 (b). Besides, consensus clustering is helpful, especially when dealing with the Ising-DP model due to its computation time. It takes roughly 24 hours to run 5,000 iterations for Scenario 4, thus we can save computation time by running shorter iterations in parallel instead of one long chain.

5.7 Results on the ADNI dataset

In the following section, the performance of the proposed models are evaluated and compared to three competitors using ADNI datasets. We present the results from analysing the ordinal outcome measures (CN vs MCI vs AD). On top of that, we analyse the different combinations of binary responses which are obtained by transforming the ordinal outcome measures into three different binary pairs: CN vs AD, CN vs MCI and MCI vs AD.

Table 5.5 reports the posterior mean of the estimated coefficient for each covariate (age, gender and APOE e4 carriers) for both the binary and ordinal responses of the ADNI dataset under each model. As depicted in Table 5.5, we find that age has a negative effect on the outcome measures whereas gender and APOE e4 carriers have a positive effect on the outcome measures, except for the binary response with CN vs MCI which gender has a negative effect on the outcome measures. There is not much difference among the models regarding the estimated coefficient for the covariates for each different combination of outcome measures. For instance, for the binary response with CN vs AD, we observe the posterior mean for age ranges from -1.008 to -0.134, gender ranges from 0.137 to 0.441, APOE e4 carriers (e3/e4) range from 0.264 to 0.966, and APOE e4 carriers (e4/e4) ranges from 0.412 to 1.971.

The numerical results, including the posterior mean of classification accuracy (CA), cross-entropy (CE), sensitivity, specificity and the number of clusters, M, are tabulated in Table 5.6. The posterior mean number of clusters, M, varies differently between the models. For the Ising and Ising-GMRF models, the maximum number of clusters is two because of the underlying property of the Ising prior while for other models, the number of clusters ranges from 6.655 to 17.984. From the numerical results, we observe that there is no clear winner in the comparison between the proposed models and the competitors.

For the binary response with CN vs AD, we see that in general, the results are better for each model compared to the probit regression, with classification accuracy increasing greatly from 65.822 to 78.481 - 82.278. For the binary response with CN vs MCI and MCI vs AD, the classification accuracy for the probit regression is almost similar to the proposed models and competitors but we observe that the probit regression produces a very low specificity (0.340) for CN vs MCI and a very low sensitivity (0.0) for MCI vs

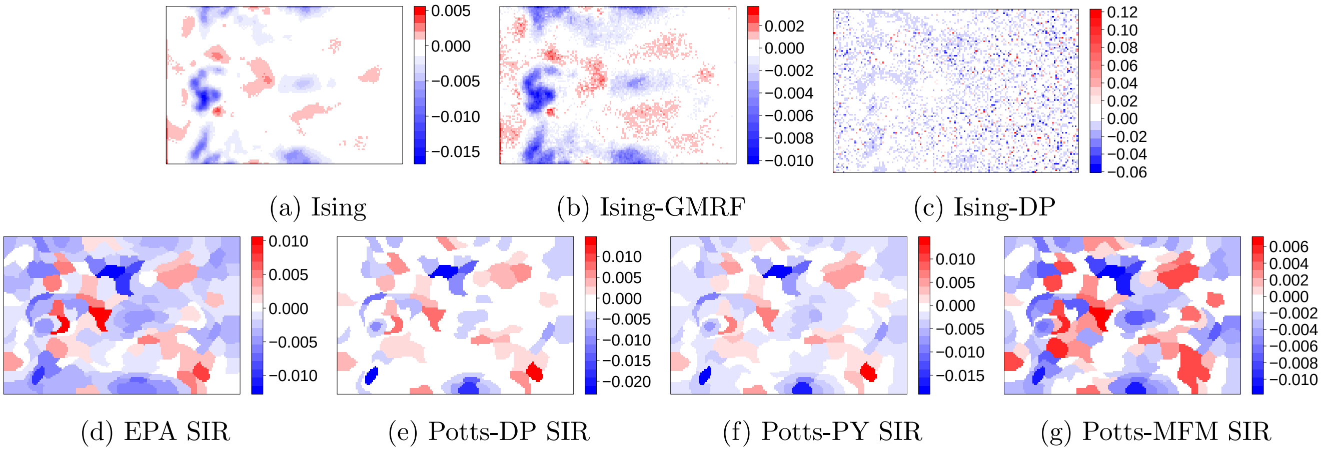
			ADNI (**CN vs AD**)		
	Intercept	Age	Gender	APOE e4 carriers (e3/e4)	APOE e4 carriers (e4/e4)
EPA SIR	-7.767 (4.115)	-1.008 (0.300)	0.432 (0.191)	0.846 (0.197)	1.422 (0.403)
Potts-DP SIR	-9.173 (4.115)	-0.991 (0.296)	0.441 (0.196)	0.833 (0.198)	1.416 (0.416)
Potts-PY SIR	-8.834 (3.634)	-0.969 (0.290)	0.430 (0.190)	0.833 (0.194)	1.407 (0.408)
Potts-MFM SIR	-9.257 (3.659)	-1.002 (0.293)	0.435 (0.190)	0.854 (0.195)	1.408 (0.402)
Probit regression	-0.672	-0.134	0.137	0.966	1.971
Ising	-0.361 (0.163)	-0.785 (0.254)	0.302 (0.168)	0.748 (0.183)	1.212 (0.362)
Ising-GMRF	-0.370 (0.170)	-0.771 (0.246)	0.295 (0.167)	0.748 (0.177)	1.232 (0.349)
Ising-DP	-4.338 (0.493)	-0.369 (0.130)	0.157 (0.075)	0.264 (0.115)	0.412 (0.190)
			ADNI (**CN vs MCI**)		
	Intercept	Age	Gender	APOE e4 carriers (e3/e4)	APOE e4 carriers (e4/e4)
EPA SIR	-2.716 (2.178)	-0.748 (0.196)	-0.109 (0.132)	0.495 (0.140)	1.127 (0.320)
Potts-DP SIR	-3.628 (1.936)	-0.733 (0.194)	-0.109 (0.133)	0.502 (0.133)	1.112 (0.315)
Potts-PY SIR	-3.650 (1.858)	-0.729 (0.192)	-0.109 (0.132)	0.502 (0.141)	1.116 (0.317)
Potts-MFM SIR	-3.941 (1.793)	-0.741 (0.193)	-0.106 (0.132)	0.504 (0.140)	1.121 (0.315)
Probit regression	0.216	-0.265	-0.245	0.596	1.371
Ising	-0.147 (0.104)	-0.649 (0.176)	-0.123 (0.127)	0.498 (0.138)	1.024 (0.296)
Ising-GMRF	-0.154 (0.105)	-0.639 (0.105)	-0.121 (0.126)	0.501 (0.137)	1.029 (0.295)
Ising-DP	-1.260 (0.596)	-0.867 (0.205)	-0.131 (0.139)	0.498 (0.147)	1.052 (0.314)
			ADNI (**MCI vs AD**)		
	Intercept	Age	Gender	APOE e4 carriers (e3/e4)	APOE e4 carriers (e4/e4)
EPA SIR	-1.658 (1.679)	-0.101 (0.173)	0.384 (0.130)	0.308 (0.138)	0.340 (0.191)
Potts-DP SIR	-2.951 (1.510)	-0.075 (0.173)	0.396 (0.130)	0.321 (0.138)	0.343 (0.191)
Potts-PY SIR	-2.929 (1.419)	-0.070 (0.172)	0.395 (0.130)	0.320 (0.137)	0.345 (0.189)
Potts-MFM SIR	-2.878 (1.416)	-0.073 (0.172)	0.386 (0.130)	0.316 (0.139)	0.337 (0.191)
Probit regression	-0.865	0.169	0.344	0.347	0.453
Ising	-0.263 (0.127)	0.044 (0.161)	0.328 (0.125)	0.294 (0.135)	0.353 (0.135)
Ising-GMRF	-0.259 (0.122)	0.040 (0.161)	0.327 (0.124)	0.296 (0.135)	0.362 (0.184)
Ising-DP	-3.981 (1.029)	-0.071 (0.183)	0.393 (0.133)	0.288 (0.147)	0.282 (0.196)
			ADNI (**Ordinal**)		
	Intercept	Age	Gender	APOE e4 carriers (e3/e4)	APOE e4 carriers (e4/e4)
EPA SIR	-2.552 (1.726)	-0.506 (0.136)	0.148 (0.096)	0.524 (0.102)	0.668 (0.160)
Potts-DP SIR	-3.479 (1.725)	-0.498 (0.137)	0.145 (0.095)	0.522 (0.102)	0.661 (0.158)
Potts-PY SIR	-3.396 (1.629)	-0.500 (0.136)	0.147 (0.096)	0.524 (0.102)	0.664 (0.161)
Potts-MFM SIR	-3.518 (1.575)	-0.503 (0.137)	0.149 (0.096)	0.524 (0.101)	0.667 (0.160)
Ordinal logistic regression	0.4315	-0.10783	0.04694	1.08326	1.60946
Ising	-0.051 (0.097)	-0.879 (0.163)	0.159 (0.099)	0.435 (0.107)	0.282 (0.171)
Ising-GMRF	-0.305 (0.119)	-0.271(0.131)	0.085 (0.091)	0.568 (0.098)	0.815 (0.154)
Ising-DP	-131.492 (34.461)	-0.249 (0.979)	0.269 (1.009)	0.369 (1.002)	0.285 (1.004)

Table 5.5: The posterior mean and standard deviation (in parentheses) of the estimated coefficient for the covariates for both the binary and ordinal responses of ADNI datasets under each model.

AD. For the ordinal response, the ordinal logistic regression classifies all the test data into one class, MCI.

For both the proposed models and the competitors, the performance when considering the binary response with CN vs MCI is slightly better than the binary response with MCI vs AD as shown in Table 5.6. When considering binary response with MCI vs AD, the sensitivity drops abundantly for all the models (0.085 - 0.285) but still has high specificity (0.896 - 1.0). All the models are better at identifying AD from CN than identifying AD from MCI. The results are not surprising as MCI is an intermediate state between CN and AD, thereby causing difficulty in distinguishing AD from CN and MCI. Lastly, the lowest classification accuracy is observed when applying the models to the ordinal responses (49.358 - 54.487).

Figure 5.12 - 5.15 plots the estimated mean coefficient maps of each model for both the binary and ordinal responses of ADNI datasets. The coefficient values for all the models are roughly similar for each case. However, the coefficient maps might look slightly different among the models as the underlying assumptions among the models differ. Still, we can see that the coefficient maps share some characteristics as the regions that are predicted to have positive (negative) effects under one model also are likely observed to have positive (negative) effects under another model. We notice that the number of coefficients with negative coefficient estimates is larger than the number of positive estimates. This matches the scientific hypothesis that AD will induce hippocampal shrinkage.



(a) Ising

(b) Ising-GMRF

(c) Ising-DP

(d) EPA SIR

(e) Potts-DP SIR

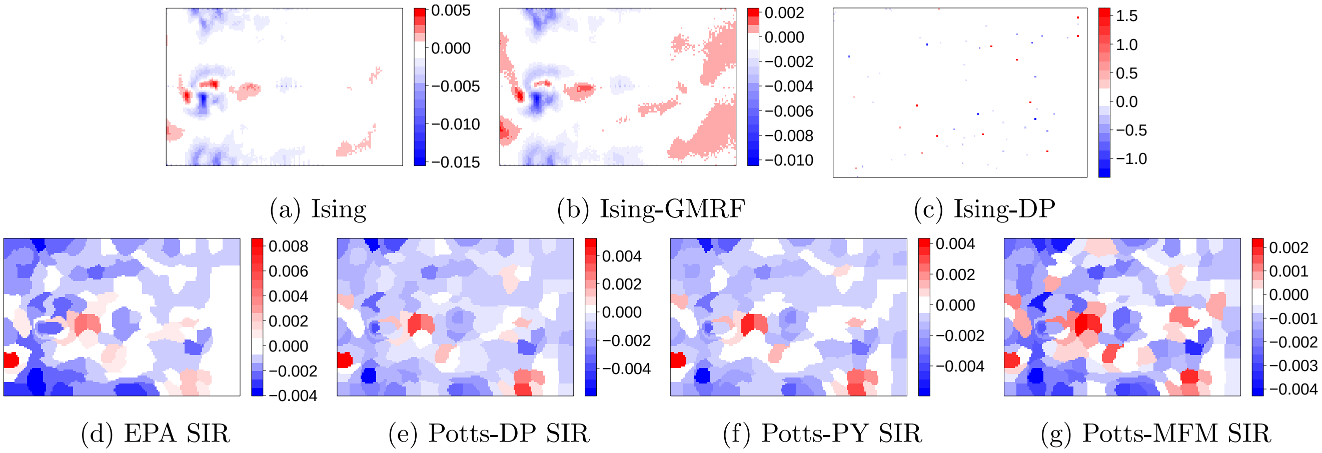
(f) Potts-PY SIR

(g) Potts-MFM SIR

Figure 5.12: The estimated mean coefficient maps of each model for the binary response of ADNI datasets (**CN vs AD**).

5.8 Conclusions

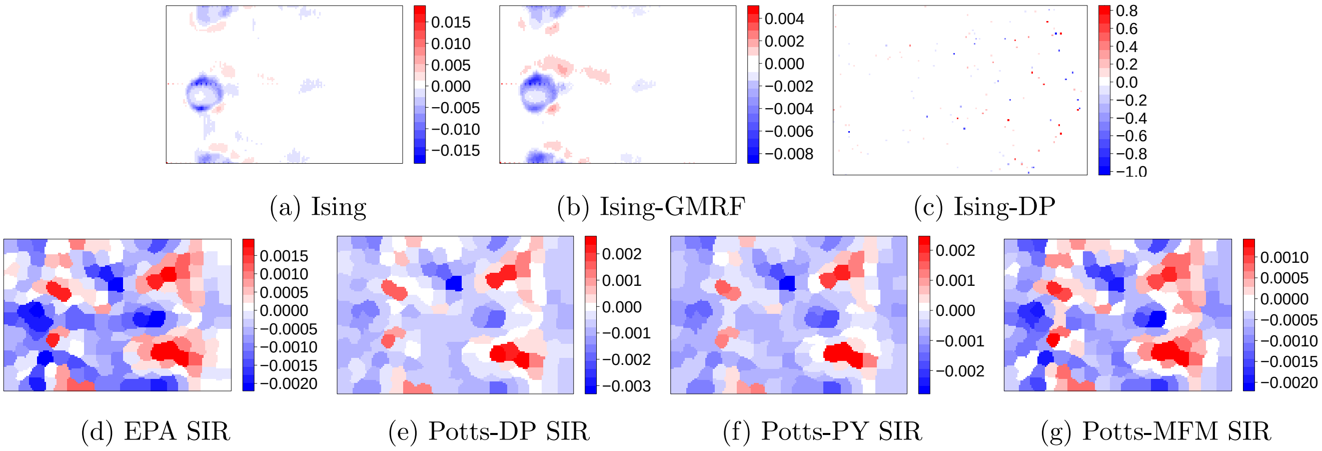
We have demonstrated the good performance of our proposed models on the simulated datasets. The proposed models are able to learn the underlying cluster structure in the data, leading to higher ARI than the competitors. For the study on AD based on neuroimaging data from the ADNI database, our proposed models still perform well but



(a) Ising (b) Ising-GMRF (c) Ising-DP

(d) EPA SIR (e) Potts-DP SIR (f) Potts-PY SIR (g) Potts-MFM SIR

Figure 5.13: The estimated mean coefficient maps of each model for the binary response of ADNI datasets (**CN vs MCI**).

