Journal of Machine Learning Research 22 (2021) 1-52 Submitted 2/20; Revised 9/20; Published 1/21

A Bayesian Contiguous Partitioning Method for Learning Clustered Latent Variables

Zhao Tang Luo ZTLUO@STAT.TAMU.EDU
Huiyan Sang HUIYAN@STAT.TAMU.EDU
Bani Mallick BMALLICK@STAT.TAMU.EDU
Department of Statistics
Texas A&M University
College Station, TX 77840, USA

Editor: XuanLong Nguyen

Abstract

This article develops a Bayesian partitioning prior model from spanning trees of a graph, by first assigning priors on spanning trees, and then the number and the positions of removed edges given a spanning tree. The proposed method guarantees contiguity in clustering and allows to detect clusters with arbitrary shapes and sizes, whereas most existing partition models such as binary trees and Voronoi tessellations do not possess such properties. We embed this partition model within a hierarchical modeling framework to detect a clustered pattern in latent variables. We focus on illustrating the method through a clustered regression coefficient model for spatial data and propose extensions to other hierarchical models. We prove Bayesian posterior concentration results under an asymptotic framework with random graphs. We design an efficient collapsed Reversible Jump Markov chain Monte Carlo (RJ-MCMC) algorithm to estimate the clustered coefficient values and their uncertainty measures. Finally, we illustrate the performance of the model with simulation studies and a real data analysis of detecting the temperature-salinity relationship from water masses in the Atlantic Ocean.

Keywords: Bayesian high dimensional regression, posterior concentration, reversible jump Markov chain Monte Carlo, tree-based methods, varying coefficient models

1. Introduction

Spanning trees have gained popularity as a flexible computing tool in computational geometry (Preparata and Shamos, 2012) and clustering analysis (Zahn, 1970; Grygorash et al., 2006), since they are capable of guaranteeing contiguous clustering configurations and detecting clusters with irregular shapes. A spanning tree of a connected graph is a subgraph connecting all vertices in the graph without cycles, in which any two vertices are connected by exactly one edge. A partition of vertices is induced when some edges in a spanning tree are removed such that vertices connected to each other form a cluster. A large body of existing literature on spanning trees is based on machine learning algorithms directly using observed points or point-level features (e.g., Assunção et al., 2006; Guo, 2008; Aydin et al., 2018), whereas the development of spanning tree based modeling and inference framework involving clustered latent variables is still at its infancy.

Our main contribution is to propose a Bayesian model-based spanning tree partitioning method, along with theoretical justifications and efficient computational algorithms, to model clustered latent variables with a focus on spatially clustered varying coefficient models. Most existing literature in spatial regression assume regression coefficients are constants or smoothly varying in space (Fotheringham et al., 2003; Gelfand et al., 2003; Mu et al., 2018). But in many applications, relationships among spatial variables may change abruptly across some boundaries. There is a great need to detect spatially clustered patterns with uncertainty measures in such relationships that allow practitioners to conduct and interpret subregional analysis. The work in this paper is among the first to develop a Bayesian approach for detecting contiguous clusters in regression coefficients.

The Bayesian Spatially Clustered Coefficient Model (BSCC) uses different spanning trees for each covariate and treats them as unknown parameters. Model specifications of space partitions are done by assigning priors on spanning trees, and then the number and the positions of removed edges given a spanning tree. As a result, it allows an adaptive spatial order for cluster detection. Indeed, we show that the sample space of partitions induced from the Bayesian random spanning tree models accommodates all possible contiguous partitions with arbitrary shapes and sizes, defined from connected components of any given graph. Most existing clustering methods which we will review in Section 2 do not possess this property. We emphasize that this property has two important implications. First, it allows us to simplify a complex combinatorial graph partitioning problem into a more compact tree based prior representation that can facilitate computation while maintaining flexibility. Second, the method enjoys great flexibility in the cluster shapes and naturally induces spatially contiguous clusters so that practitioners can interpret clusters as subregions. And the number of clusters is treated as random and determined from data.

An additional advantage of the BSCC is that the Bayesian inference allows us to assess uncertainties in the position of spatial boundaries and the estimated regression models within clusters. Moreover, although we concentrate on the Gaussian spatial regression models in this paper, the proposed partitioning prior model is generic and we propose extensions of the method for embedding in and adaption to various Bayesian hierarchical modeling frameworks that involve latent piecewise constant variables. Finally, since the method is built upon graphs such as triangular meshes, it can be used as a flexible prior on non-exchangeable partitions of data or latent variables distributed on graphs/networks in complex geometric domains.

The regression problem we consider in this article is high-dimensional in nature with n samples and np unknown regression coefficients. We prove that the proposed model achieves posterior consistency, under an asymptotic framework for piecewise constant functions defined on random graphs with a diverging number of vertices. Theoretical guarantee of Bayesian binary treed methods is developed recently (Linero and Yang, 2018; Ročková and van der Pas, 2020; Ročková and Saha, 2019). However, to the best of our knowledge, theoretical properties of spanning tree based Bayesian partition models haven't been investigated in the literature.

The inference of the proposed method is performed in a Bayesian framework, where we extend the conventional reversible jump Markov chain Monte Carlo (RJ-MCMC) algorithm (Green, 1995) by employing various computation strategies such as parallel tempering, low-rank matrix operations, Cholesky factor updates/downdates, and collapsed Gibbs sampling

that greatly improves the computation efficiency for large data sets. The RJ-MCMC procedure allows partitions and spanning trees to be updated adaptively so it can achieve high accuracy in cluster recovery and coefficient estimation, as evidenced by our numerical results that demonstrate striking improvements over competing methods.

The rest of the article is organized as follows. In Section 2, we review other related model-based clustering approaches. In Section 3, we present the Bayesian Spatially Clustered Coefficient regression model, state the theoretical results, develop computation algorithms for Bayesian model implementation, and discuss hyperparameter selection. In Section 4, we present extensions to other hierarchical model settings. Section 5 presents some simulation studies to illustrate the performance of our method. In Section 6, we apply the BSCC model to an ocean temperature and salinity data set. Section 7 concludes our method with some discussion. The proof of the main theoretical results, the detailed implementation and discussion of the RJ-MCMC algorithm, and additional simulation results are provided in the Appendix.

2. Related Work

A large body of model based spatial partition approaches have been proposed in various contexts. Methods such as Markov connected component fields (Gangnon and Clayton, 2000) and product partition models (Hegarty and Barry, 2008; Page and Quintana, 2016) take into account spatial information for clustering, but may not fully guarantee spatial contiguity or allow for arbitrary cluster shapes. Mixture models such as Dirichlet processes (e.g., Gelfand et al., 2005; Blei and Frazier, 2011; Zhang et al., 2014; Ma et al., 2020) are popular Bayesian nonparametric methods for clustering but tend to produce many small clusters. Space partitioning approaches, such as binary treed methods and Voronoi tessellations (Green and Sibson, 1978), have also been widely used in statistics to model responses locally in a region of the input space. Examples of binary treed methods include CART (Breiman et al., 1984; Chipman et al., 1998; Denison et al., 1998), BART (Chipman et al., 2010) and treed Gaussian processes (Gramacy and Lee, 2008; Konomi et al., 2014), where the input space is partitioned into non-overlapping regions by making binary splits recursively. On the other hand, Voronoi tessellation based models (e.g., Knorr-Held and Raßer, 2000; Denison and Holmes, 2001; Kim et al., 2005; Feng et al., 2016) define regions by a number of center locations such that points within a region are closer to its center than any other centers. However, both methods put considerable constraints on the shape of the regions. Voronoi tessellations imply a convexity assumption on the region shapes, and binary treed approaches only produce rectangle shaped regions. Spatial scan statistics (Kulldorff and Nagarwalla, 1995; Kulldorff, 1997) and their variants are also popular approaches to detect spatial clusters. Lin (2014) and Lin et al. (2016) consider Poisson regression models with spatially clustered intercepts using spatial scan statitstics. Lee et al. (2017) develop spatial cluster detection for regression coefficients using spatial scan statistics where the candidate clusters are often assumed to be circular windows.

Our method is motivated from Li and Sang (2019), who propose a fused lasso regularization and optimization method for spatially varying coefficient models, called the SCC, which uses a Euclidean distance based minimum spanning tree (MST) as the "spatial order" to encourage homogeneity between the regression coefficients at two adjacent locations.