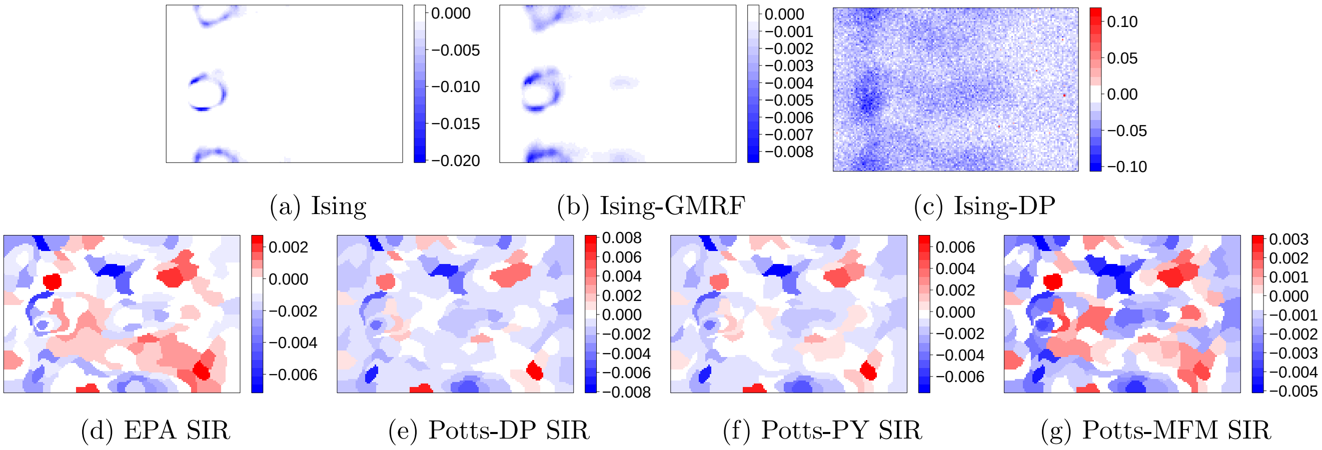


(a) Ising (b) Ising-GMRF (c) Ising-DP

(d) EPA SIR (e) Potts-DP SIR (f) Potts-PY SIR (g) Potts-MFM SIR

Figure 5.14: The estimated mean coefficient maps of each model for the binary response of ADNI datasets (**MCI vs AD**).

we would say that there is no single model that stands out from others. Most of the models do well in differentiating CN and AD, but not doing great in differentiating MCI and AD or differentiating CN and MCI.



(a) Ising (b) Ising-GMRF (c) Ising-DP

(d) EPA SIR (e) Potts-DP SIR (f) Potts-PY SIR (g) Potts-MFM SIR

Figure 5.15: The estimated mean coefficient maps of each model for the ordinal response of ADNI datasets. (**CN vs MCI vs AD**)

	ADNI (**CN vs AD**)				
	CA	CE	Sensitivity	Specificity	M
EPA SIR	78.481	37.084	0.685	0.863	6.655
Potts-DP SIR	81.012	36.582	0.714	0.886	13.536
Potts-PY SIR	81.012	36.355	0.714	0.886	17.984
Potts-MFM SIR	79.746	37.450	0.685	0.886	6.596
Probit regression	65.822	-	0.636	0.685	-
Ising	82.278	35.924	0.714	0.909	2.0
Ising-GMRF	81.012	36.224	0.685	0.909	2.0
Ising-DP	81.012	42.567	0.714	0.886	10.0
	ADNI (**CN vs MCI**)				
	CA	CE	Sensitivity	Specificity	M
EPA SIR	70.247	69.878	0.740	0.636	5.814
Potts-DP SIR	71.074	69.789	0.753	0.636	16.533
Potts-PY SIR	71.900	69.744	0.766	0.636	16.533
Potts-MFM SIR	71.074	69.803	0.753	0.636	5.999
Probit regression	70.247	-	0.909	0.340	-
Ising	69.421	69.421	0.779	0.545	2.0
Ising-GMRF	66.942	66.942	0.766	0.500	2.0
Ising-DP	69.421	68.541	0.753	0.590	10.695
	ADNI (**MCI vs AD**)				
	CA	CE	Sensitivity	Specificity	M
EPA SIR	67.857	66.996	0.114	0.935	5.455
Potts-DP SIR	67.857	67.857	0.114	0.935	12.060
Potts-PY SIR	68.750	67.016	0.114	0.948	15.108
Potts-MFM SIR	66.964	66.974	0.114	0.922	6.613
Probit regression	68.750	-	0.0	1.0	-
Ising	66.071	67.537	0.085	0.922	2.0
Ising-GMRF	68.750	67.561	0.085	0.961	2.0
Ising-DP	70.535	67.272	0.285	0.896	11.757
	ADNI (**Ordinal**)				
	CA	CE	Sensitivity	Specificity	M
EPA SIR	54.487	143.509	-	-	6.001
Potts-DP SIR	53.846	143.764	-	-	12.738
Potts-PY SIR	53.846	143.626	-	-	16.738
Potts-MFM SIR	54.487	143.833	-	-	6.500
Ordinal logistic regression	49.358	-	-	-	-
Ising	49.358	161.031	-	-	2.0
Ising-GMRF	52.564	147.775	-	-	2.0
Ising-DP	53.205	177.232	-	-	12.412

Table 5.6: The posterior mean of classification accuracy (CA), cross-entropy (CE), sensitivity, specificity and the number of clusters (M) for both the binary and ordinal responses of ADNI datasets under each model.

Chapter 6

Discussion and Future Work

We have developed novel Bayesian scalar-on-image regression models, that employ clustering and exhibit spatial dependence by leveraging the spatial coordinates of the pixels. Specifically, the models group pixels with similar effects on the response have a common coefficient within the model. To encourage groups representing spatially contiguous regions, we incorporate the spatial information directly in the prior for the random partition by utilising the Ewens-Pitman attraction distribution and the Gibbs-type random partition models. On top of that, the Bayesian shrinkage priors are utilised to identify the covariates and regions that are most relevant for the prediction. The procedure yields a coefficient image that is both sparse and spatially smooth. Also, we have derived an MCMC sampler based on this representation.

We have shown the potential of the proposed models in detecting the cluster structure on the simulated and real datasets and allows for automatic extraction of regions of interest from the image. By taking into consideration spatial dependence in the random partition model via either the EPA distribution or the Potts-Gibbs models, the proposed models produce spatially aware clustering and thus improve the predictions. The results have shown the proposed models have great potential in recovering the underlying cluster structure under a variety of configurations. Overall, our studies indicate that there is no clear winner between EPA SIR and Potts-Gibbs SIR models. The EPA SIR models allow for learning of key hyperparameters of the random partition model, which have a strong influence of the number and spatial connectivity of the clusters. However, from a computational constraint standpoint, the Potts-Gibbs SIR models are preferred as the computational cost is only linear in the number of pixels compared with the quadratic complexity of the EPA SIR model.

Finally, this thesis paves the way for a number of exciting extensions for further research. Firstly, anatomical changes in the brain associated with the disease may differ across patients due to various factors such as undiscovered genes, comorbidities or lifestyle choices. This heterogeneity can be incorporated through hierarchical extensions that allow for a latent clustering of patients with cluster-specific image partitions. Next, the model has focused on a single tissue density map per patient (based on grey matter), and extensions for tissue density maps of grey matter, white matter, and cerebrospinal fluid are of interest; in particular, this would involve developing dependent image partition models across the three maps. On a related note, it would be desirable to combine data from other imaging modalities, such as amyloid positron emission tomography (PET) imaging, fluorodeoxyglucose uptake on PET (FDG-PET) imaging and functional MRI, and allow for dependence in the random image partitions for different imaging modalities; this would make use of the variety of imaging data available for improved diagnosis. Developments in this direction could combine the random image partition model with multiple partition models for partially exchangeable data (Camerlenghi et al., 2017). A further extension of interest is the inclusion of longitudinal, as well as cross-sectional data, for dynamic modelling of disease status and prediction of conversion to the disease. Finally, for more flexible modelling of the relationship between disease status and imaging data, non-linear models may be utilised, such as Gaussian processes.

Due to the high resolution of the image, more computationally efficient approaches are essential for the application of full imaging data. Optimising the implementation of the proposed models to improve the performance and scalability should be possible based on parallel MCMC algorithms, such as the recent work of Ni et al. (2019). Fast approximate Bayesian inference specifically is another interesting direction, for example, developing novel maximum aposteriori (MAP) algorithms, based on iterated conditional modes for Dirichlet process mixtures (Raykov et al., 2016) and Bayesian hierarchical clustering algorithms (Heller and Ghahramani, 2005).

Chapter 7

Appendix

7.1 Posterior inference

The mathematical derivation of the posterior marginal densities and posterior joint densities of the random variables for the proposed models described in Section 4.3 are provided here.

7.1.1 The update of the parameters μ, β^* and σ^2 given the data y and other parameters

The derivation for the posterior distribution of the parameters $\boldsymbol{\mu}, \boldsymbol{\beta}^*$ and σ^2 are given here. From the prior distributions specified in Equation (4.5), we know the corresponding joint prior density is given by:

$$\begin{aligned}
\mathrm{pr}(\tilde{\boldsymbol{\beta}}, \sigma^2) &= \mathrm{pr}(\tilde{\boldsymbol{\beta}}|\sigma^2)\mathrm{pr}(\sigma^2) \\
&= \mathrm{N}(\tilde{\boldsymbol{\beta}}|\boldsymbol{m}_{\tilde{\beta}}, \sigma^2\Sigma_{\tilde{\beta}})\mathrm{IG}(\sigma^2|a_\sigma, b_\sigma) \\
&\propto (\sigma^2)^{-\frac{M+q}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{m}_{\tilde{\beta}})^T\Sigma_{\tilde{\beta}}(\tilde{\boldsymbol{\beta}} - \boldsymbol{m}_{\tilde{\beta}})\right\}(\sigma^2)^{-(a_\sigma+1)} \exp\left(-\frac{1}{\sigma^2}b_\sigma\right).
\end{aligned}$$

The posterior distribution is given by:

$$\begin{aligned}
\mathrm{pr}(\tilde{\boldsymbol{\beta}}, \sigma^2|\dots) \propto & f(\mathbf{y}|\tilde{\boldsymbol{\beta}}, \sigma^2)\mathrm{pr}(\tilde{\boldsymbol{\beta}}, \sigma^2) \\
= & \prod_{i=1}^{n} \mathrm{N}(y_i|\tilde{\boldsymbol{x}}_i^T\tilde{\boldsymbol{\beta}}, \sigma^2)\mathrm{N}(\tilde{\boldsymbol{\beta}}|\boldsymbol{m}_{\tilde{\beta}}, \sigma^2\Sigma_{\tilde{\beta}})\mathrm{IG}(\sigma^2|a_\sigma, b_\sigma) \\
\propto & (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}})^T(\mathbf{y} - \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}})\right\}\times \\
& (\sigma^2)^{-\frac{M+q}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{m}_{\tilde{\beta}})^T\Sigma_{\tilde{\beta}}^{-1}(\tilde{\boldsymbol{\beta}} - \boldsymbol{m}_{\tilde{\beta}})\right\}\times \\
& (\sigma^2)^{-(a_\sigma+1)} \exp\left(-\frac{1}{\sigma^2}b_\sigma\right) \\
\propto & (\sigma^2)^{-(a_\sigma+\frac{n}{2}+1)} \exp\left\{-\frac{1}{\sigma^2}(b_\sigma + \frac{1}{2}\boldsymbol{m}_{\tilde{\beta}}^T\Sigma_{\tilde{\beta}}^{-1}\boldsymbol{m}_{\tilde{\beta}} + \frac{1}{2}\mathbf{y}^T\mathbf{y})\right\}\times \\
& (\sigma^2)^{-\frac{M+q}{2}} \exp\left[-\frac{1}{2\sigma^2}\left\{\tilde{\boldsymbol{\beta}}^T(\Sigma_{\tilde{\beta}}^{-1} + \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})\tilde{\boldsymbol{\beta}} - 2\tilde{\boldsymbol{\beta}}^T(\Sigma_{\tilde{\beta}}^{-1}\boldsymbol{m}_{\tilde{\beta}} + \tilde{\boldsymbol{X}}^T\mathbf{y})\right\}\right].
\end{aligned}$$

We define $\hat{\Sigma}_{\tilde{\beta}} = (\Sigma_{\tilde{\beta}}^{-1} + \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}$ and $\hat{\boldsymbol{m}}_{\tilde{\beta}} = \hat{\Sigma}_{\tilde{\beta}}(\Sigma_{\tilde{\beta}}^{-1}\boldsymbol{m}_{\tilde{\beta}} + \tilde{\boldsymbol{X}}^T\mathbf{y})$. Thus we have

$$\begin{aligned}
\mathrm{pr}(\tilde{\boldsymbol{\beta}}, \sigma^2|\dots) \propto & (\sigma^2)^{-(a_\sigma+\frac{n}{2}+1)} \exp\left\{-\frac{1}{\sigma^2}(b_\sigma + \frac{1}{2}\boldsymbol{m}_{\tilde{\beta}}^T\Sigma_{\tilde{\beta}}^{-1}\boldsymbol{m}_{\tilde{\beta}} + \frac{1}{2}\mathbf{y}^T\mathbf{y} - \frac{1}{2}\hat{\boldsymbol{m}}_{\tilde{\beta}}^T\hat{\Sigma}_{\tilde{\beta}}^{-1}\hat{\boldsymbol{m}}_{\tilde{\beta}})\right\}\times \\
& (\sigma^2)^{-\frac{M+q}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{m}}_{\tilde{\beta}})^T\hat{\Sigma}_{\tilde{\beta}}^{-1}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{m}}_{\tilde{\beta}})\right\}.
\end{aligned}$$

We obtain two parts of the exponential. The first term is proportional to the marginal posterior density for σ^2 whereas the second term is the conditional posterior for $\tilde{\boldsymbol{\beta}}$. We recognise the posterior distribution for σ^2 to be inverse-gamma with

$$\sigma^2|\dots \sim \mathrm{IG}(\hat{a}_\sigma, \hat{b}_\sigma),$$

where $\mathrm{IG}(\hat{a}_\sigma, \hat{b}_\sigma)$ denotes the inverse-gamma distribution with updated shape $\hat{a}_\sigma = a_\sigma + \frac{n}{2}$ and scale $\hat{b}_\sigma = b_\sigma + \frac{1}{2}\boldsymbol{m}_{\tilde{\beta}}^T\Sigma_{\tilde{\beta}}^{-1}\boldsymbol{m}_{\tilde{\beta}} + \frac{1}{2}\mathbf{y}^T\mathbf{y} - \frac{1}{2}\hat{\boldsymbol{m}}_{\tilde{\beta}}^T\hat{\Sigma}_{\tilde{\beta}}^{-1}\hat{\boldsymbol{m}}_{\tilde{\beta}}$. The conditional posterior distribution for the parameters is as follows:

$$\begin{aligned}
\tilde{\boldsymbol{\beta}}|\boldsymbol{\eta}^*, \sigma^2, \dots \sim & \mathrm{N}(\hat{\boldsymbol{m}}_{\tilde{\beta}}, \sigma^2\hat{\Sigma}_{\tilde{\beta}}), \\
\sigma^2|\dots \sim & \mathrm{IG}(\hat{a}_\sigma, \hat{b}_\sigma).
\end{aligned}$$

7.1.2 The update of the parameter η^* given the other parameters

The derivation for the posterior distribution of the parameter $\boldsymbol{\eta}^*$ is given here. From the prior distributions specified in Equation (4.5), we obtain the posterior distribution for each η^*_m as follows,

$$\begin{aligned}
\mathrm{pr}(\eta^*_1, \ldots, \eta^*_M | \ldots) &= \prod_{m=1}^{M} \mathrm{IG}(\eta^*_m | a_\eta, b_\eta) \prod_{m=1}^{M} \mathrm{N}(\beta^*_m | 0, \eta^*_m \sigma^2) \\
&= \prod_{m=1}^{M} \frac{b_\eta^{a_\eta}}{\Gamma(a_\eta)} \left(\frac{1}{\eta^*_m}\right)^{a_\eta+1} \exp\left(-\frac{b_\eta}{\eta^*_m}\right) \prod_{m=1}^{M} \frac{1}{\sqrt{2\pi\eta^*_m\sigma^2}} \exp\left(-\frac{\beta^{*2}_m}{2\eta^*_m\sigma^2}\right) \\
&\propto \prod_{m=1}^{M} \left(\frac{1}{\eta^*_m}\right)^{a_\eta+1} \exp\left(-\frac{b_\eta}{\eta^*_m}\right) \left(\frac{1}{\eta^*_m}\right)^{\frac{1}{2}} \exp\left(-\frac{\beta^{*2}_m}{2\eta^*_m\sigma^2}\right) \\
&\propto \prod_{m=1}^{M} \left(\frac{1}{\eta^*_m}\right)^{a_\eta+\frac{1}{2}+1} \exp\left(-\frac{b_\eta + \frac{\beta^{*2}_m}{2b^2}}{\eta^*_m}\right).
\end{aligned}$$