The method pursues a sparse solution on the difference between the two edge-connected coefficients, where the zero element indicates that two vertices belong to the same cluster, while the non-zero element corresponds to a cut set of edges which, if removed from the MST, will partition the vertices into a number of clusters. Nevertheless, the method does not produce uncertainty measures of parameter estimations. In addition, a fixed Euclidean MST is used as the spatial order for the regression coefficient of each covariate, which leads to over-clustering especially with small sample sizes as it only induces a restricted partition space to which the actual partition may not belong. In contrast, the Bayesian method developed in this work seeks to find the true spatial order by treating different spanning trees for each covariate as unknown parameters. We will show in Section 5 that this has a significant impact on the results, evidenced by the nearly 80% reduction in the mean square error of BSCC compared with that of SCC in simulation studies.

Most recently, Teixeira et al. (2015, 2019) also develop a Bayesian spatial partitioning model based on spanning trees for the clustering of spatial and spatial temporal responses, respectively. The idea is to construct a random partition model based on random spanning trees, where probabilistic prior models are assigned to the spanning trees and the edge removal probabilities. Their methods have shown a superior performance in terms of clustering accuracy for a number of spatial and spatial temporal clustering tasks, indicating a great potential of the random spanning tree methods. Following a similar spirit, the proposed model offers a new random spanning tree model which complements and differs from theirs in several main aspects. First, we extend beyond a single spanning tree partition model for spatial response data to a general hierarchical model setting for the multiple partitions of latent variables. Second, Teixeira et al. (2019) assume a uniform prior on the spanning tree space and an approximate sampler is used to sample a spanning tree in their MCMC algorithm. We overcome this issue by assigning uniform priors to edge weights in the original graph, which induces priors on the spanning tree space. An exact sampler based on this prior setting is proposed in this paper. Third, they model the prior probability of a partition given a spanning tree by assigning a Beta-distributed prior on the edge inclusion probability without discussing the choice of its hyperparameters. We argue, from a theoretical point of view, that such choice needs careful considerations as it reflects penalty on the number of clusters and has profound effect on the asymptotic behavior of posterior distributions. In this work, we explicitly assign a penalized complexity prior on the number of partitions for which we prove the posterior consistency and design a tailored efficient RJ-MCMC algorithm. In addition, the posterior inference of their partitions relies on a pre-specified threshold of the edge inclusion probability, whereas our method allows us to directly obtain posterior samples of partitions. Finally, we derive a number of original non-asymptotic (e.g., Proposition 2) and asymptotic theories (e.g., Theorem 3), which provide a rigorous justification for the use of random spanning tree models.

3. Methodology

We begin with a varying coefficient regression model in the spatial context to illustrate our Bayesian partitioning method, and outline extensions to other hierarchical models with latent clustered variables in Section 4.

Let $[\{\mathbf{x}(\mathbf{s}_i), y(\mathbf{s}_i)\},\ i = 1, \ldots, n]$ be the spatial data observed at locations $\mathbf{s}_1, \ldots, \mathbf{s}_n \in \mathcal{D} \subset \mathbb{R}^d$, where $\mathbf{x}(\mathbf{s}_i) = \{x_1(\mathbf{s}_i), \ldots, x_p(\mathbf{s}_i)\}^{\mathrm{T}} \in \mathbb{R}^p$ is a vector of covariates and $y(\mathbf{s}_i)$ is a scalar of response. We consider a model

$$y(\mathbf{s}_i) = \mathbf{x}^{\mathrm{T}}(\mathbf{s}_i)\boldsymbol{\beta}(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), \quad \epsilon(\mathbf{s}_i) \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2), \tag{1}$$

where $\boldsymbol{\beta}(\mathbf{s}_i) = \{\beta_1(\mathbf{s}_i), \ldots, \beta_p(\mathbf{s}_i)\}^{\mathrm{T}}$ are unknown coefficients quantifying the relationships between the response and covariates, and $\epsilon(\mathbf{s}_i)$ are independently and identically distributed (i.i.d.) random noises. Clearly, this is a high-dimensional regression problem as there are n samples and np unknown regression coefficients. Assumptions need to be made on $\boldsymbol{\beta}(\mathbf{s}_i)$ to regularize this ill-posed problem. Previous spatial high-dimensional regression models often assume sparsity (Chu et al., 2011) or smoothness in $\boldsymbol{\beta}(\mathbf{s}_i)$ (Gelfand et al., 2003; Mu et al., 2018).

In this paper, we are interested in detecting clustering patterns in $\boldsymbol{\beta}(\mathbf{s}_i)$. For each individual $\beta_m(\mathbf{s}_i)$ $(m = 1, \ldots, p)$, we assume there is a covariate-specific unknown disjoint partition such that $\beta_m(\mathbf{s}_i)$ is a spatially piecewise constant, i.e., $\beta_m(\mathbf{s}_i) = \beta_m(\mathbf{s}_j)$ if $\mathbf{s}_i$ and $\mathbf{s}_j$ are in the same cluster. Alternatively, one may assume there is a single common unknown partition for the whole vector $\boldsymbol{\beta}(\mathbf{s}_i)$, i.e., $\{\beta_1(\mathbf{s}_i), \ldots, \beta_p(\mathbf{s}_i)\}^{\mathrm{T}} = \{\beta_1(\mathbf{s}_j), \ldots, \beta_p(\mathbf{s}_j)\}^{\mathrm{T}}$ if $\mathbf{s}_i$ and $\mathbf{s}_j$ are in the same cluster. The advantage of the first assumption is that it allows us to make inference for the partition in each covariate. We adopt this assumption in this article since one may expect different cluster structures in coefficients for different covariates, but it is straightforward to extend our method to the second one.

In the Bayesian framework, we need to assign priors for the unknown partitions and to sample from the space of partitions for inference. In many spatial applications, as aforementioned, it is desired to consider partitions of locations with spatially contiguous clusters such that only adjacent locations are clustered together. When a complete order of regression coefficients is available, such as in time series problems (Kowal et al., 2019), we could obtain contiguous clusters easily by finding change points in the ordered coefficients. However, it is known that spatial data do not have a natural order. In this paper, we propose to use spanning tree as the spatial order for cluster detection and by treating it as an unknown parameter, our method can adaptively learn the spanning tree order and detect changes in the tree-ordered coefficients.

Below, we give formal definitions for contiguous partitions and clusters, and construct a spanning tree model for such partitions.

3.1 A Prior Model for Contiguous Partitions

Consider an undirected graph $\mathcal{G}_0 = (\mathcal{V}_0, \mathcal{E}_0)$, where $\mathcal{V}_0 = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ is the vertex set and the edge set $\mathcal{E}_0$ is a subset of $\{(\mathbf{s}_i, \mathbf{s}_j) : \mathbf{s}_i, \mathbf{s}_j \in \mathcal{V}_0, \mathbf{s}_i \neq \mathbf{s}_j\}$. Note that in $\mathcal{E}_0$, $(\mathbf{s}_i, \mathbf{s}_j)$ is an unordered pair. Given a spatial data set, we can construct an undirected graph $\mathcal{G}_0$ to represent the relationship of spatial adjacency or neighborhood. For regularly spaced data, a lattice graph is a common choice. For irregularly spaced data, one straightforward way for construction is to connect a vertex with all its neighbors within a certain radius. Another approach is the Delaunay triangulation (Lee and Schachter, 1980), which constructs triangles with a vertex set $\mathcal{V}_0$ such that no vertex is inside the circumcircle of any triangle. In practice, edges longer than a certain threshold are removed to ensure spatial proximity of

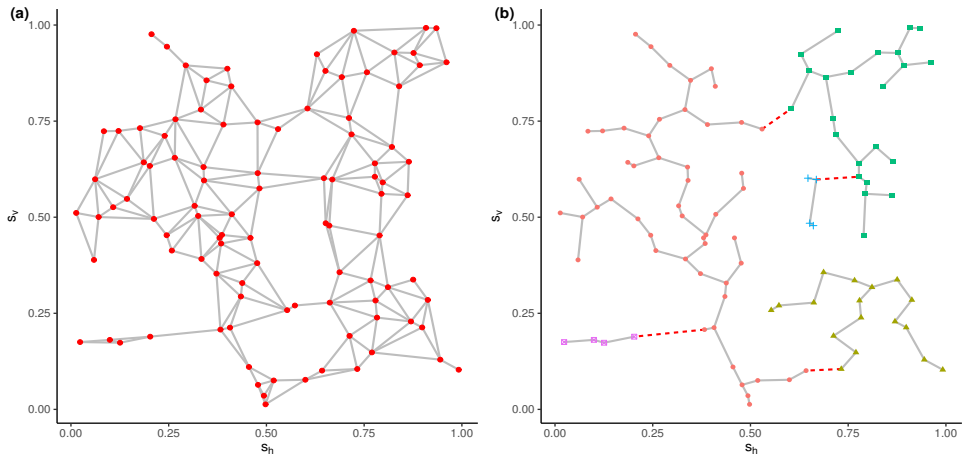

Figure 1: (a) A graph constructed by the Delaunay triangulation, with edges longer than 0.2 removed. (b) An example of a partition with 5 clusters induced by removing the set of red dashed edges from a spanning tree of the graph in (a). Different clusters are marked by different colors.

neighboring vertices. Figure 1(a) demonstrates an example of the Delaunay triangulation. We will show in Section 3.3 that spatial graphs constructed by these two approaches achieve nice theoretical properties.