We identify this as an inverse-gamma distribution with updated shape $\hat{a}_\eta = a_\eta + 1/2$ and scale $\hat{b}_\eta = b_\eta + \beta^{*2}_m / 2\sigma^2$. And the posterior distribution for b_η is given by:

$$\begin{aligned}
\mathrm{pr}(b_\eta | \cdots) &= \mathrm{G}(b_\eta | a_o, b_o) \prod_{m=1}^{M} \mathrm{IG}(\eta^*_m | a_\eta, b_\eta) \\
&= \frac{b_\eta^{a_o-1} \exp\left(-b_o b_\eta\right) b_o^{a_o}}{\Gamma(a_o)} \prod_{m=1}^{M} \frac{b_\eta^{a_\eta}}{\Gamma(a_\eta)} \left(\frac{1}{\eta^*_m}\right)^{a_\eta+1} \exp\left(-\frac{b_\eta}{\eta^*_m}\right) \\
&\propto b_\eta^{a_o-1} \exp\left(-b_o b_\eta\right) \prod_{m=1}^{M} b_\eta^{a_\eta} \exp\left(-\frac{b_\eta}{\eta^*_m}\right) \\
&\propto b_\eta^{a_o+Ma_\eta-1} \exp\left\{-b_\eta \left(b_o + \sum_{m=1}^{M} \frac{1}{\eta^*_m}\right)\right\}.
\end{aligned}$$

We identify this as an gamma distribution with updated shape $\hat{a}_o = a_o + Ma_\eta$ and rate $\hat{b}_o = b_o + \sum_{m=1}^{M} \frac{1}{\eta^*_m}$. The marginal posterior distribution for the parameters is as follows:

$$\begin{aligned}
\eta^*_m | \cdots &\sim \mathrm{IG}(\hat{a}_\eta, \hat{b}_\eta), \quad \text{for all } m = 1, \ldots, M, \\
b_\eta | \cdots &\sim \mathrm{G}\left(\hat{a}_o, \hat{b}_o\right).
\end{aligned}$$

7.1.3 The update of the bond variables r_{jk} given the partition π_p

The derivation for the posterior distribution of the bond variables, r_{jk}, for $1 \leq j < k \leq p$is given here. To sample from the Potts-Gibbs models, we use the GSW sampler. As previously mentioned in Section 4.3.1, the GSW involves introducing auxiliary binary bond variables, r_{jk}. Based on this, the augmented model can be defined as follows:

$$\begin{aligned}
\mathrm{pr}(\pi_p, \boldsymbol{r}) &= \mathrm{pr}(\pi_p)\mathrm{pr}(\boldsymbol{r}|\pi_p), \\
\mathrm{pr}(\boldsymbol{r}|\pi_p) &= \prod_{1 \le j < k \le p} \mathrm{pr}(r_{jk}|\pi_p), \\
\mathrm{pr}(r_{jk}|\pi_p) &= \begin{cases} \exp(-\upsilon\zeta_{jk}\mathbb{1}_{z_j = z_k}) = q_{jk} & r_{jk} = 0 \\ 1 - \exp(-\upsilon\zeta_{jk}) & r_{jk} = 1, \end{cases}
\end{aligned} \tag{7.1}$$

where $\boldsymbol{r} = (r_{jk})_{1 \le j < k \le p}$.

Since the bonds are independent of the data given the partition π_p, we have $\mathrm{pr}(\boldsymbol{r}|\pi_p, \boldsymbol{y}) = \mathrm{pr}(\boldsymbol{r}|\pi_p)$. Thus, the bond variables can be updated independently using Equation 7.1.

7.1.4 The update of the partition π_p given the data y and other parameters

The derivation for the posterior distribution of the partition π_p is given here. The posterior distribution of $\mathrm{pr}(\pi_p)$ is given by:

$$\mathrm{pr}(\pi_p | \ldots) \propto f(\mathbf{y}|\pi_p, \boldsymbol{\eta}^*, \tilde{\boldsymbol{\beta}}, \sigma^2)\mathrm{pr}(\pi_p|\Theta).$$

For $o = 1, \ldots, O$, each nested cluster A_o is removed in turn and assigned to cluster $(m = 1, \ldots, M^{-A_o}, M^{-A_o} + 1)$ with probability as follows:

$$\mathrm{pr}(A_o \in C_m^{-A_o} | \ldots) \propto f(\mathbf{y}|\pi_p^{A_o \to m}, \boldsymbol{\eta}^*, \tilde{\boldsymbol{\beta}}, \sigma^2)\mathrm{pr}(A_o \in C_m^{-A_o} | \pi_p^{-A_o}, \Theta),$$

for all $m = 1, \ldots, M^{-A_o}, M^{-A_o} + 1$. If the EPA distribution is selected, the $\mathrm{pr}(\pi_p^{A_o \to m}|\Theta)$ can be evaluated using Equation (3.1) and (3.2). This probability is evaluated at the partition $\pi_p^{A_o \to m}$, where each nested cluster is a singleton for the EPA distribution. For the Potts-Gibbs models, the conditional distribution $\mathrm{pr}(\pi_p|\boldsymbol{r},\boldsymbol{y})$ can be expressed as:

$$\begin{aligned}
\mathrm{pr}(\pi_p|\boldsymbol{r},\boldsymbol{y}) &\propto \mathrm{pr}(\pi_p)\mathrm{pr}(\boldsymbol{r}|\pi_p)f(\boldsymbol{y}|\pi_p) \\
&= \mathcal{B}(z_1,\cdots z_p)\mathrm{pr}(|C_1|,\ldots,|C_M|)\prod_{1\le j<k\le p}\mathrm{pr}(r_{jk}|\pi_p)\times f(\boldsymbol{y}|\pi_p) \\
&= \prod_{1\le j<k\le p}\exp\left(\upsilon\mathbb{1}_{z_j=z_k}\right)\mathrm{pr}(|C_1|,\ldots,|C_M|)\times \\
&\quad \prod_{1\le j<k\le p}(1-q_{jk})^{r_{jk}}\exp(-\upsilon\zeta_{jk}\mathbb{1}_{z_j=z_k})^{1-r_{jk}}f(\boldsymbol{y}|\pi_p) \\
&= \prod_{1\le j<k\le p}\exp\left(\upsilon\mathbb{1}_{z_j=z_k}\right)\mathrm{pr}(|C_1|,\ldots,|C_M|)\times \\
&\quad \prod_{1\le j<k\le p}\left\{1-\exp(-\upsilon\zeta_{jk}\mathbb{1}_{z_j=z_k})\right\}^{r_{jk}}\exp(-\upsilon\zeta_{jk}\mathbb{1}_{z_j=z_k})^{1-r_{jk}}f(\boldsymbol{y}|\pi_p) \\
&= \prod_{1\le j<k\le p}\exp\left(\upsilon\mathbb{1}_{z_j=z_k}\right)\mathrm{pr}(|C_1|,\ldots,|C_M|)\times \\
&\quad \prod_{1\le j<k\le p}(1-\{\exp(\upsilon\zeta_{jk}\mathbb{1}_{z_j=z_k})\}^{-1})^{r_{jk}}\exp(-\upsilon\zeta_{jk}\mathbb{1}_{z_j=z_k})^{1-r_{jk}}f(\boldsymbol{y}|\pi_p) \\
&= \prod_{1\le j<k\le p}\exp\left(\upsilon\mathbb{1}_{z_j=z_k}\right)\mathrm{pr}(|C_1|,\ldots,|C_M|)\times \\
&\quad \prod_{1\le j<k\le p}\left\{\frac{\exp(\upsilon\zeta_{jk}\mathbb{1}_{z_j=z_k})-1}{\exp(\upsilon\zeta_{jk}\mathbb{1}_{z_j=z_k})}\right\}^{r_{jk}}\exp(-\upsilon\zeta_{jk}\mathbb{1}_{z_j=z_k})^{1-r_{jk}}f(\boldsymbol{y}|\pi_p) \\
&= f(\boldsymbol{y}|\pi_p)\mathrm{pr}(|C_1|,\ldots,|C_M|)\times \\
&\quad \prod_{1\le j<k\le p}\left\{\exp\left(\upsilon\mathbb{1}_{z_j=z_k}\right)\right\}^{1-\zeta_{jk}(1-r_{jk})-\zeta_{jk}r_{jk}}\left\{\exp\left(\upsilon\zeta_{jk}\mathbb{1}_{z_j=z_k}\right)-1\right\}^{r_{jk}} \\
&= f(\boldsymbol{y}|\pi_p)\mathrm{pr}(|C_1|,\ldots,|C_M|)\times \\
&\quad \prod_{1\le j<k\le p}\left[\exp\left\{\upsilon(1-\zeta_{jk})\mathbb{1}_{z_j=z_k}\right\}\right]\left\{\exp(\upsilon\zeta_{jk}\mathbb{1}_{z_j=z_k})-1\right\}^{r_{jk}}.
\end{aligned}$$

Thus, the nested cluster A_o is assigned to an existing cluster $m=1,\cdots,M^{-A_o}$ with probability proportional to

$$
\begin{aligned}
&\propto \frac{f(\boldsymbol{y}|\pi_p^{A_o\to m})}{f(\boldsymbol{y}|\pi_p^{-A_o})}\text{pr}\left(|C_1^{-A_o}|,\ldots,|C_m^{-A_o}+A_o|,\ldots,|C_{M^{-A_o}}^{-A_o}|\right)\times\\
&\qquad\prod_{\{(j,k)|j\in A_o,k\in C_m^{-A_o},r_{jk}=0\}}\left[\exp\left\{\upsilon(1-\zeta_{jk})\mathbb{1}_{z_j=z_k}\right\}\right]\left\{\exp(\upsilon\zeta_{jk}\mathbb{1}_{z_j=z_k})-1\right\}^{r_{jk}}\\
&\propto \frac{f(\boldsymbol{y}|\pi_p^{A_o\to m})}{f(\boldsymbol{y}|\pi_p^{-A_o})}\text{pr}\left(|C_1^{-A_o}|,\ldots,|C_m^{-A_o}+A_o|,\ldots,|C_{M^{-A_o}}^{-A_o}|\right)\times\\
&\qquad\prod_{\{(j,k)|j\in A_o,k\in C_m^{-A_o},r_{jk}=0\}}\exp\left\{\upsilon(1-\zeta_{jk})\right\}.
\end{aligned}
$$

Note that the final term represents the predictive probability for the Potts component of the Potts-Gibbs models, as detailed in Table 4.1. Or else, the nested cluster A_o is assigned to a new cluster $m = M^{-A_o}+1$ with probability proportional to

$$\propto \frac{f(\boldsymbol{y}|\pi_p^{A_o\to m})}{f(\boldsymbol{y}|\pi_p^{-A_o})}\text{pr}\left(|C_1^{-A_o}|,\ldots,|C_{M^{-A_o}}^{-A_o}|,|A_o|\right)$$

Marginal likelihood

To sample the new image partition π_p given the latent variables $\boldsymbol{\eta}^*$ and the data, we marginalise out the parameters $\tilde{\boldsymbol{\beta}}$ and σ^2. The marginal likelihood for the continuous case is obtained as follows:

$$
\begin{aligned}
&f\left(\boldsymbol{y}|\pi_p^{A_o\to m},\boldsymbol{\eta}^*\right)\propto\frac{|\Sigma_{\tilde{\beta}}^{-1}|^{\frac{1}{2}}}{|\Sigma_{\tilde{\beta}}^{-1}+\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}|^{\frac{1}{2}}}\left(b_\sigma+S^2/2\right)^{-(a_\sigma+\frac{n}{2})}\\
&\log\left\{f\left(\boldsymbol{y}|\pi_p^{A_o\to m},\boldsymbol{\eta}^*\right)\right\}\propto\frac{1}{2}\log\left(|\Sigma_{\tilde{\beta}}^{-1}|\right)-\frac{1}{2}\log\left(|\Sigma_{\tilde{\beta}}^{-1}+\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}|\right)-\left(a_\sigma+\frac{n}{2}\right)\log\left\{\left(b_\sigma+\frac{S^2}{2}\right)\right\},
\end{aligned}
$$

where

$$
\begin{aligned}
S^2&=(\boldsymbol{y}-\tilde{\boldsymbol{X}}\boldsymbol{m}_{\tilde{\beta}})^T(\boldsymbol{y}-\tilde{\boldsymbol{X}}\boldsymbol{m}_{\tilde{\beta}})-(\boldsymbol{y}-\tilde{\boldsymbol{X}}\boldsymbol{m}_{\tilde{\beta}})^T\tilde{\boldsymbol{X}}(\Sigma_{\tilde{\beta}}^{-1}+\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T(\boldsymbol{y}-\tilde{\boldsymbol{X}}\boldsymbol{m}_{\tilde{\beta}})\\
&=(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{m}_\mu)^T(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{m}_\mu)-(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{m}_\mu)^T\tilde{\boldsymbol{X}}(\Sigma_{\tilde{\beta}}^{-1}+\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T(\boldsymbol{y}-\boldsymbol{W}\boldsymbol{m}_\mu).
\end{aligned}
$$

Note that $b_\sigma + S^2/2 = \widehat{b}_\sigma$. On the other hand, for the binary case, we fix $\sigma^2 = 1$, resulting in a multivariate normal distribution for $f\left(\boldsymbol{y}|\pi_p^{A_o\to m},\boldsymbol{\eta}^*\right)$ instead of a multivariate student-t distribution:

$$f\left(\boldsymbol{y}|\pi_p^{A_o\to m},\tilde{\boldsymbol{\beta}}\right)=\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(y_i-\tilde{\boldsymbol{x}}_i^T\tilde{\boldsymbol{\beta}})^2\right\},$$

with $f\left(\boldsymbol{y}|\pi_p^{A_o\to m}, \boldsymbol{\eta}^*\right)$:

$$
\begin{aligned}
&= \int f\left(\boldsymbol{y}|\pi_p^{A_o\to m}, \tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}^*\right) \mathrm{pr}(\tilde{\boldsymbol{\beta}})\, d\tilde{\boldsymbol{\beta}} \\
&= \int \prod_{i=1}^{n} \mathrm{N}(y_i|\tilde{\boldsymbol{x}}_i^T\tilde{\boldsymbol{\beta}}) N(\tilde{\boldsymbol{\beta}}|\boldsymbol{m}_{\tilde{\beta}}, \Sigma_{\tilde{\beta}})\, d\tilde{\boldsymbol{\beta}} \\
&= \int \frac{1}{\sqrt{2\pi}^{n}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}})^T(\mathbf{y}-\tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}})\right\} \frac{|\Sigma_{\tilde{\beta}}|^{-\frac{1}{2}}}{\sqrt{2\pi}^{M+q}} \exp\left\{-\frac{1}{2}(\tilde{\boldsymbol{\beta}}-\boldsymbol{m}_{\tilde{\beta}})^T\Sigma_{\tilde{\beta}}^{-1}(\tilde{\boldsymbol{\beta}}-\boldsymbol{m}_{\tilde{\beta}})\right\} d\tilde{\boldsymbol{\beta}} \\
&= \frac{|\Sigma_{\tilde{\beta}}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{y}^T\mathbf{y} - \frac{1}{2}\boldsymbol{m}_{\tilde{\beta}}^T\Sigma_{\tilde{\beta}}^{-1}\boldsymbol{m}_{\tilde{\beta}}\right)}{(2\pi)^{\frac{n}{2}+\frac{M+q}{2}}} \int \exp\left[-\frac{1}{2}\left\{\tilde{\boldsymbol{\beta}}^T(\Sigma_{\tilde{\beta}}^{-1}+\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})\tilde{\boldsymbol{\beta}} - 2\tilde{\boldsymbol{\beta}}^T(\Sigma_{\tilde{\beta}}^{-1}\boldsymbol{m}_{\tilde{\beta}}+\tilde{\boldsymbol{X}}^T\mathbf{y})\right\}\right] d\tilde{\boldsymbol{\beta}}.
\end{aligned}
$$

We use the defined terms $\hat{\Sigma}_{\tilde{\beta}} = (\Sigma_{\tilde{\beta}}^{-1} + \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}$ and $\hat{\boldsymbol{m}}_{\tilde{\beta}} = \hat{\Sigma}_{\tilde{\beta}}(\Sigma_{\tilde{\beta}}^{-1}\boldsymbol{m}_{\tilde{\beta}} + \tilde{\boldsymbol{X}}^T\mathbf{y})$ to simplify further the formula, so $f\left(\boldsymbol{y}|\pi_p^{A_o\to m}, \boldsymbol{\eta}^*\right)$:

$$
\begin{aligned}
&= \frac{|\Sigma_{\tilde{\beta}}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{y}^T\mathbf{y} - \frac{1}{2}\boldsymbol{m}_{\tilde{\beta}}^T\Sigma_{\tilde{\beta}}^{-1}\boldsymbol{m}_{\tilde{\beta}}\right)}{(2\pi)^{\frac{n}{2}+\frac{M+q}{2}}} \int \exp\left\{-\frac{1}{2}\left(\tilde{\boldsymbol{\beta}}^T\hat{\Sigma}_{\tilde{\beta}}\tilde{\boldsymbol{\beta}} - 2\tilde{\boldsymbol{\beta}}^T\hat{\Sigma}_{\tilde{\beta}}^{-1}\hat{\boldsymbol{m}}_{\tilde{\beta}}\right)\right\} d\tilde{\boldsymbol{\beta}} \\
&= \frac{|\Sigma_{\tilde{\beta}}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{y}^T\mathbf{y} - \frac{1}{2}\boldsymbol{m}_{\tilde{\beta}}^T\Sigma_{\tilde{\beta}}^{-1}\boldsymbol{m}_{\tilde{\beta}}\right)}{(2\pi)^{\frac{n}{2}+\frac{M+q}{2}}} \int \exp\left\{-\frac{1}{2}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{m}}_{\tilde{\beta}})^T\hat{\Sigma}_{\tilde{\beta}}^{-1}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{m}}_{\tilde{\beta}}) + \frac{1}{2}\hat{\boldsymbol{m}}_{\tilde{\beta}}^T\hat{\Sigma}_{\tilde{\beta}}^{-1}\hat{\boldsymbol{m}}_{\tilde{\beta}}\right\} d\tilde{\boldsymbol{\beta}} \\
&= \frac{|\Sigma_{\tilde{\beta}}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{y} - \frac{1}{2}\boldsymbol{m}_{\tilde{\beta}}^T\Sigma_{\tilde{\beta}}^{-1}\boldsymbol{m}_{\tilde{\beta}}\right)}{(2\pi)^{\frac{n}{2}+\frac{M+q}{2}}} \exp\left(\frac{1}{2}\hat{\boldsymbol{m}}_{\tilde{\beta}}^T\hat{\Sigma}_{\tilde{\beta}}^{-1}\hat{\boldsymbol{m}}_{\tilde{\beta}}\right) \int \exp\left\{-\frac{1}{2}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{m}}_{\tilde{\beta}})^T\hat{\Sigma}_{\tilde{\beta}}^{-1}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{m}}_{\tilde{\beta}})\right\} d\tilde{\boldsymbol{\beta}} \\
&= \frac{|\Sigma_{\tilde{\beta}}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{y}^T\mathbf{y} - \frac{1}{2}\boldsymbol{m}_{\tilde{\beta}}^T\Sigma_{\tilde{\beta}}^{-1}\boldsymbol{m}_{\tilde{\beta}}\right)}{(2\pi)^{\frac{n}{2}+\frac{M+q}{2}}} \exp\left(\frac{1}{2}\hat{\boldsymbol{m}}_{\tilde{\beta}}^T\hat{\Sigma}_{\tilde{\beta}}^{-1}\hat{\boldsymbol{m}}_{\tilde{\beta}}\right)(2\pi)^{\frac{M+q}{2}}|\hat{\Sigma}_{\tilde{\beta}}|^{\frac{1}{2}} \\
&= \frac{|\Sigma_{\tilde{\beta}}|^{-\frac{1}{2}}|\hat{\Sigma}_{\tilde{\beta}}|^{-\frac{1}{2}}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\mathbf{y}^T\mathbf{y} - \frac{1}{2}\boldsymbol{m}_{\tilde{\beta}}^T\Sigma_{\tilde{\beta}}^{-1}\boldsymbol{m}_{\tilde{\beta}} + \frac{1}{2}\hat{\boldsymbol{m}}_{\tilde{\beta}}^T\hat{\Sigma}_{\tilde{\beta}}^{-1}\hat{\boldsymbol{m}}_{\tilde{\beta}}\right) \\
&\propto |\Sigma_{\tilde{\beta}}|^{-\frac{1}{2}}|\hat{\Sigma}_{\tilde{\beta}}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{y}^T\mathbf{y} - \frac{1}{2}\boldsymbol{m}_{\tilde{\beta}}^T\Sigma_{\tilde{\beta}}^{-1}\boldsymbol{m}_{\tilde{\beta}} + \frac{1}{2}\hat{\boldsymbol{m}}_{\tilde{\beta}}^T\hat{\Sigma}_{\tilde{\beta}}^{-1}\hat{\boldsymbol{m}}_{\tilde{\beta}}\right).
\end{aligned}
$$

Thus, the log marginal likelihood for the binary case is:

$$
\log\left\{f\left(\boldsymbol{y}|\pi_p^{A_o\to m}, \boldsymbol{\eta}^*\right)\right\} \propto -\frac{1}{2}\log(|\Sigma_{\tilde{\beta}}|) + \frac{1}{2}\log(|\hat{\Sigma}_{\tilde{\beta}}|) - \frac{1}{2}\mathbf{y}^T\mathbf{y} - \frac{1}{2}\boldsymbol{m}_{\tilde{\beta}}^T\Sigma_{\tilde{\beta}}^{-1}\boldsymbol{m}_{\tilde{\beta}} + \frac{1}{2}\hat{\boldsymbol{m}}_{\tilde{\beta}}^T\hat{\Sigma}_{\tilde{\beta}}^{-1}\hat{\boldsymbol{m}}_{\tilde{\beta}}.
$$

Note that the term that we can cancel when we normalise the unnormalised log marginal likelihood is $\mathbf{y}^T\mathbf{y}$.

7.2 The computation of the coefficients $V_p(M)$ for the Potts-MFM model

The computation of the coefficients $V_p(M)$ for the Potts-MFM model is described in this section. The coefficients can be approximated as:

$$V_p(M) \sim \frac{M_{(M)}}{(\gamma M)^{(p)}} p_L(M|\psi) \sim \frac{M!}{p!} \frac{\Gamma(\gamma M)}{p^{\gamma M - 1}} p_L(M|\psi),$$

when $p \to \infty$ (see Section 3.3.6).

To compute the conditional distribution $\mathrm{pr}(\pi_p^{A_o \to m} | \pi_p^{-A_o}, \Theta)$ for the Potts-MFM model, we need to evaluate the ratio of the coefficients $V_p(M^{-A_o}+1)$ and $V_p(M^{-A_o})$:

$$\begin{aligned}
\frac{V_p(M^{-A_o}+1)}{V_p(M^{-A_o})} &= \frac{\frac{(M^{-A_o}+1)!}{p!} \frac{\Gamma(\gamma(M^{-A_o}+1))}{p^{\gamma(M^{-A_o}+1)-1}} p_L(M^{-A_o}+1)}{\frac{M^{-A_o}!}{p!} \frac{\Gamma(\gamma M^{-A_o})}{p^{\gamma M^{-A_o}-1}} p_L(M^{-A_o})} \\
&= \frac{(M^{-A_o}+1)!}{M^{-A_o}!} \frac{\Gamma(\gamma(M^{-A_o}+1))}{\Gamma(\gamma M^{-A_o})} \frac{p^{\gamma M^{-A_o}-1}}{p^{\gamma(M^{-A_o}+1)-1}} \frac{p_L(M^{-A_o}+1)}{p_L(M^{-A_o})} \\
&= (M^{-A_o}+1) \frac{\Gamma(\gamma(M^{-A_o}+1))}{\Gamma(\gamma M^{-A_o})} \frac{1}{p^\gamma} \frac{p_L(M^{-A_o}+1)}{p_L(M^{-A_o})}.
\end{aligned}$$

If we choose the zero-truncated Poisson (ZTP) distribution for $p_L(l)$:

$$p_L(l) = \frac{\lambda^l}{(\exp \lambda - 1) l!},$$

where λ is the parameter, then we have:

$$\begin{aligned}
\frac{p_L(M^{-A_o}+1)}{p_L(M^{-A_o})} &= \frac{\frac{\lambda^{(M^{-A_o}+1)}}{(\exp\lambda - 1)(M^{-A_o}+1)!}}{\frac{\lambda^{M^{-A_o}}}{(\exp\lambda - 1) M^{-A_o}!}} \\
&= \frac{\lambda}{M^{-A_o}+1}.
\end{aligned}$$

Substituting the ratio of the coefficients $p_L(M^{-A_o}+1)$ and $p_L(M^{-A_o})$ into the original expression, we obtain:

$$\begin{aligned}
\frac{V_p(M^{-A_o}+1)}{V_p(M^{-A_o})} &= (M^{-A_o}+1) \frac{\Gamma(\gamma(M^{-A_o}+1))}{\Gamma(\gamma M^{-A_o})} \frac{1}{p^\gamma} \frac{\lambda}{M^{-A_o}+1} \\
&= \frac{\lambda}{p^\gamma} \frac{\Gamma(\gamma(M^{-A_o}+1))}{\Gamma(\gamma M^{-A_o})}.
\end{aligned}$$

7.3 Algorithm pseudocode

Here we list the pseudocode to describe how different algorithms and procedures work for our implementation of the proposed models: EPA SIR and Potts-Gibbs SIR models, which are the EPA sampling algorithm (Algorithm 1), GSW sampling algorithm (Algorithm 2) and the full algorithm (Algorithm 3).

Algorithm 1: EPA Sampling Algorithm

```
Input: $\alpha, \delta, \tau, h$
Output: A MCMC chain of simulated values of $\pi_p$
for $t \leftarrow 1$ to $T$ do
    for $j \leftarrow 1$ to $p$ do
        Remove $j$ from $C_{z_j}$.
        if $z_j \neq z_k \forall k \neq j$ then
            Relabel $z_j = M_t^{-j} + 1$.
            Set $\beta^*_{M_t^{-j}+1} = \beta^*_{z_j}$.
            Sample $\eta_{M_t^{-j}+2}, \ldots, \eta_{M_t^{-j}+h}$ independently from the prior distribution (Equation (4.5)).
            Sample $\beta^*_{M_t^{-j}+2}, \ldots, \beta^*_{M_t^{-j}+h}$ independently from its prior distribution (Equation (4.5)).
        else
            Sample $\eta_{M_t^{-j}+1}, \ldots, \eta_{M_t^{-j}+h}$ independently from the prior distribution (Equation (4.5)).
            Sample $\beta^*_{M_t^{-j}+1}, \ldots, \beta^*_{M_t^{-j}+h}$ independently from its prior distribution (Equation (4.5)).
        end
        Draw a sample of which cluster the unit $j$ should belong to, $C_m^{-j}$ :
        for $m \leftarrow 1$ to $M_t^{-j} + h$ do
            Calculate the log probability of $z_j = m$ as described in Section 4.3 using Equation (3.1)–(3.2) (with the second condition in the corresponding probabilities divided by $h$) at the partition $\pi_p^{A_o \rightarrow m}$.
        end
        Transform unnormalised log probabilities into probabilities.
        if $m \leq M_t^{-j}$ then
            Update $z_j$ according to $z_j = m$.
        else
            A new cluster has been created.
            Append this new cluster to the vector of non-empty clusters.
        end
    end
    A new partition is created.
end
```

Algorithm 2: GSW Sampling Algorithm

```
Input: $\upsilon, \phi, \kappa, \tau, h$
Output: A MCMC chain of simulated values of $\pi_p$
for $t \leftarrow 1$ to $T$ do
    Create nested clusters $A_1, \ldots, A_{O_t}$:
    for $m \leftarrow 1$ to $M_t$ do
        for $j \leftarrow 1$ to $|C_m|$ do
            for $k$ in $C_m$ and $j \sim k$ and $j < k$ do
                $$r_{jk} \sim \text{Ber}\left(1 - \exp(-\upsilon\zeta_{jk}\mathbb{1}_{z_j=z_k})\right).$$
            end
        end
    end
    for $o \leftarrow 1$ to $O_t$ do
        Remove $A_o$ from $\pi_p$.
        if $z_{A_o} \neq z_{A_l} \forall l \neq o$ then
            Relabel $z_{A_o} = M_t^{-A_o} + 1$.
            Set $\beta^*_{M_t^{-A_o}+1} = \beta^*_{z_{A_o}}$.
            Sample $\eta^*_{M_t^{-A_o}+2}, \ldots, \eta^*_{M_t^{-A_o}+h}$ independently from the prior distribution
              (Equation (4.5)).
            Sample $\beta^*_{M_t^{-A_o}+2}, \ldots, \beta^*_{M_t^{-A_o}+h}$ independently from its prior distribution
              (Equation (4.5)).
        else
            Sample $\eta^*_{M_t^{-A_o}+1}, \ldots, \eta^*_{M_t^{-A_o}+h}$ independently from the prior distribution
              (Equation (4.5)).
            Sample $\beta^*_{M_t^{-A_o}+1}, \ldots, \beta^*_{M_t^{-A_o}+h}$ independently from its prior distribution
              (Equation (4.5)).
        end
        Draw a sample of which cluster this $A_o$ should belong to, $C_m^{-A_o}$:
        for $m \leftarrow 1$ to $M_t^{-A_o} + h$ do
            Calculate the log probability using $\text{pr}(A_o \in C_m^{-A_o} | \ldots)$ from Table 4.1.
        end
        Transform unnormalised log probabilities into probabilities.
        if $m \leq M_t^{-A_o}$ then
            Update $z_{A_o}$ according to $z_{A_o} = m$.
        else
            A new cluster has been created.
            Append this new cluster to the vector of non-empty clusters.
        end
    end
    A new partition is created.
end
```

Algorithm 3: Full Algorithm

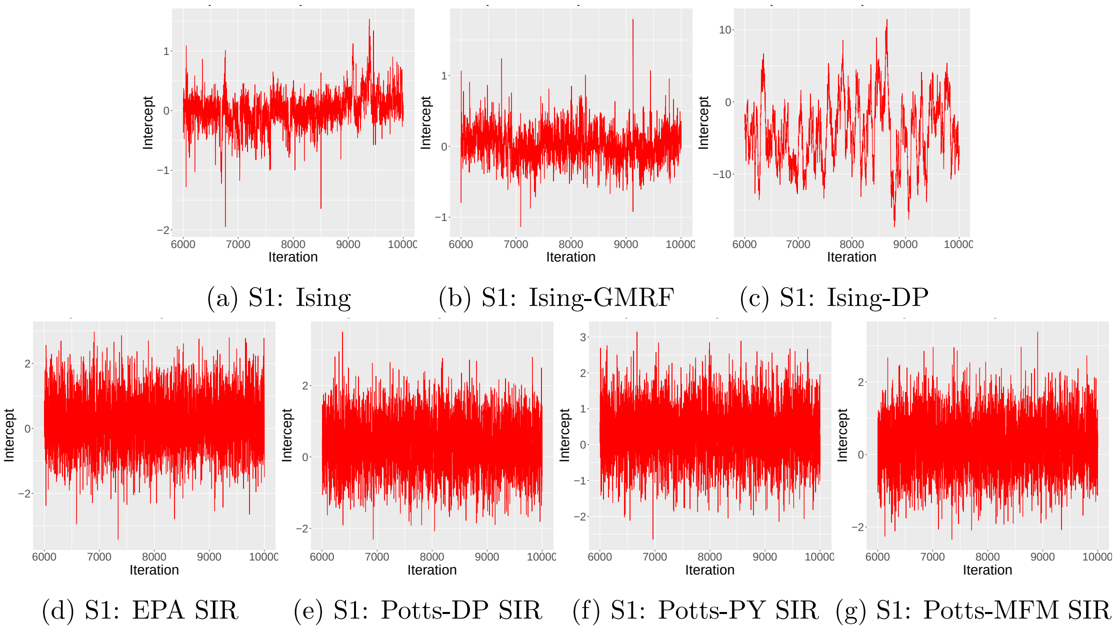
```
Input: $(a_\sigma, b_\sigma), (a_\eta, a_o, b_o), (\boldsymbol{m}_\mu, \boldsymbol{c}_\mu), \Theta$
Output: A MCMC chain of simulated values of $\pi_p, \sigma^2, \tilde{\boldsymbol{\beta}}, \boldsymbol{\eta}$
Initialisation: $\pi_p^0$, $\left(\sigma^2\right)^0, \boldsymbol{\eta}^0, \tilde{\boldsymbol{\beta}}^0$
for $t \leftarrow 1$ to $T$ do
    Image partition, $\pi_p$: First enters either EPA sampling (Algorithm 1) or GSW sampling (Algorithm 2).
    Noise variance, $\sigma^2$: Generate a new value for $\sigma^2$ from the marginal posterior distribution:
        $$\left(\sigma^2\right)^t \sim \mathrm{IG}\left(\hat{a}_\sigma, \hat{b}_\sigma^t\right).$$
    Coefficients, $\tilde{\boldsymbol{\beta}}$: Conditional on the newly updated parameter value $\sigma^2$, generate a new value for $\tilde{\boldsymbol{\beta}}$ from the posterior conditional distribution:
        $$\tilde{\boldsymbol{\beta}}^t | \left(\sigma^2\right)^t, \cdots \sim \mathrm{N}\left(\hat{\boldsymbol{m}}_{\tilde{\beta}}^t, (\sigma^2)^t \hat{\Sigma}_{\tilde{\beta}}^t\right).$$
    Local shrinkage parameters, $\boldsymbol{\eta}^*$: Conditional on the newly updated parameter values $\tilde{\boldsymbol{\beta}}$ and $\sigma^2$, generate a new value for $\boldsymbol{\eta}^*$ from the posterior conditional distribution:
        $$\eta_m^{*t} | \tilde{\boldsymbol{\beta}}^t, \left(\sigma^2\right)^t, \cdots \sim \mathrm{IG}(\hat{a}_\eta, \hat{b}_\eta^t), \quad \text{for all } m = 1, \ldots, M_t.$$
    Global shrinkage parameter, $b_\eta$: Conditional on the newly updated parameter value $\boldsymbol{\eta}^*$, generate a new value for $b_\eta$ from the marginal posterior distribution:
        $$b_\eta^t | \boldsymbol{\eta}^*, \cdots \sim \mathrm{G}\left(\hat{a}_o, \hat{b}_o^t\right).$$
end
```

7.4 Experiments

Traceplot, posterior inclusion probability (PIP) and posterior predictive check (PPC) of the experiments are provided here.