In graph theory, a sequence of edges $\{(\mathbf{s}_{i_0}, \mathbf{s}_{i_1}), \ldots, (\mathbf{s}_{i_{t-1}}, \mathbf{s}_{i_t})\} \subseteq \mathcal{E}_0$ is called a path of length t between $\mathbf{s}_{i_0}$ and $\mathbf{s}_{i_t}$ if all $\mathbf{s}_{i_j}$'s are distinct. It is called a cycle if $\mathbf{s}_{i_0} = \mathbf{s}_{i_t}$ and all other vertices are distinct. A graph $\mathcal{G}_0$ is said to be connected if for any two vertices there exists a path between them. In this article we assume $\mathcal{G}_0$ is always connected. A subgraph $(\mathcal{V}, \mathcal{E}), \mathcal{V} \subseteq \mathcal{V}_0, \mathcal{E} \subseteq \mathcal{E}_0$ is called a connected component of $\mathcal{G}_0$ if it is connected and there is no path between any vertex in $\mathcal{V}$ and any vertex in $\mathcal{V}_0 \setminus \mathcal{V} := \{\mathbf{s} \in \mathcal{V}_0 : \mathbf{s} \notin \mathcal{V}\}$, the difference between sets $\mathcal{V}_0$ and $\mathcal{V}$. Now one can define spatially contiguous partitions and clusters formally based on the notion of connected components (Teixeira et al., 2019).

Definition 1 *Given an undirected graph $\mathcal{G}_0 = (\mathcal{V}_0, \mathcal{E}_0)$, a subset $\mathcal{C} \subseteq \mathcal{V}_0$ is a spatially contiguous cluster if there exists a connected subgraph $(\mathcal{C}, \mathcal{E}_\mathcal{C}), \mathcal{E}_\mathcal{C} \subseteq \mathcal{E}_0$. A spatially contiguous partition of $\mathcal{V}_0$ is a collection of disjoint spatially contiguous clusters $\pi = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ such that $\cup_{j=1}^k \mathcal{C}_j = \mathcal{V}_0$.*

For conciseness, henceforth, we refer to spatially contiguous partitions and clusters as partitions and clusters, respectively. Our goal is to develop a partition model for a given spatial graph. However, it is a long-standing challenging task since the number of all possible partitions grows rapidly as the number of locations. Following the similar ideas as in Teixeira et al. (2015, 2019) and Li and Sang (2019), we consider a much more compact representation of spatially contiguous partitions based on spanning trees.

A spanning tree of a graph $\mathcal{G}_0$ is defined as a subgraph $\mathcal{T} = (\mathcal{V}_0, \mathcal{E}_\mathcal{T}), \mathcal{E}_\mathcal{T} \subseteq \mathcal{E}_0$ that connects all vertices without any cycle. Therefore, a spanning tree has $|\mathcal{V}_0|$ vertices and

$|\mathcal{V}_0| - 1$ edges, where $|\mathcal{V}_0|$ denotes the cardinality of set $\mathcal{V}_0$. By definition, there can be multiple spanning trees for a given graph. Suppose that weights w_e are assigned to each edge $e \in \mathcal{E}_0$, and then an MST is a spanning tree $(\mathcal{V}_0, \mathcal{E}_\mathcal{T}), \mathcal{E}_\mathcal{T} \subseteq \mathcal{E}_0$ that has the minimal sum of weight $\sum_{e\in\mathcal{E}_\mathcal{T}} w_e$.

A partition with $k+1$ clusters can also be defined by a spanning tree and a subset of edges $\mathcal{E}_k \subseteq \mathcal{E}_\mathcal{T}$ of cardinality k. Specifically, as shown in Figure 1(b), if a set of k edges is removed from a spanning tree $\mathcal{T}$, we create a subgraph of $\mathcal{T}$ that has $k+1$ connected components, and the vertex set of each component forms a cluster. Throughout the paper, we say a partition is *induced* by a spanning tree $\mathcal{T}$ if the partition can be obtained by removing a subset of edges from $\mathcal{E}_\mathcal{T}$.

Below, we show the sample space of partitions induced from random spanning trees accommodates all possible contiguous partitions.

Proposition 2 *Let $\mathcal{G}_0 = (\mathcal{V}_0, \mathcal{E}_0)$ be a connected graph and $\pi = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ be an arbitrary spatially contiguous partition of $\mathcal{V}_0$. There exists at least one spanning tree $\mathcal{T} = (\mathcal{V}_0, \mathcal{E}_\mathcal{T}), \mathcal{E}_\mathcal{T} \subseteq \mathcal{E}_0$ and a subset $\mathcal{E}_{k-1} \subseteq \mathcal{E}$ of cardinality $k-1$ that induce π.*

Proposition 2 implies that we can represent any partition by a spanning tree and a subset of its edge set. It is notable that there is no assumption on the shape and size of each cluster in the partition. The detailed proof of Proposition 2 is provided in Appendix A.1.

The above discussion suggests that the prior model specification for partitions boils down to assigning prior models for spanning trees and the removed edge set given a spanning tree. Conditional on a spanning tree $\mathcal{T}$ and the number of clusters k, we can impose a prior on the space of partitions induced by the spanning tree, or equivalently, on the selection of $(k-1)$-sized subsets of $\mathcal{E}_\mathcal{T}$. Then we can assign a prior on the space of all possible spanning trees and a prior on the number of clusters.

Formally, let $\mathcal{T}^{(m)}$ be a spanning tree of $\mathcal{G}_0$ that can induce $\pi^{(m)}$, the partition associated with the mth covariate. Conditional on $\mathcal{T}^{(m)}$ and k_m, we assume independent uniform priors on all possible $\pi^{(m)}$'s with k_m clusters that are induced by $\mathcal{T}^{(m)}$ (also see Teixeira et al. 2015, 2019 for an alternative prior model on partitions):

$$p\left\{\pi^{(m)} \mid k_m, \mathcal{T}^{(m)}\right\} \propto \mathbf{1}\{\pi^{(m)} \text{ is induced by } \mathcal{T}^{(m)} \text{ and has } k_m \text{ clusters}\}, \tag{2}$$

independently for $m = 1, \ldots, p$, where $\mathbf{1}(\cdot)$ is an indicator function. From the perspective of variable selection, our prior is equivalent to assigning equal probability to all possible selections of $k_m - 1$ edges from the edge set of size $n-1$.

To specify the prior on $\mathcal{T}^{(m)}$, we let $\mathbf{w}^{(m)} = \{w_{ij}^{(m)}\}_{(\mathbf{s}_i,\mathbf{s}_j)\in\mathcal{E}_0}$ be a vector of edge weights associated with the mth covariate, where $w_{ij}^{(m)}$ is the weight for edge $(\mathbf{s}_i, \mathbf{s}_j)$. We assign independent and identical Uniform$(0,1)$ prior on $w_{ij}^{(m)}$ and let $\mathcal{T}^{(m)}$ be the MST given $\mathbf{w}^{(m)}$, i.e.,

$$\mathcal{T}^{(m)} = \text{MST}\{\mathbf{w}^{(m)}\}, \quad w_{ij}^{(m)} \overset{i.i.d.}{\sim} \text{Uniform}(0,1), \tag{3}$$

where MST$(\mathbf{w})$ means an MST of the graph $\mathcal{G}_0$ based on edge weights $\mathbf{w}$ given by Prim's algorithm. Recall that an MST is a spanning tree that has minimal sum of edge weights and it is determined by the edge weights of the original graph. Also note that for any given spanning tree of the original graph, there exists a set of edge weights such that the resulting

MST produces that spanning tree. Therefore, the prior on edge weights induces a prior model on the resulting spanning tree. Note, however, that our induced prior on the space of spanning trees is not uniform, in contrast to the prior in Teixeira et al. (2015, 2019), who use an approximate sampler to update spanning trees. Our prior setting leads to an *exact* update of $\mathcal{T}^{(m)}$ in our RJ-MCMC algorithm (see Section 3.4 for details).

Finally, we assign the following prior to the number of clusters for each coefficient, following the setup of Knorr-Held and Raßer (2000) and Feng et al. (2016):

$$\text{pr}(k_m = k) \propto (1-c)^k, \quad \text{for } k = 1, \ldots, n,\ 0 \leq c < 1 \tag{4}$$

independently for all m. This prior is a geometric distribution truncated to the support $\{1, \ldots, n\}$ with prior mean $E(k_m) = 1/c - n(1-c)^n/\{1-(1-c)^n\}$ when $0 < c < 1$; when $c = 0$ the prior becomes a truncated discrete uniform distribution with prior mean $E(k_m) = (1+n)/2$. It is noted that this prior has a geometrically decaying probability with hyperparameter c controlling the decaying rate, and hence serves as a prior to penalize the model with a large number of clusters. If c is closer to 1 we have a stronger penalization for the large number of clusters. The choice of c plays a crucial role in high-dimensional settings. We will show in Section 3.3 that a theoretically viable choice is to let $-\log(1-c)$ grow at the same rate as $\log(|\mathcal{V}_0|)$. It is possible to assign a prior on k_m conditional on $\mathcal{T}^{(m)}$; however, when there is no *a priori* information about the true partitions and the spanning trees that induce them, we assume that the priors for k_m are independent of $\mathcal{T}^{(m)}$.

3.2 Bayesian Hierarchical Spatially Clustered Coefficient Models

Let $\pi^{(m)} = \{\mathcal{C}_1^{(m)}, \ldots, \mathcal{C}_{k_m}^{(m)}\}\,(m = 1, \ldots, p)$ be the spatial partition of the regression coefficient associated with the mth covariate, $\boldsymbol{\beta}^{(m)} = \{\beta_1^{(m)}, \ldots, \beta_{k_m}^{(m)}\}^{\mathrm{T}}$ be the vector of all different values of the mth coefficient, where $\beta_j^{(m)}$ is the coefficient value associated with cluster $\mathcal{C}_j^{(m)}$. With a slight abuse of notation, we denote $\mathbf{s}_i \in \mathcal{C}_{j_1}^{(1)} \cap \cdots \cap \mathcal{C}_{j_p}^{(p)}$ for some $j_1, \ldots, j_p$, if the regression coefficient at $\mathbf{s}_i$ for the mth covariate belongs to $\mathcal{C}_{j_m}^{(m)}$. Choosing conjugate priors for other model parameters, our hierarchical model can be written as

$$y(\mathbf{s}_i) \mid \{\boldsymbol{\beta}^{(m)}\}_{m=1}^p, \sigma^2, \lambda, \left\{\pi^{(m)}, k_m, \mathbf{w}^{(m)}\right\}_{m=1}^p \overset{ind.}{\sim} \mathrm{N}\left\{\sum_{m=1}^p \beta_{j_m}^{(m)} x_m(\mathbf{s}_i),\ \sigma^2\right\}, \tag{5a}$$

$$\boldsymbol{\beta}^{(m)} \mid \sigma^2, \lambda, \pi^{(m)}, k_m \overset{ind.}{\sim} \mathrm{N}_{k_m}\left(\mathbf{0}, \lambda^{-1}\sigma^2 \boldsymbol{\Sigma}_m\right), \tag{5b}$$

$$\left\{\pi^{(m)}, k_m, \mathbf{w}^{(m)}\right\}_{m=1}^p \sim \prod_{m=1}^p p\left\{\pi^{(m)} \mid k_m, \mathbf{w}^{(m)}\right\} p(k_m) p\{\mathbf{w}^{(m)}\}, \tag{5c}$$

$$\sigma^2 \sim \mathrm{IG}(a_0/2, b_0/2), \tag{5d}$$

$$\lambda \sim \mathrm{Gamma}(c_0/2, d_0/2), \tag{5e}$$

where N_{k_m} represents the k_m-dimensional multivariate normal distribution, $\boldsymbol{\Sigma}_m$ is a $k_m \times k_m$ covariance matrix, $\mathrm{IG}(a, b)$ is the inverse-Gamma distribution, $\mathrm{Gamma}(a, b)$ is the Gamma distribution in shape-rate parameterization, and a_0, b_0, c_0, d_0 are hyperparameters. The

notation "*ind.*" means that we assume (5a) holds independently for all $i = 1, \ldots, n$ and place independent prior (5b) on $\boldsymbol{\beta}^{(m)}$ for all $m = 1, \ldots, p$. The priors in (5c), (5d), and (5e) are also assumed to be mutually independent. We allow the prior of $\boldsymbol{\beta}^{(m)}$ to accommodate spatial dependence among clusters if one assumes spatial structures in $\boldsymbol{\Sigma}_m$. In the case where there is no prior information on the spatial dependence structure of $\boldsymbol{\beta}^{(m)}$, one can set $\boldsymbol{\Sigma}_m = \mathbf{I}_{k_m}$, the $k_m \times k_m$ identity matrix. We only consider this independent case in this article for simplicity. Note that it is also possible to choose other priors for $\boldsymbol{\beta}^{(m)}$, σ^2, and λ. For example, one can place non-informative priors on σ^2 and λ. And we specify independent and identical priors for the partitions of each regression coefficient, $\{\pi^{(m)}, k_m, \mathbf{w}^{(m)}\}$, following the method described in Section 3.1.

3.3 Theoretical Properties

To ease notations, we present our theoretical results for $p = 1$ case,

$$y(\mathbf{s}_i) = x(\mathbf{s}_i)\beta(\mathbf{s}_i) + \epsilon(\mathbf{s}_i),$$

where $x(\mathbf{s}_i), \beta(\mathbf{s}_i) \in \mathbb{R}$, though the result can be extended to a more general case. In this subsection, we let x_i and β_i denote $x(\mathbf{s}_i)$ and $\beta(\mathbf{s}_i)$, respectively. Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^{\mathrm{T}}$. Given a spanning tree $\mathcal{T} = (\mathcal{V}_0, \mathcal{E}_\mathcal{T})$, we define $G^*_\mathcal{T} = \{(\mathbf{s}_i, \mathbf{s}_j) \in \mathcal{E}_\mathcal{T} : \beta^*_i - \beta^*_j \neq 0\}$, where β^*_i is the true value of β_i with the corresponding true partition denoted as π^*. We assume that the number of clusters in π^*, denoted by k^*, is *fixed*. $G^*_\mathcal{T}$ represents the edges of $\mathcal{T}$ that have nonzero jumps in $\boldsymbol{\beta}^*$, the true value of $\boldsymbol{\beta}$. When π^* is induced by $\mathcal{T}$ so that there is exactly one jump in $\mathcal{E}_\mathcal{T}$ that crosses two distinct clusters, $|G^*_\mathcal{T}| + 1$ equals k^*. Otherwise, $|G^*_\mathcal{T}|$ will be larger than $k^* - 1$. Indeed, in this case, we get a nested partition of the true π^* when $G^*_\mathcal{T}$ is removed from $\mathcal{E}_\mathcal{T}$. We let $\mathbb{T}_n$ be the set of all spanning trees of the graph $\mathcal{G}_0$ with n vertices, and define $g^*_n = \max_{\mathcal{T} \in \mathbb{T}_n} |G^*_\mathcal{T}| + 1$ such that $g^*_n - 1$ is the maximum number of edges that have nonzero jumps in $\boldsymbol{\beta}^*$ among all possible spanning trees.

We adopt the following asymptotic notations. Given two positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n \succ b_n$ means $\lim_{n\to\infty}(a_n/b_n) = \infty$ and $a_n \asymp b_n$ means $0 < \liminf_{n\to\infty}(a_n/b_n) \leq \limsup_{n\to\infty}(a_n/b_n) < \infty$. We also denote the L_2 norm by $\|\cdot\|$.

Our results on posterior consistency rely on the following assumptions as $n \to \infty$:

(C1) x_i is non-random, and $|x_i| \leq M_0$ for some $M_0 > 0$ and any i.

(C2) $\log(\max_{1\leq i\leq n} |\beta^*_i|/\sigma^*) = O(\log n)$, where σ^* is the fixed true value of σ as n grows.

(C3) The graph satisfies $g^*_n \prec n/\log n$. Let P_n be the number of all unique partitions nested in π^* that have at most $g^*_n q_n$ clusters for a given sequence $q_n \to \infty$. We assume that $\log P_n = O(g^*_n \log n)$.