7.4.1 Traceplot

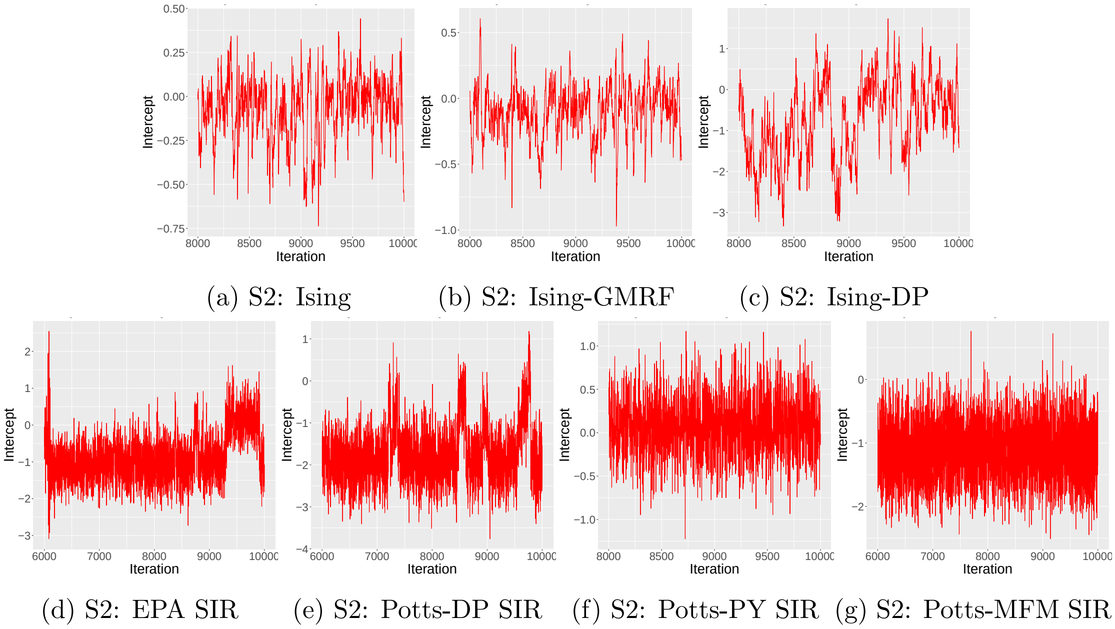
The following are the traceplots of the intercept from the posterior distribution of both the simulated and real data sets. As can be observed, the Ising-DP model has a relatively poor mixing rate in most of the scenarios compared to the other models. For the Ising-DP model, we have also tried to run for longer iterations (100k) for the simulated data set for Scenario 3 and Scenario 4. However, the traceplots of the intercept for both scenarios as displayed in Figure 7.5 still do not show any sign of convergence, due to the computation budget, we do not proceed to run it further.



(a) S1: Ising (b) S1: Ising-GMRF (c) S1: Ising-DP

(d) S1: EPA SIR (e) S1: Potts-DP SIR (f) S1: Potts-PY SIR (g) S1: Potts-MFM SIR

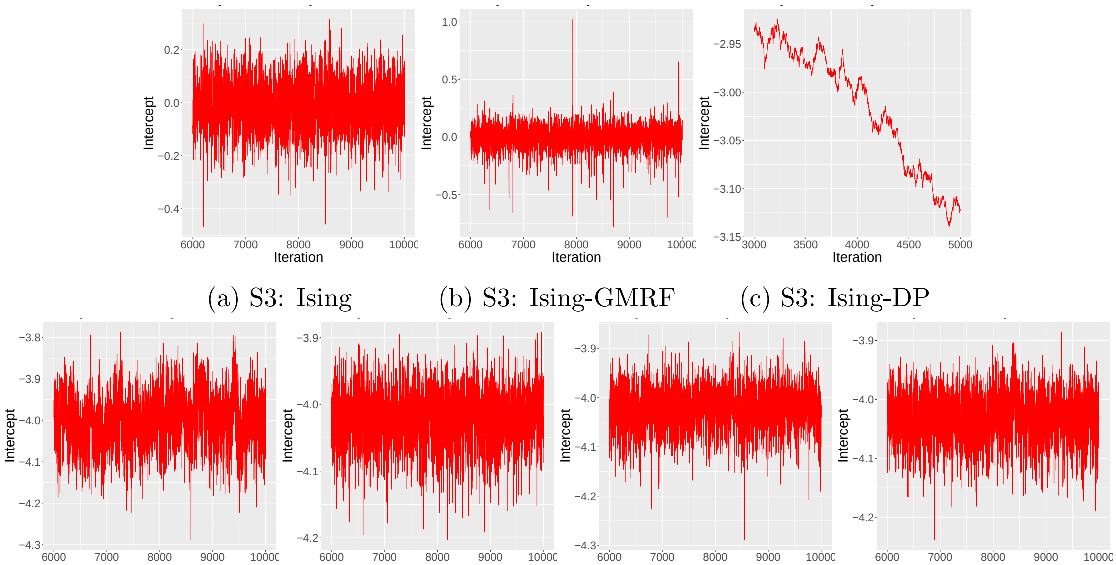
Figure 7.1: Figures showing the traceplots of the intercept of the simulated data sets for **Scenario 1** under each model.



(a) S2: Ising (b) S2: Ising-GMRF (c) S2: Ising-DP

(d) S2: EPA SIR (e) S2: Potts-DP SIR (f) S2: Potts-PY SIR (g) S2: Potts-MFM SIR

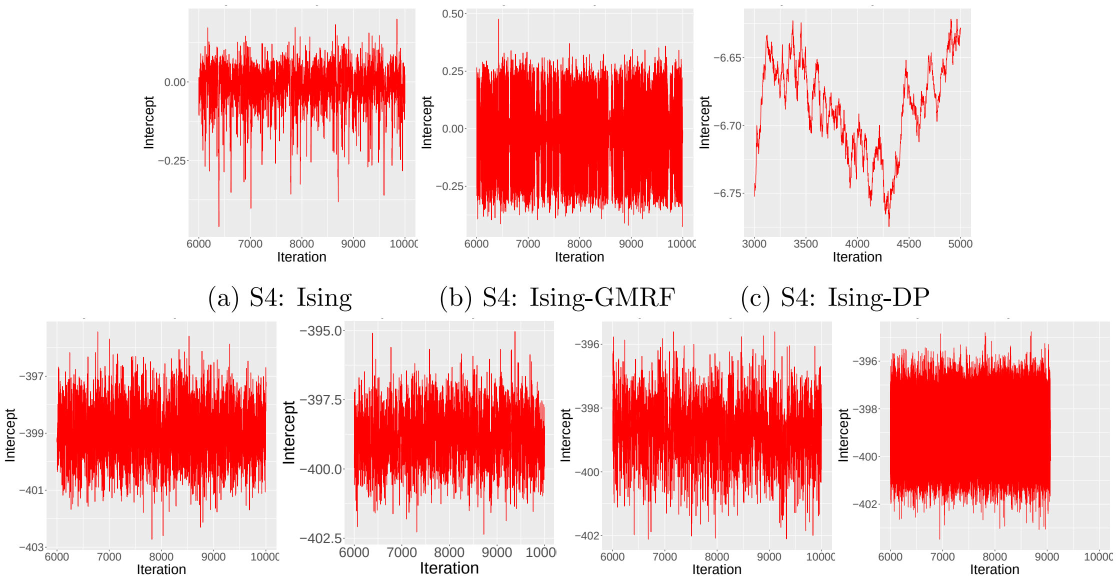
Figure 7.2: Figures showing the traceplots of the intercept of the simulated data sets for **Scenario 2** under each model.



(a) S3: Ising (b) S3: Ising-GMRF (c) S3: Ising-DP

(d) S3: EPA SIR (e) S3: Potts-DP SIR (f) S3: Potts-PY SIR (g) S3: Potts-MFM SIR

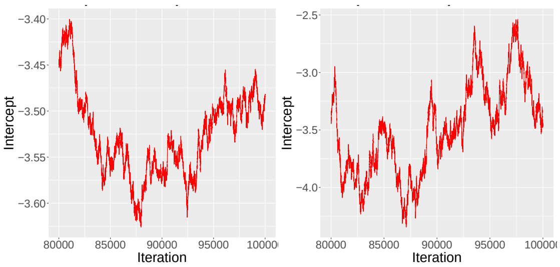
Figure 7.3: Figures showing the traceplots of the intercept of the simulated data sets for **Scenario 3** under each model.



(a) S4: Ising (b) S4: Ising-GMRF (c) S4: Ising-DP

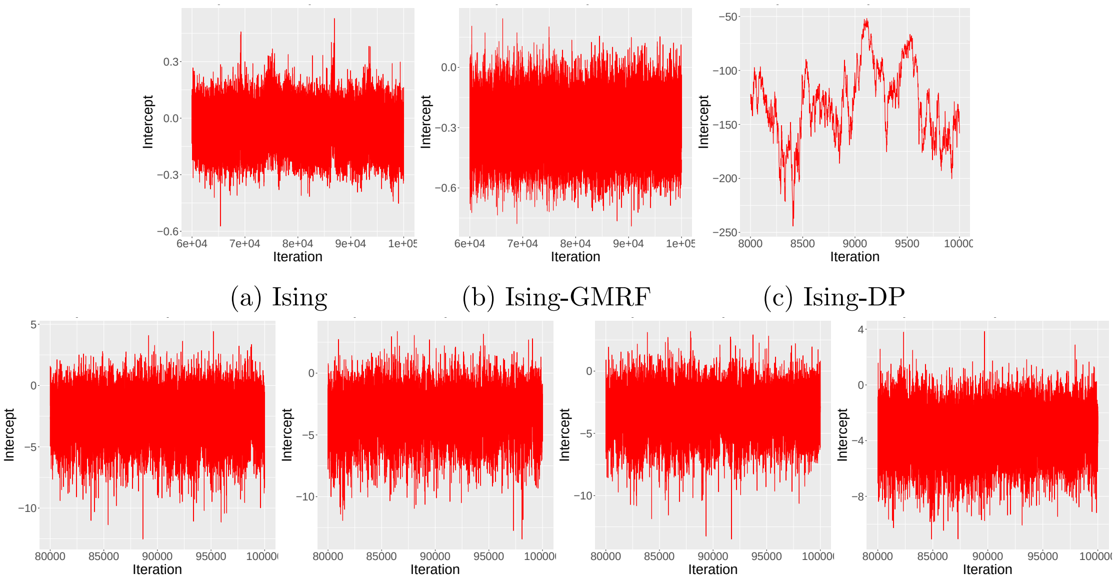
(d) S4: EPA SIR (e) S4: Potts-DP SIR (f) S4: Potts-PY SIR (g) S4: Potts-MFM SIR

Figure 7.4: Figures showing the traceplots of the intercept of the simulated data sets for **Scenario 4** under each model.



(a) S3: Ising-DP (b) S4: Ising-DP

Figure 7.5: Figures showing the traceplots of the intercept of the simulated data sets for **Scenario 3** and **Scenario 4** for the IsingDP with longer iteration (100k).



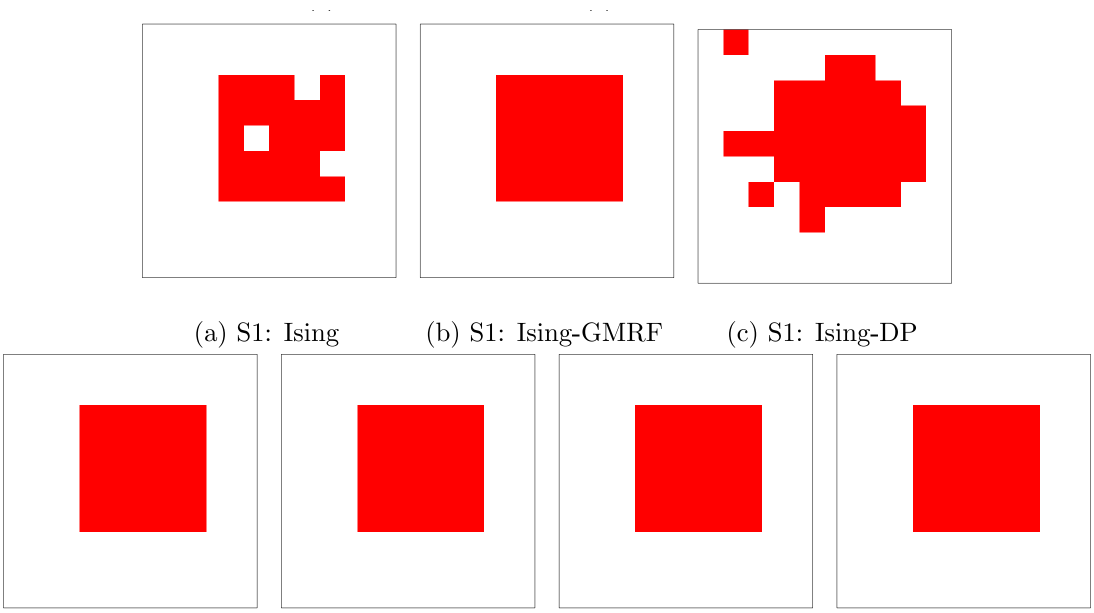
(a) Ising (b) Ising-GMRF (c) Ising-DP

(d) EPA SIR (e) Potts-DP SIR (f) Potts-PY SIR (g) Potts-MFM SIR

Figure 7.6: Figures showing the traceplots of the intercept of the real data sets for **ordinal response** under each model.