(C4) $1 - c \asymp n^{-\alpha}$ for some constant $\alpha > 0$.

Assumption (C1) is a commonly adopted assumption which states that the covariate space is bounded. Assumption (C2) constrains the asymptotic growth rate of the magnitude of the true coefficients (see, e.g., Song and Cheng, 2020). Assumption (C3) restricts

the number of edges that have nonzero jumps in $\boldsymbol{\beta}^*$ for any possible spanning tree, and essentially excludes graphs that are too dense. We will show that $g_n^* \prec n/\log n$ is satisfied by commonly used spatial designs and graphs with probability tending to 1 in Proposition 5. The second part of Assumption (C3) constrains the complexity of the space of partitions to ensure the existence of test functions in our proof. Assumption (C4) imposes restriction on the tail behavior of our penalized complexity prior such that it provides enough probability mass around the true model. Similar conditions on prior hyperparameters are common in Bayesian high-dimensional regression literature (see, e.g., Armagan et al., 2013; Yang et al., 2016).

The following theorem states that if Assumptions (C1)-(C4) hold, the posterior distribution of the predicted responses from BSCC model concentrates around the true means asymptotically.

Theorem 3 *(Posterior consistency for fixed spatial graph designs) Let $\boldsymbol{\mu}$ and $\boldsymbol{\mu}^*$ be n-dimensional vectors such that $\mu_i = x_i\beta_i$ and $\mu_i^* = x_i\beta_i^*$. Under Assumptions (C1)-(C4), there exists a constant $M_1 > 0$ and $\varepsilon_n \asymp \sqrt{g_n^* \log n/n}$ such that the posterior distribution satisfies*

$$\Pi_n\left(\frac{1}{\sqrt{n}}\|\boldsymbol{\mu} - \boldsymbol{\mu}^*\| \geq M_1\sigma^*\varepsilon_n \mid \mathbf{y}\right) \longrightarrow 0$$

with probability one.

The detailed proof is provided in Appendix A.2.

We verify that the first part of Assumption (C3) holds with probability tending to 1 for some common choices of spatial designs and spatial graphs. In the spatial context, we consider an asymptotic framework for piecewise constant functions that are defined on spatial random graphs with a diverging number of vertices in $\mathbb{R}^2$. Before giving the proposition, we will first describe a formulation for the sampling region and a nonuniform random spatial design for irregularly spaced data, and then a technical definition of piecewise constant functions will be introduced.

Below, we state assumptions on the sampling region $\mathcal{D}$ and the sampling design of n points $\mathbf{s}_1^D, \cdots, \mathbf{s}_n^D$ in $\mathcal{D}$.

> (C5) *Spatial sampling region.* Assume $\mathcal{D}$ is homeomorphic to the unit square with the Euclidean metric and a bi-Lipschitz homeomorphism $\mathcal{F}_D : \mathcal{D} \to [0,1]^2$. Under this assumption, $\mathcal{S}_n = (\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_n)$, where $\mathbf{s}_i = \mathcal{F}_D(\mathbf{s}_n^D)$ for $i = 1, \cdots, n$ is the mapping of the original sampling point to $[0,1]^2$. This condition allows us to consider a study region with a variety of shapes as long as it is topologically equivalent to a unit square.
>
> (C6) *Spatial design and spatial graph.* Given $n \in \mathbb{N}$, we assume $\mathcal{S}_n$ is a sequence of n independent points where each point is distributed on $[0,1]^2$ with a probability density function p_s such that $0 < p_s^{\min} \leq p_s(\mathbf{s}) \leq p_s^{\max} < \infty$. We assume the spatial graph on $\mathcal{S}_n$ is constructed by (i) the radius-based nearest neighbor (R-NN) graph with a radius $\gamma_1 \asymp \sqrt{\log n/n}$ and $\gamma_1 > \gamma_0$, where γ_0 is the maximum edge length of the MST on $\mathcal{S}_n$; or (ii) the Delaunay triangulation graph where the edges are removed if they are longer than γ_2, where $\gamma_2 \asymp \sqrt{\log n/n}$ and $\gamma_2 > \gamma_0$. We will refer to it as the *restricted Delaunay triangulation* in the proof.

Notice that $|G^*_{\mathcal{T}}|$ is essentially the number of edges across the cluster boundaries of the true coefficient, which is viewed as a piecewise constant function defined on the spatial domain $\mathcal{D}$. To bound $\max_{\mathcal{T}\in\mathbb{T}_n}|G^*_{\mathcal{T}}|$, we work with the following definition of piecewise constant functions, in which the cluster boundary set is introduced.

Definition 4 *(see, e.g., Willett et al. 2006) We say that a function $g : [0,1]^2 \to \mathbb{R}$ is piecewise constant if there exists a cluster boundary set $\mathcal{B}_g$ such that:*

1. The cluster boundary set $\mathcal{B}_g$ has a υ_n-covering number $N(\mathcal{B}_g, \upsilon_n, \|\cdot\|) \leq M_2\upsilon_n^{-1}$, for some constant $M_2 > 0$.

2. The function g is locally constant on $[0,1]^2\backslash\mathcal{B}_g$, i.e., $g(\mathbf{s}) = g(\mathbf{s}')$ if $\mathbf{s}$ and $\mathbf{s}'$ belong to the same connected component of $[0,1]^2\backslash\mathcal{B}_g$.

The next proposition states that the condition $g^*_n \prec n/\log n$ is met under Assumption (C5) and (C6) with high probability. The proof is delayed to Appendix A.3.

Proposition 5 *Assume further the true regression coefficient $\beta^{*,D}$ is a function $\beta^{*,D}(\mathbf{s}^D) : \mathcal{D} \to \mathbb{R}$ such that $\beta^*(\mathbf{s}) : [0,1]^2 \to \mathbb{R}$ is piecewise-constant on $[0,1]^2$ with the boundary set $\mathcal{B}_{\beta^*}$. Under Assumptions (C5) and (C6), there exist positive constants $M_3, M_4 > 0$, such that $g^*_n \leq M_3\sqrt{n\log n}$ holds with probability at least $1 - \exp\left(-M_4\sqrt{n\log n}\right)$.*

Combining Theorem 3 and Proposition 5 gives the following posterior concentration result under the random spatial graph in Assumption (C6). The proof is given in Appendix A.4.

Corollary 6 *(Posterior consistency for random spatial graph designs) Let $\tilde{P}_n$ be the number of all unique partitions nested in π^* that have at most $M_3 q_n\sqrt{n\log n}$ clusters, where π^* is the true partition corresponding to $\beta^*(\mathbf{s})$ in Proposition 5 given $\mathcal{S}_n$. Assume that $\log\tilde{P}_n \leq M_5 n^{1/2}\log^{3/2} n$ with probability tending to one for some constant $M_5 > 0$ not depending on $\mathcal{S}_n$. Under Assumptions (C1), (C2) and (C4)-(C6), there exists a constant $M_6 > 0$ and $\tilde{\varepsilon}_n \asymp n^{-1/4}\log^{3/4} n$ such that the posterior distribution satisfies*

$$\Pi_n\left(\frac{1}{\sqrt{n}}\|\boldsymbol{\mu} - \boldsymbol{\mu}^*\| \geq M_6\sigma^*\tilde{\varepsilon}_n \mid \mathbf{y}, \mathcal{S}_n\right) \longrightarrow 0$$

in probability.

3.4 Computational Strategies

We extend conventional RJ-MCMC algorithm to sample the partitions, the values of coefficients, and other parameters simultaneously. Standard RJ-MCMC algorithm may suffer from poor mixing and slow convergence, because of the potentially multimodal posterior (which is common in many partition models such as Chipman et al. 1998) and the large space of spanning trees. We propose several strategies to address computational issues.

Let $\mathbf{y} = \{y(\mathbf{s}_1), \ldots, y(\mathbf{s}_n)\}^{\mathrm{T}}$ be the vector of responses, $\tilde{\boldsymbol{\beta}} = [\{\boldsymbol{\beta}^{(1)}\}^{\mathrm{T}}, \ldots, \{\boldsymbol{\beta}^{(p)}\}^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^K$ be the stacked vector of coefficients, where $K = \sum_{m=1}^p k_m$, and $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{X}}_1 \cdots \widetilde{\mathbf{X}}_p] \in \mathbb{R}^{n\times K}$ be the design matrix associated with $\tilde{\boldsymbol{\beta}}$, where each sub-matrix $\widetilde{\mathbf{X}}_m \in \mathbb{R}^{n\times k_m}$ is

constructed in the following way. The (i,j)th element of $\widetilde{\mathbf{X}}_m$ is set to be $x_m(\mathbf{s}_i)$ if the ith location belongs to cluster $\mathcal{C}_j^{(m)}$ for some $j \in \{1,\ldots,k_m\}$; otherwise, it is set to be zero.