7.4.2 Binary posterior inclusion map

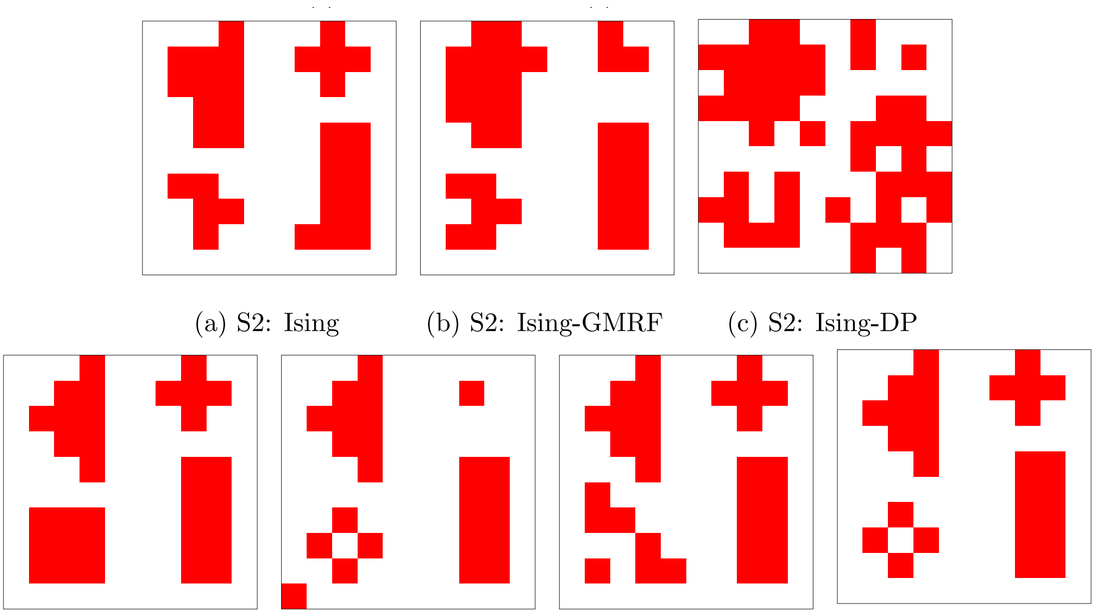
The binary posterior inclusion map as explained in Section 4.4 of both the simulated and real data sets is shown below.



(a) S1: Ising (b) S1: Ising-GMRF (c) S1: Ising-DP

(d) S1: EPA SIR (e) S1: Potts-DP SIR (f) S1: Potts-PY SIR (g) S1: Potts-MFM SIR

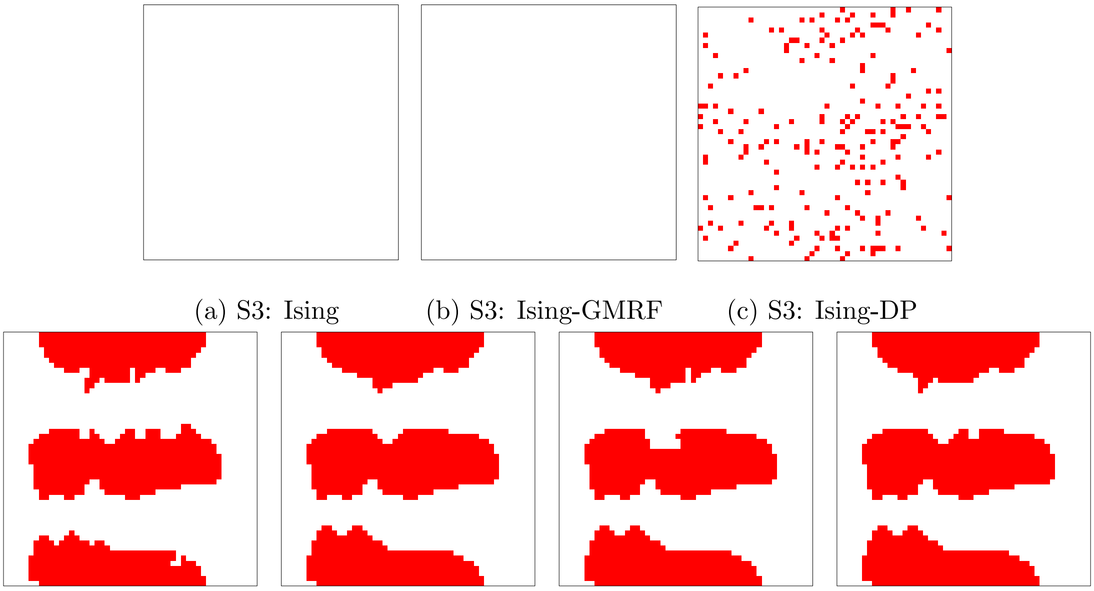
Figure 7.7: Figures showing the binary posterior inclusion maps of the simulated data sets for **Scenario 1** under each model.



(a) S2: Ising (b) S2: Ising-GMRF (c) S2: Ising-DP

(d) S2: EPA SIR (e) S2: Potts-DP SIR (f) S2: Potts-PY SIR (g) S2: Potts-MFM SIR

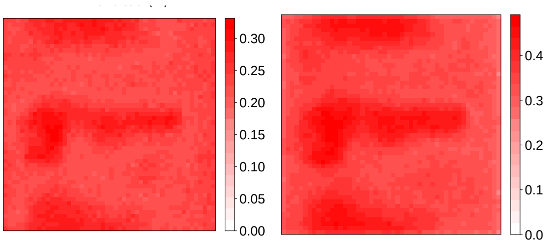
Figure 7.8: Figures showing the binary posterior inclusion maps of the simulated data sets for **Scenario 2** under each model.



(a) S3: Ising (b) S3: Ising-GMRF (c) S3: Ising-DP

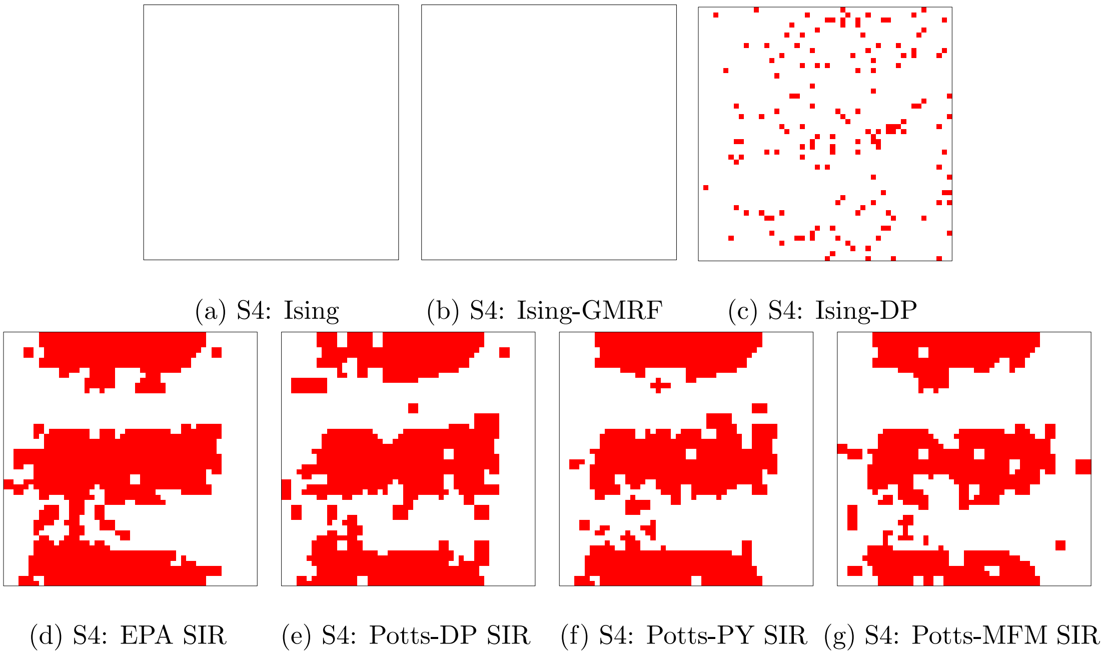
(d) S3: EPA SIR (e) S3: Potts-DP SIR (f) S3: Potts-PY SIR (g) S3: Potts-MFM SIR

Figure 7.9: Figures showing the binary posterior inclusion maps of the simulated data sets for **Scenario 3** under each model.



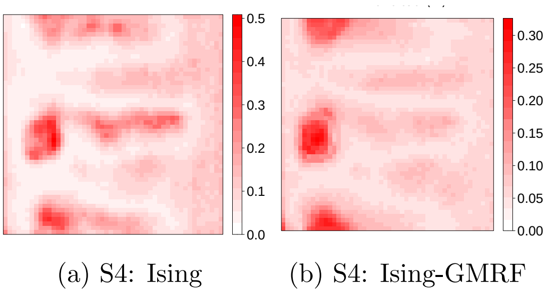
(a) S3: Ising (b) S3: Ising-GMRF

Figure 7.10: Figures showing the posterior inclusion probability (PIP) plots of the simulated data sets for **Scenario 3** under the Ising and Ising-GMRF models as the binary posterior inclusion maps do not show much information for both models.


(a) S4: Ising (b) S4: Ising-GMRF (c) S4: Ising-DP

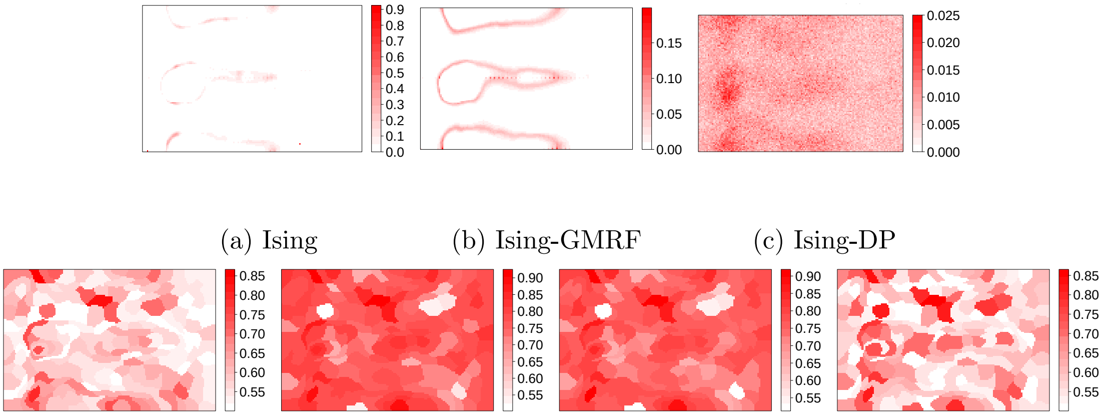
(d) S4: EPA SIR (e) S4: Potts-DP SIR (f) S4: Potts-PY SIR (g) S4: Potts-MFM SIR

Figure 7.11: Figures showing the binary posterior inclusion maps of the simulated data sets for **Scenario 4** under each model.



(a) S4: Ising (b) S4: Ising-GMRF

Figure 7.12: Figures showing the posterior inclusion probability (PIP) plots of the simulated data sets for **Scenario 4** under the Ising and Ising-GMRF models as the binary posterior inclusion maps do not show much information for both models.



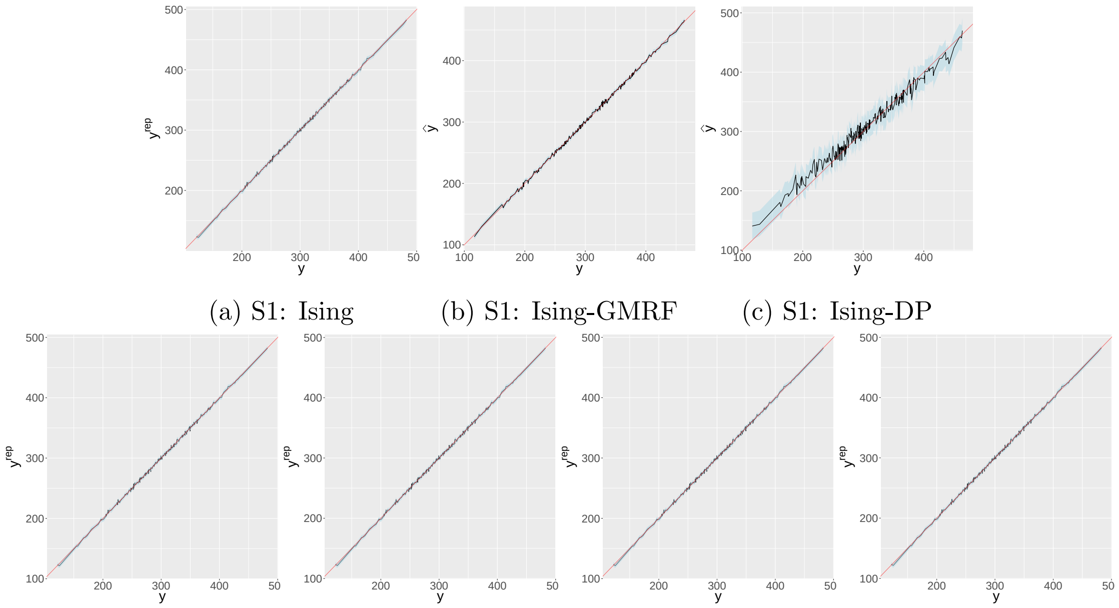
(a) Ising (b) Ising-GMRF (c) Ising-DP

(d) EPA SIR (e) Potts-DP SIR (f) Potts-PY SIR (g) Potts-MFM SIR

Figure 7.13: Figures showing the posterior inclusion probability (PIP) plots of the real data sets for **ordinal response** under each model.

7.4.3 Posterior predictive checking

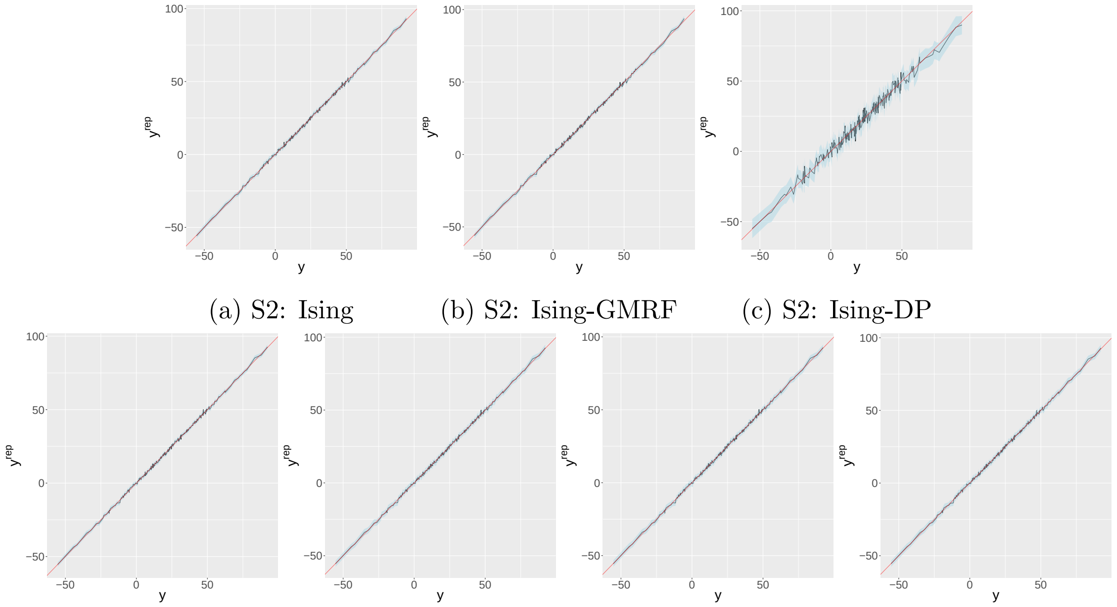
The posterior predictive check (PPC) plots as explained in Section 4.4 of both the simulated and real data sets are shown below.