We first rewrite the data model and the prior model for $\tilde{\boldsymbol{\beta}}$ in matrix forms as

$$\begin{aligned}\mathbf{y} \mid \tilde{\boldsymbol{\beta}}, \sigma^2, \lambda, \left\{\pi^{(m)}, k_m, \mathbf{w}^{(m)}\right\}_{m=1}^{p} &\sim \mathrm{N}_n\left(\widetilde{\mathbf{X}}\tilde{\boldsymbol{\beta}}, \sigma^2\mathbf{I}_n\right)\\ \tilde{\boldsymbol{\beta}} \mid \sigma^2, \lambda, \left\{\pi^{(m)}, k_m, \mathbf{w}^{(m)}\right\}_{m=1}^{p} &\sim \mathrm{N}_K\left(\mathbf{0}, \lambda^{-1}\sigma^2\mathbf{I}_K\right)\end{aligned}$$

Integrating out $\tilde{\boldsymbol{\beta}}$, the marginal distribution of $\mathbf{y}$ becomes

$$\mathbf{y} \mid \sigma^2, \lambda, \left\{\pi^{(m)}, k_m, \mathbf{w}^{(m)}\right\}_{m=1}^{p} \sim \mathrm{N}_n\left(\mathbf{0}, \sigma^2\mathbf{P}_\lambda\right), \tag{6}$$

where $\mathbf{P}_\lambda = \mathbf{I}_n + \lambda^{-1}\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^{\mathrm{T}}$. It allows us to sample from the collapsed posterior distribution of $\left[\left\{\pi^{(m)}, k_m, \mathbf{w}^{(m)}\right\}_{m=1}^{p}, \sigma^2, \lambda\right]$ as follows

$$\begin{aligned}p\left[\left\{\pi^{(m)}, k_m, \mathbf{w}^{(m)}\right\}_{m=1}^{p}, \sigma^2, \lambda \;\middle|\; \mathbf{y}\right] \propto\\ (\sigma^2)^{-n/2}|\mathbf{P}_\lambda|^{-1/2}\exp\left(-\frac{1}{2\sigma^2}\mathbf{y}^{\mathrm{T}}\mathbf{P}_\lambda^{-1}\mathbf{y}\right)\cdot(\sigma^2)^{-a_0/2-1}\exp\left(-\frac{b_0}{2\sigma_y^2}\right)\times\\ \lambda^{c_0/2-1}\exp\left(-\frac{d_0}{2}\lambda\right)\cdot\prod_{m=1}^{p}\left\{\binom{n-1}{k_m-1}^{-1}\cdot(1-c)^{k_m}\right\}.\end{aligned} \tag{7}$$

Standard uncollapsed MCMC can lead to poor mixing due to the strong dependence of $\tilde{\boldsymbol{\beta}}$. This collapsed posterior greatly improves the efficiency and mixing in searching the posterior of partitions.

Since the number of clusters in each partition is unknown, we employ the reversible jump MCMC (Green, 1995) to sample from the posterior in (7). Within each iteration of RJ-MCMC, we further iterate through each covariate from $m = 1$ to p. In each inner iteration, one of the following four possible moves is performed.

(a) *Birth*: Fixing the spanning tree $\mathcal{T}^{(m)}$, add a new cluster to $\pi^{(m)}$ by splitting an existing cluster.

(b) *Death*: Fixing $\mathcal{T}^{(m)}$, randomly remove an existing cluster by merging it into an adjacent cluster.

(c) *Change*: Fixing $\mathcal{T}^{(m)}$, randomly remove an existing cluster by merging it into an adjacent cluster, and then add a new cluster by splitting an existing cluster, so that the number of clusters remains unchanged.

(d) *Hyper*: Update parameters σ^2, λ, and $\mathbf{w}^{(m)}$ (and hence $\mathcal{T}^{(m)}$). Specifically, σ^2 is updated by a Gibbs step, $\mathbf{w}^{(m)}$ is updated by sampling a set of edge weights such that the resulting MST can induce the current sample of $\pi^{(m)}$ using an *exact* algorithm derived below, and λ is updated by a Metropolis-Hastings procedure with a symmetric random walk proposal.

The exact update of $\mathcal{T}^{(m)}$ is done by a Metropolis-Hastings algorithm to sample edge weights followed by Prim's algorithm. From (7) we have the full conditional of $\mathbf{w}^{(m)}$ proportional to

$$\mathbf{1}\left[\pi^{(m)} \text{ is induced by MST}\{\mathbf{w}^{(m)}\} \text{ and } 0 < w_{ij}^{(m)} < 1 \text{ for all } (\mathbf{s}_i, \mathbf{s}_j) \in \mathcal{E}_0\right]. \tag{8}$$

We propose a new $\mathbf{w}^{(m)}$ by sampling $w_{ij}^{(m)}$ from i.i.d. Uniform$(1/2, 1)$ if $\mathbf{s}_i$ and $\mathbf{s}_j$ are in different clusters and sampling $w_{ij}^{(m)}$ from i.i.d. Uniform$(0, 1/2)$ if $\mathbf{s}_i$ and $\mathbf{s}_j$ are in a same cluster. The resulting spanning tree from Prim's algorithm based on the proposed edge weights is guaranteed to induce the current partition $\pi^{(m)}$ (Teixeira et al., 2015). The acceptance probability for $\mathbf{w}^{(m)}$ is always 1. To see this, first notice that (8) remains the same for the proposed weights, and thus the likelihood ratio is 1. The prior ratio is also 1 since we assume a uniform prior on $\mathbf{w}^{(m)}$. Due to the design of proposal distribution, the proposal ratio is again 1 as the sets of cross-cluster edges and within-cluster edges are preserved. The sample of $\mathcal{T}^{(m)}$ is the MST generated by Prim's algorithm. Note that this sampler is exact in the sense that there is no approximation in this sampling scheme. The induced chain of spanning trees is irreducible, as suggested by the following proposition.

Proposition 7 *For any spanning tree $\mathcal{T}$ of $\mathcal{G}_0$ that induces a partition π, the spanning tree sampling algorithm described above generates $\mathcal{T}$ with strictly positive probability.*

The proof of Proposition 7 is postponed to Appendix A.5.

We set the probability for each move to be $r_B(k) = 0.425, r_D(k) = 0.425, r_C(k) = 0.1$, and $r_H(k) = 0.05$, respectively. Adjustments are made for boundary cases when $k_m = 1$ or n. The choice of these probabilities works well empirically in our studies. But we remark that these probabilities can be modified if desired. For the first three moves, a new partition is accepted with probability $\alpha_1 = \min(1,\ \mathcal{A}\cdot\mathcal{P}\cdot\mathcal{L})$, where $\mathcal{A}, \mathcal{P}, \mathcal{L}$ are the prior ratio, proposal ratio, and likelihood ratio, respectively. For the fourth move, hyper, the spanning tree is updated adaptively to the current estimate of the partition, thus allowing for the search of spanning trees that can induce the true partitions. The RJ-MCMC algorithm is summarized in Algorithm 1 and detailed in Appendix B.

After obtaining samples of $\boldsymbol{\theta} = \left[\left\{\pi^{(m)}, k_m, \mathbf{w}^{(m)}\right\}_{m=1}^{p}, \sigma^2, \lambda\right]$, it is straightforward to obtain a sample of $\tilde{\boldsymbol{\beta}}$ by sampling from $p(\tilde{\boldsymbol{\beta}} \mid \boldsymbol{\theta}, \mathbf{y})$, which takes the following closed form

$$\tilde{\boldsymbol{\beta}} \mid \boldsymbol{\theta}, \mathbf{y} \sim \mathrm{N}_K\left\{(\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}} + \lambda\mathbf{I}_K)^{-1}\widetilde{\mathbf{X}}^{\mathrm{T}}\mathbf{y},\ \sigma^2(\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}} + \lambda\mathbf{I}_K)^{-1}\right\}.$$

One computation bottleneck is the evaluation of the likelihood function in (6), which involves the inversion of the $n \times n$ matrix $\mathbf{I}_n + \lambda^{-1}\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^{\mathrm{T}}$. Recall that the dimension of $\widetilde{\mathbf{X}}$ is $n \times K$, where K is the summed number of clusters over all covariates. As K is typically much smaller than n, we take advantage of the low-rank structure and apply the Sherman-Woodbury-Morrison formula to reduce the problem to computing $\mathbf{y}^{\mathrm{T}}\widetilde{\mathbf{X}}(\lambda\mathbf{I}_K + \widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^{\mathrm{T}}\mathbf{y}$.

The update of the above quadratic form in each MCMC iteration can be further simplified by the fact that most columns of $\widetilde{\mathbf{X}}$ are unchanged in a birth, death, or change step. For instance, in a birth step, $\widetilde{\mathbf{X}}$ is changed by adding one column and modifying another, which can be done by removing one column and adding two. The Cholesky decomposition of $\lambda\mathbf{I}_K + \widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}}$ can therefore be updated efficiently from the Cholesky factor

Algorithm 1: RJ-MCMC algorithm

Initialize partitions and edge weights $\left\{\pi^{(m)}, k_m, \mathbf{w}^{(m)}\right\}_{m=1}^{p}$;
for $t \leftarrow 1$ **to** T **do**
  for $m \leftarrow 1$ **to** p **do**
    Propose a *birth*, *death*, *change*, or *hyper* step with certain probabilities ;
    if *birth step* **then**
      Propose a new cluster by splitting an existing cluster in $\pi^{(m)}$;
    else if *death step* **then**
      Randomly remove an existing cluster by merging it to a neighboring cluster in $\pi^{(m)}$;
    else if *change step* **then**
      Randomly remove an existing cluster by merging it to a neighboring cluster, then propose a new cluster by splitting an existing cluster ;
    else if *hyper step* **then**
      Update σ^2 using Gibbs sampling ;
      Update $\mathbf{w}^{(m)}$ (and hence $\mathcal{T}^{(m)}$) by a Metropolis-Hastings step ;
      Update λ by a Metropolis-Hastings step ;
    Accept proposed change with probability α_1;
Discard samples from burn-in period;

at the previous step following the supernodal sparse Cholesky update/downdate algorithms (Chen et al., 2008; Osborne, 2010). $\widetilde{\mathbf{X}}^{\mathrm{T}}\mathbf{y}$ can also be updated by changing one element and adding/removing another. The overall time complexity to update the quadratic term is $O(nK)$, whereas directly evaluating it requires $O(nK^2)$ operations.

Finally, it is common to have multimodal posterior distributions for some parameters near cluster boundaries. We employ parallel tempering (Geyer, 1991) to better explore the posterior and improve mixing. Specifically, we run d chains in parallel with the likelihood function tempered by different "temperatures". The target distribution of the jth chain is

$$p_j(\boldsymbol{\theta} \mid \mathbf{y}) \propto \{\ell(\boldsymbol{\theta} \mid \mathbf{y})\}^{\nu_j} p(\boldsymbol{\theta}),$$

where $\ell(\boldsymbol{\theta} \mid \mathbf{y})$ is the likelihood, $p(\boldsymbol{\theta})$ is the prior, and $1 = \nu_1 > \cdots > \nu_d > 0$ are called the inverse temperatures. Note that the first chain has the same target distribution as the conventional RJ-MCMC algorithm does. We choose the inverse temperatures from the sigmoidal temperature ladder used in Gramacy and Taddy (2010) and Payne et al. (2020). Every a certain number of iterations (which is called a swap interval), all chains swap their parameters $\boldsymbol{\theta}$ with their neighboring chains with some probabilities. For a swap attempt between the jth and the $(j+1)$th chains, the acceptance probability is given by

$$\alpha_2 = \min\left\{1,\ \frac{p_j(\boldsymbol{\theta}_{j-1} \mid \mathbf{y}) \cdot p_{j-1}(\boldsymbol{\theta}_j \mid \mathbf{y})}{p_j(\boldsymbol{\theta}_j \mid \mathbf{y}) \cdot p_{j-1}(\boldsymbol{\theta}_{j-1} \mid \mathbf{y})}\right\},$$

where $\boldsymbol{\theta}_j$ is the parameter in the jth chain. The draws from the first chain are the MCMC samples from the desired posterior distribution. Generally, a chain with lower inverse temperature has higher acceptance rates in reversible jump moves, allowing it to reach regions

that are hard to visit by chains with higher inverse temperatures. Samples from these regions can then be passed to chains with higher inverse temperatures by the swap procedure, which speeds up the exploration of the posterior sample space.

3.5 Selection of c

The hyperparameter c has profound effect on the asymptotic behavior of posterior distributions and thus it is rather important to carefully specify the order of c with respect to the sample size n. Following Assumption (C4), we set $1 - c = n^{-\alpha}$ so that the posterior consistency result in Theorem 3 can be guaranteed. In practice the constant α is unknown and the selection of c boils down to choosing appropriate positive α.

We propose to use Watanabe-Akaike information criterion (WAIC; Watanabe, 2010) to select α, which takes the form

$$\text{WAIC} = -2\sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} \ell(\boldsymbol{\theta}^s|y_i)\right) + 2p_{\text{WAIC}},$$

where y_i is a shorthand for $y(\mathbf{s}_i)$, $\boldsymbol{\theta}^s$ is the sth ($s = 1, \ldots, S$) MCMC sample of the parameters, and p_{WAIC} is a term quantifying model complexity. In addition to the widely used complexity term

$$p_{\text{WAIC}_1} = 2\sum_{i=1}^{n}\left\{\log\left(\frac{1}{S}\sum_{s=1}^{S} \ell(\boldsymbol{\theta}^s|y_i)\right) - \frac{1}{S}\sum_{s=1}^{S}\log \ell(\boldsymbol{\theta}^s|y_i)\right\},$$

a numerically more stable alternative

$$p_{\text{WAIC}_2} = V_{s=1}^{S} \log \ell(\boldsymbol{\theta}^s|y_i),$$

where $V_{s=1}^{S}$ represents the unbiased sample variance, is also recommended (Gelman et al., 2014). An α that leads to lower WAIC is preferred. Note that WAIC is applicable because our model assumes conditional independence of $\mathbf{y}$ given the parameters and the spatial dependence is modelled via the latent partition structure of the parameters.

4. Extensions to Other Hierarchical Models

The preceding Bayesian spanning tree partitioning prior model can be extended to other hierarchical model settings. Let $\{y_i,\ i = 1, \ldots, n\}$ be the observations at each vertex of an undirected graph $\mathcal{G}_0 = (\mathcal{V}_0, \mathcal{E}_0)$, where $\mathcal{G}_0$ encodes prior knowledge on the relationships among vertices to encourage neighboring vertices sharing identical models. Examples of such graphs can go beyond spatial domains to more complex domains such as brain networks, road networks or social networks.

Given a partition $\pi = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ of the vertices, we let $\mathbf{y}_{c_1}, \cdots, \mathbf{y}_{c_k}$ denote the corresponding partition of observations. Conditional on the vector of latent cluster-specific model parameters, denoted as $\boldsymbol{\theta}_{(j)}$, $j = 1, \ldots, k$, and the vector of global model parameters $\boldsymbol{\eta}$, we assume a conditionally independent data-level model for $\mathbf{y}_{c_1}, \cdots, \mathbf{y}_{c_k}$ as follows

$$\prod_{j=1}^{k} f(\mathbf{y}_{c_j} \mid \boldsymbol{\theta}_{(j)}, \boldsymbol{\eta}, \pi)$$

The Bayesian approach then proceeds by assigning prior models for $\boldsymbol{\theta}_j$ and $\boldsymbol{\eta}$ conditional on the graph partition π. Finally, the Bayesian spanning tree partitioning prior model introduced in Section 3.1 is adopted to model π.

There are many general settings in which the above hierarchical model with clustered latent variables arises as the data-level model can take various forms. One example is to consider generalized linear models (GLMs) for non-Gaussian data, which were also considered in Teixeira et al. (2015, 2019) for a spatial Poisson count response data. Commonly used non-Gaussian data level models include: (i) binary response at locations, modeled using logit or probit regression, and (ii) count data at locations, modeled using Poisson regression. We model the link function of mean responses using a clustered varying coefficient model,

$$g(E(y_i)) = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_{(j)}, \text{ for } i \in \mathcal{C}_j \tag{9}$$

The prior models for the partitions can be assigned in the same way as in Section 3.1. If one simplifies the model by assuming a single common unknown partition for the whole vector of regression coefficients, a prior model such as a multivariate normal can be assigned for each $\boldsymbol{\beta}_{(j)}$ independently. For this single partition case, in addition to our prior model, one may also consider the spanning tree partitioning prior proposed in Teixeira et al. (2015, 2019).