(a) S1: Ising (b) S1: Ising-GMRF (c) S1: Ising-DP

(d) S1: EPA SIR (e) S1: Potts-DP SIR (f) S1: Potts-PY SIR (g) S1: Potts-MFM SIR

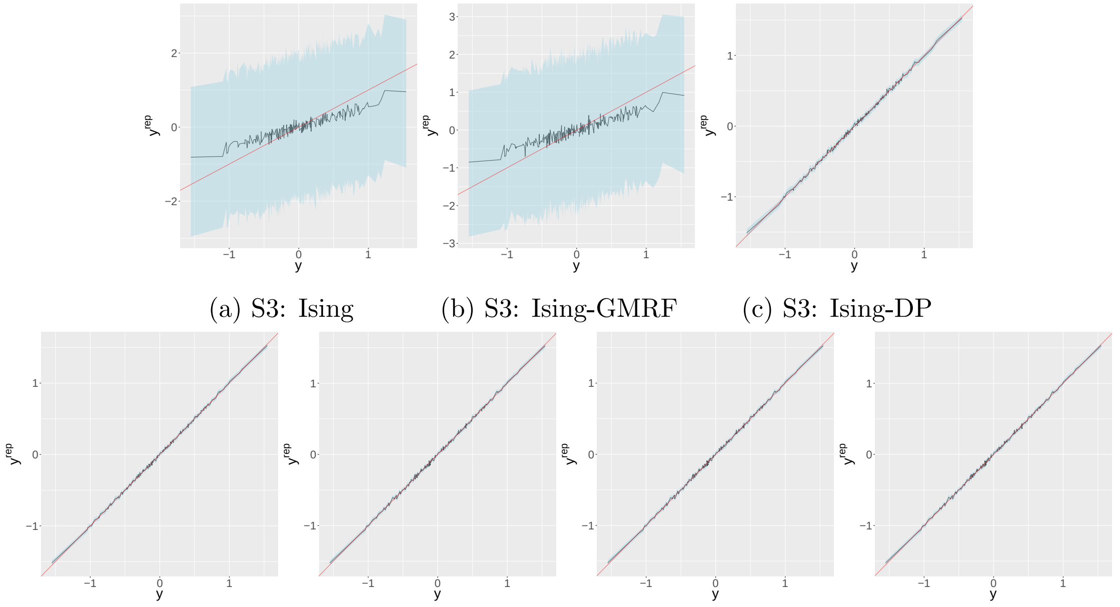
Figure 7.14: Figures showing the posterior predictive check (PPC) plots of the simulated data sets for **Scenario 1** under each model.



(a) S2: Ising (b) S2: Ising-GMRF (c) S2: Ising-DP

(d) S2: EPA SIR (e) S2: Potts-DP SIR (f) S2: Potts-PY SIR (g) S2: Potts-MFM SIR

Figure 7.15: Figures showing the posterior predictive check (PPC) plots of the simulated data sets for **Scenario 2** under each model.



(a) S3: Ising (b) S3: Ising-GMRF (c) S3: Ising-DP

(d) S3: EPA SIR (e) S3: Potts-DP SIR (f) S3: Potts-PY SIR (g) S3: Potts-MFM SIR

Figure 7.16: Figures showing the posterior predictive check (PPC) plots of the simulated data sets for **Scenario 3** under each model.

Bibliography

Barbu Adrian and Song-Chun Zhu. Generalizing Swendsen-Wang to sampling arbitrary posterior probabilities. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(8):1239–1253, 2005. ISSN 0162-8828.

James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *J. Am. Stat. Assoc.*, 88(422):669–679, 1993.

Nicola Amoroso, Domenico Diacono, Annarita Fanizzi, Marianna La Rocca, Alfonso Monaco, Angela Lombardi, Cataldo Guaragnella, Roberto Bellotti, Sabina Tangaro, Alzheimer's Disease Neuroimaging Initiative, et al. Deep learning reveals Alzheimer's disease onset in MCI subjects: Results from an international challenge. *J. Neurosci. Methods*, 302:3–9, 2018.

Liana G Apostolova, Ivo D Dinov, Rebecca A Dutton, Kiralee M Hayashi, Arthur W Toga, Jeffrey L Cummings, and Paul M Thompson. 3D comparison of hippocampal atrophy in amnestic mild cognitive impairment and Alzheimer's disease. *Brain*, 129(11):2867–2873, 2006.

Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):898–916, 2010.

John Ashburner and Karl J Friston. Voxel-based morphometry—the methods. *NeuroImage*, 11 (6):805–821, 2000.

Cecilia Balocchi and Shane T Jensen. Spatial modeling of trends in crime over time in Philadelphia. *Ann. Appl. Stat.*, 13(4):2235–2259, 2019.

Adrian Barbu and Song-Chun Zhu. Generalizing Swendsen-Wang for image analysis. *J. Comput. Graph. Stat.*, 16(4):877–900, 2007. ISSN 1061-8600.

Daniel Barry and John A Hartigan. A Bayesian analysis for change point problems. *J. Am. Stat. Assoc.*, 88(421):309–319, 1993. ISSN 0162-1459. URL `http://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10594323`.

José Miguel Bernardo, María Jesús Bayarri, James Orvis Berger, Alexander Philip Dawid, David Heckerman, Adrian FM Smith, and Mike West. Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Stat.*, 7:733–742, 2003.

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc., B: Stat. Methodol.*, 36(2):192–236, 1974. ISSN 0035-9246.

Julian Besag. On the statistical analysis of dirty pictures. *J. R. Stat. Soc., B: Stat. Methodol.*, 48(3):259–279, 1986.

David M Blei and Peter I Frazier. Distance dependent Chinese restaurant processes. *J. Mach. Learning Res.*, 12:2461–2488, 2011.

Heiko Braak and Eva Braak. Neuropathological stageing of Alzheimer-related changes. *Acta. Neuropathol.*, 82(4):239–259, 1991. ISSN 0001-6322.

Wray Buntine and Marcus Hutter. A Bayesian view of the Poisson-Dirichlet process. *arXiv preprint arXiv:1007.0296*, 2010.

Federico Camerlenghi, Antonio Lijoi, and Igor Prünster. Bayesian prediction with multiple-samples information. *J. Multivar. Anal.*, 156:18–28, 2017.

Ismaël Castillo, Johannes Schmidt-Hieber, and Aad van Der Vaart. Bayesian linear regression with sparse priors. *Ann. Stat.*, 43(5):1986, 2015. ISSN 00905364. URL `http://search.proquest.com/docview/1787036022/`.

Gilles Celeux, Mohammed El Anbari, Jean-Michel Marin, and Christian P Robert. Regularization in regression: Comparing Bayesian and frequentist methods in a poorly informative situation. *Bayesian Anal.*, 7:477–502, 2012.

Annalisa Cerquetti. Generalized Chinese restaurant construction of exchangeable Gibbs partitions and related results. *arXiv preprint arXiv:0805.3853*, 2008.

Sotirios P Chatzis. A Markov random field-regulated Pitman–Yor process prior for spatially constrained data clustering. *Pattern Recognit.*, 46(6):1595–1603, 2013.

Sotirios P Chatzis and Gabriel Tsechpenakis. The infinite hidden Markov random field model. *IEEE trans. neural netw.*, 21(6):1004–1014, 2010.

Stephen Coleman, Paul DW Kirk, and Chris Wallace. Consensus clustering for Bayesian mixture models. *bioRxiv*, pages 2020–12, 2021.

Antonio Convit, Josepheen De Asis-Cruz, Mony J De Leon, Chaim Y Tarshish, Susan Desanti, and Henry Rusinek. Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to Alzheimer's disease. *Neurobiol. Aging*, 21: 19–26, 2000.

R Cameron Craddock, Paul E Holtzheimer III, Xiaoping P Hu, and Helen S Mayberg. Disease state prediction from resting state functional connectivity. *Magn. Reson. Med.*, 62(6):1619–1628, 2009. ISSN 0740-3194.

Richard Yi Da Xu, Francois Caron, and Arnaud Doucet. Bayesian nonparametric image segmentation using a generalized Swendsen-Wang algorithm. *arXiv preprint arXiv:1602.03048*, 2016.

David B Dahl, Ryan Day, and Jerry W Tsai. Random partition distribution indexed by pairwise information. *J. Am. Stat. Assoc.*, 112:721–732, 2017.

Christos Davatzikos, Dinggang Shen, Ruben C Gur, Xiaoying Wu, Dengfeng Liu, Yong Fan, Paul Hughett, Bruce I Turetsky, and Raquel E Gur. Whole-brain morphometric study of schizophrenia revealing a spatially complex set of focal abnormalities. *Arch. Gen. Psychiatry*, 62(11):1218–1227, 2005. ISSN 0003-990X.

Delphine Debois, Marc Ongena, Hélène Cawoy, and Edwin De Pauw. MALDI-FTICR MS imaging as a powerful tool to identify paenibacillus antibiotics involved in the inhibition of plant pathogens. *J. Am. Soc. Mass Spectrom.*, 24(8):1202–1213, 2013. ISSN 1044-0305.

David GT Denison and Christofer C Holmes. Bayesian partitioning for estimating disease risk. *Biometrics*, 57(1):143–149, 2001.

Bruno Dubois, Howard H Feldman, Claudia Jacova, Steven T DeKosky, Pascale Barberger-Gateau, Jeffrey Cummings, André Delacourte, Douglas Galasko, Serge Gauthier, Gregory Jicha, et al. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. *Lancet Neurol.*, 6:734–746, 2007.

David B Dunson, Natesh Pillai, and Ju-Hyun Park. Bayesian density regression. *J. R. Stat. Soc., B: Stat. Methodol.*, 69(2):163–183, 2007.

Jean-Baptiste Durand, Florence Forbes, Cong Duc Phan, Long Truong, Hien Nguyen, and Fatoumata Dama. Bayesian nonparametric spatial prior for traffic crash risk mapping: A case study of Victoria, Australia. 2021.

Daniele Durante. Conjugate bayes for probit regression via unified skew-normal distributions. *Biometrika*, 106(4):765–779, 2019. ISSN 0006-3444.

Yong Fan, Susan M Resnick, Xiaoying Wu, and Christos Davatzikos. Structural and functional biomarkers of prodromal Alzheimer's disease: A high-dimensional pattern classification study. *NeuroImage (Orlando, Fla.)*, 41(2):277–285, 2008. ISSN 1053-8119.

Xiangnan Feng, Tengfei Li, Xinyuan Song, and Hongtu Zhu. Bayesian scalar on image regression with nonignorable nonresponse. *J. Am. Stat. Assoc.*, 115(532):1574–1597, 2020. ISSN 0162-1459.

Thomas S Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1 (2):209–230, 03 1973. doi: 10.1214/aos/1176342360. URL `https://doi.org/10.1214/aos/1176342360`.

Thomas S Ferguson. Prior distributions on spaces of probability measures. *Ann. Statist.*, 2(4): 615–629, 07 1974. doi: 10.1214/aos/1176342752. URL `https://doi.org/10.1214/aos/1176342752`.

Bruce Ferwerda, Markus Schedl, and Marko Tkalcic. Using Instagram picture features to predict users' personality. In *International Conference on Multimedia Modeling*. Springer, 2016.

Chris Fraley and Adrian E Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. *J. Classif.*, 24(2):155–181, 2007.

Alan E Gelfand, Peter Diggle, Peter Guttorp, and Montserrat Fuentes. *Handbook of Spatial Statistics*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Taylor & Francis, 2010. ISBN 9781420072877. URL `http://books.google.com/books?id=EFbbcMFZ2mMC`.

Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for Bayesian models. *Stat. Comput.*, 24(6):997–1016, 2014.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI-6(6):721–741, 1984. ISSN 0162-8828.

Edward I George and Robert E McCulloch. Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.*, 88(423):881–889, 1993.

Samuel J Gershman and David M Blei. A tutorial on Bayesian nonparametric models. *J. Math. Psychol.*, 56:1–12, 2012.

Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

Soumya Ghosh, Andrei Ungureanu, Erik Sudderth, and David Blei. Spatial distance dependent Chinese restaurant processes for image segmentation. In *Advances in Neural Information Processing Systems*, pages 1476–1484, 2011.

Alexander Gnedin and Jim Pitman. Exchangeable Gibbs partitions and Stirling triangles. *J. Math. Sci. (N.Y.)*, 138(3):5674–5685, 2006. ISSN 1072-3374.

Jeff Goldsmith, Lei Huang, and Ciprian M Crainiceanu. Smooth scalar-on-image regression via spatial Bayesian variable selection. *J. Comput. Graph. Stat.*, 23:46–64, 2014.

Jim E Griffin and Mark FJ Steel. Order-based dependent Dirichlet processes. *J. Am. Stat. Assoc.*, 101(473):179–194, 2006.

Alexander Gundlach-Graham, Marcel Burger, Steffen Allner, Gunnar Schwarz, Hao AO Wang, Luzia Gyr, Daniel Grolimund, Bodo Hattendorf, and Detlef Günther. High-speed, high-resolution, multielemental laser ablation-inductively coupled plasma-time-of-flight mass spectrometry imaging: Part i. instrumentation and two-dimensional imaging of geological samples. *Anal. Chem. (Washington)*, 87(16):8250–8258, 2015. ISSN 0003-2700.

Katherine A Heller and Zoubin Ghahramani. Bayesian hierarchical clustering. In *International Conference on Machine Learning*, pages 297–304, 2005.

J Vernon Henderson, Adam Storeygard, and David N Weil. *Measuring Economic Growth from Outer Space.* National Bureau of Economic Research, Cambridge, Mass, 2009.

Dave Higdon, Jenise Swall, and John Kern. Non-stationary spatial modeling. In *Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting*, pages 761–768. Clarendon Press - Oxford, 1999.

Chris C Holmes and Leonhard Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Anal.*, 1(1 A):145–168, 2006. ISSN 1936-0975.

Guanyu Hu, Junxian Geng, Yishu Xue, and Huiyan Sang. Bayesian spatial homogeneity pursuit of functional data: An application to the U.S. income distribution. *Bayesian Anal.*, 1(1):1–27, 2022.

Lei Huang, Jeff Goldsmith, Philip T Reiss, Daniel S Reich, and Ciprian M Crainiceanu. Bayesian scalar-on-image regression with application to association between intracranial DTI and cognitive outcomes. *NeuroImage*, 83:210–223, 2013.

Noelle J Hum, Perrin E Chamberlin, Brittany L Hambright, Anne C Portwood, Amanda C Schat, and Jennifer L Bevan. A picture is worth a thousand words: A content analysis of Facebook profile photographs. *Comput. Hum. Behav.*, 27(5):1828–1833, 2011. ISSN 0747-5632.

Bradley T Hyman, Gary W Van Hoesen, Antonio R Damasio, and Clifford L. Barnes. Alzheimer's disease: Cell-specific pathology isolates the hippocampal formation. *Science*, 225(4667):1168–1170, 1984. ISSN 00368075.

Alzheimer's Disease International. World Alzheimer report 2019: Attitudes to dementia. *Alzheimer's Dis. Int.: London*, 2019.

Hemant Ishwaran and Lancelot F James. Gibbs sampling methods for stick-breaking priors. *J. Am. Stat. Assoc.*, 96(453):161–173, 2001.

Ernst Ising. *Beitrag zur theorie des ferro-und paramagnetismus*. PhD thesis, Grefe & Tiedemann, 1924.

Jyoti Islam and Yanqing Zhang. A novel deep learning based multi-class classification method for Alzheimer's disease detection using brain MRI data. In *International Conference on Brain Informatics*, pages 213–222. Springer, 2017.

Ronghui Ju, Chenhui Hu, Quanzheng Li, et al. Early diagnosis of Alzheimer's disease based on resting-state brain networks and deep learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 16(1):244–257, 2017.

Jian Kang, Brian J Reich, and Ana Maria Staicu. Scalar-on-image regression via the soft-thresholded Gaussian process. *Biometrika*, 105(1):165–184, 2018. ISSN 14643510. doi: 10.1093/biomet/asx075.

Yunhwan Kim and Jang Hyun Kim. Using computer vision techniques on Instagram to link users' personalities and genders to the features of their photos: An exploratory study. *Inf. Process Manag.*, 54(6):1101–1114, 2018. ISSN 0306-4573.

David S Knopman, Steven T DeKosky, JL Cummings, H Chui, J Corey-Bloom, N Relkin, GW Small, B Miller, and JC Stevens. Practice parameter: diagnosis of dementia (an evidence-based review): report of the quality standards subcommittee of the american academy of neurology. *Neurology*, 56(9):1143–1153, 2001.

Ramesh M Korwar and Myles Hollander. Contributions to the theory of Dirichlet Processes. *Ann. Probab.*, 1(4):705 – 711, 1973. doi: 10.1214/aop/1176996898. URL `https://doi.org/10.1214/aop/1176996898`.

Changwoo Lee, Zhao Tang Luo, and Huiyan Sang. T-LoHo: A Bayesian Regularization Model for Structured Sparsity and Smoothness on Graphs. *Advances in Neural Information Processing Systems*, 34:598–609, 2021.

Fan Li, Tingting Zhang, Quanli Wang, Marlen Z Gonzalez, Erin L Maresh, and James A Coan. Spatial Bayesian variable selection and grouping for high-dimensional scalar-on-image regression. *Ann. Appl. Stat.*, 9:687–713, 2015.

Stan Z Li. *Markov random field modeling in image analysis*. Springer Science & Business Media, 2009.

Antonio Lijoi and Igor Prünster. Models beyond the Dirichlet process. *Collegio Carlo Alberto, Carlo Alberto Notebooks*, 01 2009. doi: 10.1017/CBO9780511802478.004.

Weiming Lin, Tong Tong, Qinquan Gao, Di Guo, Xiaofeng Du, Yonggui Yang, Gang Guo, Min Xiao, Min Du, Xiaobo Qu, et al. Convolutional neural networks-based MRI image analysis for the Alzheimer's disease prediction from mild cognitive impairment. *Front. Neurosci.*, 12: 777, 2018.

Manhua Liu, Danni Cheng, Weiwu Yan, and Alzheimer's Disease Neuroimaging Initiative. Classification of Alzheimer's disease by combination of convolutional and recurrent neural networks using FDG-PET images. *Front. Neuroinform.*, 12:35, 2018.

Albert Y Lo. On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Stat.*, pages 351–357, 1984.

Hongliang Lü, Julyan Arbel, and Florence Forbes. Bayesian nonparametric priors for hidden Markov random fields. *Stat. Comput.*, 30(4):1015–1035, 2020. ISSN 0960-3174.

Steven N MacEachern. ASA proceedings of the section on Bayesian statistical science. In *American Statistical Association*, pages 50–55, 1999.

Katherine A Maloof, Alexis N Reinders, and Kevin R Tucker. Applications of mass spectrometry imaging in the environmental sciences. *Curr. Opin. Environ. Sci. Health*, 18:54–62, 2020. ISSN 2468-5844.

Ryan Martin, Raymond Mess, and Stephen G Walker. Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli*, 23(3):1822–1847, 2017. ISSN 13507265.

Peter McCullagh. Regression models for ordinal data. *J. R. Stat. Soc., B: Stat. Methodol.*, 42 (2), 1980. ISSN 00359246. URL `http://www.jstor.org/stable/2984952`.

Peter McCullagh and John A Nelder. *Generalized linear models*. Routledge, 2019.

Marina Meilă. Comparing clusterings—an information based distance. *J. Multivar. Anal.*, 98 (5):873–895, 2007. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2006.11.013. URL `https://www.sciencedirect.com/science/article/pii/S0047259X06002016`.

Jeffrey W Miller and Matthew T Harrison. Mixture models with a prior on the number of components. *J. Am. Stat. Assoc.*, 113(521):340–356, 2018. doi: 10.1080/01621459.2016.1255636. URL `https://doi.org/10.1080/01621459.2016.1255636`. PMID: 29983475.

Greg Mori. Guiding model search using segmentation. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1417–1423. IEEE, 2005.

Peter Müller, Fernando A Quintana, and Gary L Rosner. A product partition model with regression on covariates. *J. Comput. Graph. Stat.*, 20(1):260–278, 2011. ISSN 1061-8600. URL `http://www.tandfonline.com/doi/abs/10.1198/jcgs.2011.09066`.

Nikhil Naik, Ramesh Raskar, and César A Hidalgo. Cities are physical too: Using computer vision to measure the quality and impact of urban appearance. *Am. Econ. Rev.*, 106(5): 128–132, 2016. ISSN 0002-8282.

Nikhil Naik, Scott Duke Kominers, Ramesh Raskar, Edward L Glaeser, and César A Hidalgo. Computer vision uncovers predictors of physical urban change. *Proc. Natl. Acad. Sci.*, 114 (29):7571–7576, 2017. ISSN 0027-8424.

Radford M Neal. Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Stat.*, 9(2):249–265, 2000. ISSN 1061-8600. URL `http://www.tandfonline.com/doi/abs/10.1080/10618600.2000.10474879`.

John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *J. R. Stat. Soc., A: General*, 135(3):370–384, 1972.

Yang Ni, Peter Müller, Maurice Diesendruck, Sinead Williamson, Yitan Zhu, and Yuan Ji. Scalable Bayesian nonparametric clustering and classification. *J. Comput. Graph. Stat.*, pages 1–45, 2019.

Emma Nichols, Jaimie D Steinmetz, Stein Emil Vollset, Kai Fukutaki, Julian Chalek, Foad Abd-Allah, Amir Abdoli, Ahmed Abualhasan, Eman Abu-Gharbieh, Tayyaba Tayyaba Akram, et al. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: An analysis for the global burden of disease study 2019. *Lancet Public Health*, 7(2): e105–e125, 2022.

Peter Orbanz and Joachim M Buhmann. Nonparametric Bayesian image segmentation. *International Journal of Computer Vision*, 77(1):25–45, 2008. ISSN 0920-5691.

Saffron J O'Neill. Image matters: Climate change imagery in US, UK and Australian newspapers. *Geoforum*, 49:10–19, 2013. ISSN 0016-7185.

Saffron J O'Neill, Maxwell Boykoff, Simon Niemeyer, and Sophie A Day. On the use of imagery for climate change engagement. *Glob. Environ. Change*, 23(2):413–421, 2013. ISSN 0959-3780.

Garritt L Page and Fernando A Quintana. Spatial product partition models. *Bayesian Anal.*, 11(1):265–298, 2016.

Garritt L Page and Fernando A Quintana. Calibrating covariate informed product partition models. *Stat. Comput.*, 28(5):1009–1031, 2018. ISSN 0960-3174.

Marco Palma, Shahin Tavakoli, Julia Brettschneider, and Thomas E Nichols. Quantifying uncertainty in brain-predicted age using scalar-on-image quantile regression. *NeuroImage (Orlando, Fla.)*, 219:116938–116938, 2020. ISSN 1053-8119.

Tianyu Pan, Guanyu Hu, and Weining Shen. Identifying latent groups in spatial panel data using a Markov random field constrained product partition model. *arXiv preprint arXiv:2012.10541*, 2020.

Mihael Perman, Jim Pitman, and Marc Yor. Size-biased sampling of Poisson point processes and excursions. *Probab. Theory Relat. Fields*, 92(1):21–39, 1992.

Jim Pitman. Some developments of the Blackwell-Macqueen urn scheme. *Lecture notes-monograph series*, 30:245–267, 1996. ISSN 0749-2170.

Jim Pitman. Poisson-Kingman partitions. *Lecture notes-monograph series*, 40:1–34, 2003. ISSN 0749-2170.

Jim Pitman. *Combinatorial stochastic processes: Ecole d'ete de probabilit/'es de Saint-Flour XXXII - 2002.* Lecture notes in mathematics ; 1875. Springer, Berlin, 1st ed. 2006. edition, 2006. ISBN 1-280-62579-1.

Renfrey Burnard Potts and Cyril Domb. Some generalized order-disorder transformations. *Math. Proc. Cambridge Philos.*, 48(1):106, January 1952. doi: 10.1017/S0305004100027419.

Annapaola Prestia, Anna Caroli, Sara K Wade, Wiesjie M Van Der Flier, Rik Ossenkoppele, Bart Van Berckel, Frederik Barkhof, Charlotte E Teunissen, Anders Wall, Stephen F Carter, et al. Prediction of AD dementia by biomarkers following the NIA-AA and IWG diagnostic criteria in MCI patients from three European memory clinics. *Alzheimers Dement.*, 11:1191–1120, 2015.

Fernando A Quintana, Peter Müller, Alejandro Jara, and Steven N MacEachern. The dependent Dirichlet process and related models. *Stat. Sci.*, 37(1):24–41, 2022.

Carl Edward Rasmussen and Christopher K I Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning).* The MIT Press, 2005. ISBN 026218253X.

Yordan P Raykov, Alexis Boukouvalas, and Max A Little. Simple approximate MAP Inference for Dirichlet processes. *Electron. J. Stat*, 10:3548–3578, 2016.

Philip T Reiss and R Todd Ogden. Functional generalized linear models with images as predictors. *Biometrics*, 66(1):61–69, 2010.

Philip T Reiss, Maarten Mennes, Eva Petkova, Lei Huang, Matthew J Hoptman, Bharat B Biswal, Stanley J Colcombe, Xi-Nian Zuo, and Michael P Milham. Extracting information from functional connectivity maps via function-on-scalar regression. *NeuroImage*, 56:140–148, 2011.

Philip T Reiss, Lan Huo, Yihong Zhao, Clare Kelly, and R Todd Ogden. Wavelet-domain regression and predictive inference in psychiatric neuroimaging. *Ann. Appl. Stat.*, 9(2):1076, 2015.

David Rossell and Donatello Telesca. Nonlocal priors for high-dimensional estimation. *J. Am. Stat. Assoc.*, 112(517):254–265, 2017. ISSN 0162-1459. URL `http://www.tandfonline.com/doi/abs/10.1080/01621459.2015.1130634`.

Veronika Ročková and Edward I George. The spike-and-slab LASSO. *J. Am. Stat. Assoc.*, 113 (521):431–444, 2018. ISSN 0162-1459. URL `http://www.tandfonline.com/doi/abs/10.1080/01621459.2016.1260469`.

Håvard Rue. *Gaussian Markov random fields: Theory and applications.* Monographs on statistics and applied probability (Series) 104. Chapman & Hall/CRC Press, Boca Raton, 2005. ISBN 1584884320.

Najmeh Neysani Samany. Automatic landmark extraction from geo-tagged social media photos using deep neural network. *Cities*, 93:1–12, 2019. ISSN 0264-2751.

Christof Seiler, Xavier Pennec, and Susan Holmes. Random spatial structure of geometric deformations and Bayesian nonparametrics. In *Geometric Science of Information*, pages 120–127. Springer, Berlin, Heidelberg, 2013.

Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Stat. Sin.*, 4:639–650, 1994.

Ting Shen, Jiehui Jiang, Jiaying Lu, Min Wang, Chuantao Zuo, Zhihua Yu, and Zhuangzhi Yan. Predicting Alzheimer disease from mild cognitive impairment with a deep belief network based on 18F-FDG-PET images. *Mol. Imaging*, 18:1536012119877285, 2019.

Jie Shi, Paul M Thompson, Boris Gutman, and Yalin Wang. Surface fluid registration of conformal representation: Application to detect disease burden and genetic influence on hippocampus. *NeuroImage*, 78:111–134, 2013. ISSN 1053-8119.

Jie Shi, Natasha Lepore, Boris Gutman, Paul Thompson, Leslie Baxter, Richard Caselli, and Yalin Wang. Genetic influence of apolipoprotein E4 genotype on hippocampal morphometry: An N = 725 surface-based Alzheimer's disease neuroimaging initiative study. *Hum. Brain Mapp.*, 35(8):3903–3918, 2014a. ISSN 1065-9471.

Jie Shi, Natasha Leporé, Boris A Gutman, Paul M Thompson, Leslie C Baxter, Richard J Caselli, and Yalin Wang. Genetic influence of apolipoprotein E4 genotype on hippocampal morphometry: An N = 725 surface-based Alzheimer's disease neuroimaging initiative study. *Hum. Brain Mapp.*, 35(8):3903–3918, 2014b. ISSN 1065-9471.

Jun Shi, Xiao Zheng, Yan Li, Qi Zhang, and Shihui Ying. Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. *IEEE J. Biomed. Health Inform.*, 22(1):173–183, 2017.

Minsuk Shin, Anirban Bhattacharya, and Valen E Johnson. Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *Stat. Sin.*, 28(2):1053, 2018. ISSN 1017-0405.

Michael Smith and Ludwig Fahrmeir. Spatial Bayesian variable selection with application to functional magnetic resonance imaging. *J. Am. Stat. Assoc.*, 102(478):417–431, 2007. ISSN 0162-1459.

Michael Smith, Benno Pütz, Dorothee Auer, and Ludwig Fahrmeir. Assessing brain activity through spatial Bayesian variable selection. *NeuroImage*, 20(2):802–815, 2003.

Stephen M Smith, Mark Jenkinson, Heidi Johansen-Berg, Daniel Rueckert, Thomas E Nichols, Clare E Mackay, Kate E Watkins, Olga Ciccarelli, M Zaheer Cader, Paul M Matthews, et al. Tract-based spatial statistics: Voxelwise analysis of multi-subject diffusion data. *NeuroImage*, 31(4):1487–1505, 2006.

Qifan Song and Faming Liang. Nearly optimal Bayesian shrinkage for high dimensional regression. *arXiv preprint arXiv:1712.08964*, 2017.

Larry R Squire and Stuart Zola-Morgan. The medial temporal lobe memory system. *Science*, 253(5026), 1991. ISSN 0036-8075.

Daqiang Sun, Theo GM van Erp, Paul M Thompson, Carrie E Bearden, Melita Daley, Leila Kushan, Molly E Hardt, Keith H Nuechterlein, Arthur W Toga, and Tyrone D Cannon. Elucidating a magnetic resonance imaging-based neuroanatomic biomarker for psychosis: Classification analysis using probabilistic brain atlas and machine learning algorithms. *Biol. Psychiatry (1969)*, 66(11):1055–1060, 2009. ISSN 0006-3223.

Robert H Swendsen and Jian-Sheng Wang. Nonuniversal critical dynamics in Monte Carlo simulations. *Phys. Rev. Lett.*, 58(2):86–88, 1987. ISSN 0031-9007.

Stéphanie Van Der Pas, Jean-Bernard Salomond, and Johannes Schmidt-Hieber. Conditions for posterior contraction in the sparse normal means problem. *Electron. J. Stat.*, 10(1), 2016. ISSN 1935-7524. URL `https://hal.archives-ouvertes.fr/hal-01316155`.

Marianne AA Van Walderveen, Wouter Kamphorst, Philip Scheltens, Jan-Hein TM Van Waesberghe, Rivka Ravid, Jacob Valk, Chris H Polman, and Frederik Barkhof. Histopathologic correlate of hypointense lesions on T1-weighted spin-echo MRI in multiple sclerosis. *Neurology*, 50(5):1282–1288, 1998. ISSN 0028-3878.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.*, 27(5):1413–1432, 2017.

Sara Wade and Zoubin Ghahramani. Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Anal.*, 13(2):559–626, Jun 2018. ISSN 1936-0975. doi: 10.1214/17-ba1073. URL `http://dx.doi.org/10.1214/17-BA1073`.

Hanna Wallach, Shane Jensen, Lee Dicker, and Katherine Heller. An alternative prior process for nonparametric Bayesian clustering. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 892–899, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL `http://proceedings.mlr.press/v9/wallach10a.html`.

Xiao Wang, Hongtu Zhu, and Alzheimer's Disease Neuroimaging Initiative. Generalized scalar-on-image regression models via total variation. *J. Am. Stat. Assoc.*, 112(519):1156–1168, 2017.

Xuejing Wang, Bin Nan, Ji Zhu, and Robert Koeppe. Regularized 3D functional regression for brain image data via Haar wavelets. *Ann. Appl. Stat.*, 8(2):1045, 2014.

Sumio Watanabe and Manfred Opper. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.*, 11 (12), 2010.

Claudia Wehrhahn, Samuel Leonard, Abel Rodriguez, and Tatiana Xifara. A Bayesian approach to disease clustering using restricted Chinese restaurant processes. *Electron. J. Statist.*, 14(1): 1449–1478, 2020. ISSN 1935-7524.

Gerhard Winkler. *Image analysis, random fields and Markov chain Monte Carlo methods: A mathematical introduction*, volume 27. Springer Science & Business Media, 2003.

Raphael Wittenberg, Bo Hu, Luis Barraza-Araiza, and Amritpal Rehill. Projections of older people with dementia and costs of dementia care in the United Kingdom, 2019–2040. *London: London School of Economics*, 2019.

Henrike Wolf, Martin Grunwald, Frithjof Kruggel, Steffi G Riedel-Heller, S. Angerho, Ali Hojjatoleslami, Anke Hensel, Thomas Arendt, and Hermann-Josef Gertz. Hippocampal volume discriminates between normal cognition; questionable and mild dementia in the elderly. *Neurobiol. Aging*, 22:177–186, 2001.

Yun Yang, Martin Wainwright, and Michael Jordan. On the computational complexity of high-dimensional Bayesian variable selection. *Ann. Stat.*, 44(6):2497, 2016. ISSN 00905364. URL `http://search.proquest.com/docview/1845703715/`.

Peng Zhao, Hou-Cheng Yang, Dipak K Dey, and Guanyu Hu. Bayesian spatial homogeneity pursuit regression for count value data. *arXiv preprint arXiv:2002.06678*, 2020.

Tao Zhou, Kim-Han Thung, Xiaofeng Zhu, and Dinggang Shen. Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. *Hum. Brain Mapp.*, 40(3):1001–1016, 2019.