Another example is to consider a locally stationary Gaussian process model, in a similar spirit of the treed Gaussian process approach (Gramacy and Lee, 2008; Konomi et al., 2014). Conditional on the partition, data within each cluster is modeled as a stationary Gaussian process with latent cluster-specific covariance parameters ϕ_j and a global nugget effect τ^2, that is,

$$y_i = \mu_j + \omega_i^{(j)} + \epsilon_i, \text{ for } i \in \mathcal{D}_j \tag{10}$$

where $\omega_i^{(j)}$ is modeled as a zero mean Gaussian process with covariance function $C(\cdot; \phi_j)$, and $\mathcal{D}_j$ is a subregion in the input space such that the nearest observed location from any input point within $\mathcal{D}_j$ belongs to $\mathcal{C}_j$. Given a partition, $[\{\mu_j, \phi_j\}, \tau^2]$ are assigned with prior models following the typical Bayesian stationary Gaussian process conventions (Banerjee et al., 2014).

The RJ-MCMC algorithm presented in Section 3.4 can be adapted to sample the partitions and other parameters of the above models from their posterior distributions

$$p\left[\{\pi, k, \mathbf{w}\}, \{\boldsymbol{\theta}_{(j)}\}_{j=1:k}, \boldsymbol{\eta} \mid \mathbf{y}\right] \propto \left\{\prod_{j=1}^{k} f(\mathbf{y}_{c_j} \mid \boldsymbol{\theta}_{(j)}, \boldsymbol{\eta}, \pi)\right\} p(\{\boldsymbol{\theta}_{(j)}\}_{j=1:k} \mid \pi) p(\pi, k, \mathbf{w}) p(\boldsymbol{\eta}) \tag{11}$$

We remark that, in the Gaussian regression model, we marginalize out local cluster-specific parameters when sampling partitions to speed up mixing. But in the general case, the collapsed likelihood function may not be achievable. Nevertheless, in the birth, death and change moves in the RJ-MCMC algorithm, the calculation of the likelihood ratio can still be simplified since it only involves a subset of data that have changes in cluster memberships. Data augmentation tricks such as Albert and Chib (1993) for probit models and Polson et al. (2013) for logistic regression can also be applied to derive MCMC algorithms.

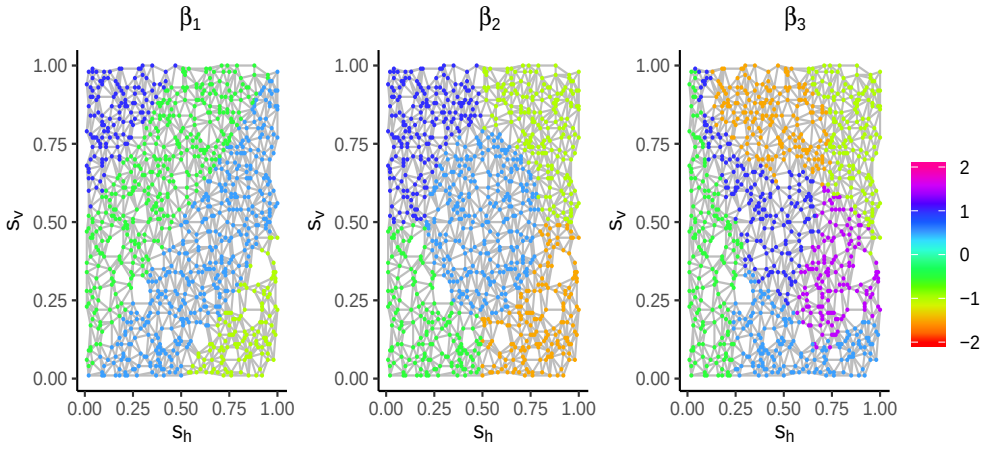

Figure 2: Spatial structures of true coefficients and the Delaunay triangulation used in BSCC.

5. Simulation Studies

5.1 Simulation Setup

In this section, we assess the performance of the BSCC method by some simulation studies. For the ease of comparison with SCC, we use the same simulation setting as in Li and Sang (2019). 1000 spatial locations are generated uniformly in a square domain $[0,1]\times[0,1]$. We generate responses at each location from a linear model with an intercept term and two covariates

$$y(\mathbf{s}_i) = x_1(\mathbf{s}_i)\beta_1(\mathbf{s}_i) + x_2(\mathbf{s}_i)\beta_2(\mathbf{s}_i) + \beta_3(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), \quad \epsilon(\mathbf{s}_i) \overset{i.i.d.}{\sim} \mathrm{N}(0,\sigma^2). \tag{12}$$

We set the true coefficients to be constant within each cluster, the true value of σ to be 0.1, and the numbers of clusters to be 4 for β_1, 5 for β_2 and 6 for β_3, respectively. We consider different clustering patterns for each coefficient, which are shown in Figure 2. In particular, the shapes of true clusters for β_3 are designed to be highly irregular, with the goal of examining the capacity of the BSCC to capture irregular cluster boundaries.

The two covariates are generated such that there is a spatial correlation among locations. Since in practice many spatial covariates are correlated with each other, we also introduce linear dependence between $x_1(\mathbf{s}_i)$ and $x_2(\mathbf{s}_i)$. Specifically, let $\{\zeta_1(\mathbf{s}_i)\}$ and $\{\zeta_2(\mathbf{s}_i)\}$ be two independent realizations of a spatial Gaussian process with zero mean and an isotropic exponential covariance function given by $\mathrm{cov}\{\zeta_m(\mathbf{s}_i),\zeta_m(\mathbf{s}_j)\} = \exp(-\|\mathbf{s}_i-\mathbf{s}_j\|/\phi), m = 1,2$, where ϕ is the range parameter controlling the strength of spatial correlation. Then $x_1(\mathbf{s}_i)$ and $x_2(\mathbf{s}_i)$ are obtained by a linear transformation given by $x_1(\mathbf{s}_i) = \zeta_1(\mathbf{s}_i), \quad x_2(\mathbf{s}_i) = r\zeta_1(\mathbf{s}_i) + \sqrt{1-r^2}\zeta_2(\mathbf{s}_i)$. We consider a moderate collinearity case by setting $r = 0.75$. For spatial correlation within each covariate, three cases are considered, namely, a weak correlation with $\phi = 0.1$, a moderate correlation with $\phi = 0.3$, and a strong correlation with $\phi = 1$. For each value of ϕ, the simulations are repeated 100 times with a same set of true values of coefficients.

We construct the initial graph using the Delaunay triangulation, removing edges longer than 0.1. We consider four candidates $\alpha = 0.0075, 0.0150, 0.1000, 0.3333$, which give $c = 0.05, 0.1, 0.5, 0.9$, respectively. The other hyperparameters are set to be $a_0 = b_0 = 1$ and $c_0 = d_0 = 10^{-6}$, and the standard deviation for the random walk proposal in the hyper step of our RJ-MCMC algorithm is chosen to be 0.9. For each simulated data set, we run $d = 8$ tempered chains in parallel with the lowest inverse temperature $t_d = 0.35$. We run each chain for $100,000$ iterations, discarding the first $50,000$. We set the thinning interval to be 20 iterations and the swap interval to be 100. A total of $2,500$ posterior samples are collected.

As is common in many Bayesian partition models (e.g., Denison et al., 1998; Gramacy and Lee, 2008; Payne et al., 2020), we use the maximum a posteriori (MAP) estimator for point estimation. The posterior distribution used here is the full $p\left[\boldsymbol{\beta}, \{\pi^{(m)}, k_m, \mathbf{w}^{(m)}\}_{m=1}^{p}, \sigma^2, \lambda \mid \mathbf{y}\right]$ derived from (5) (instead of the collapsed version in Equation 7). We also calculate the 95% highest posterior density (HPD) interval for each $\beta_m(\mathbf{s}_i)$ from the MCMC samples.

Most existing software for spatial clustering is designed for spatial response data or spatial points. The BSCC method is compared with the frequentist SCC method (Li and Sang, 2019) and a Dirichlet process mixture (DPM) model for spatial regressions proposed by Ma et al. (2020), due to the lack of other available software for multiple regressions with spatially clustered coefficients. In SCC a fixed MST is used and the tuning parameter for penalization is chosen by BIC. The original DPM model in Ma et al. (2020) includes a term for spatial random effects modeled by a Gaussian process. For fair comparison, we drop this term since our model doesn't include these smoothly varying effects (the results of the original version of DPM models are included in Appendix C.4). The DPM model is essentially a Bayesian linear varying coefficient model with a Dirichlet process prior on the coefficients to capture cluster patterns. Inference of the DPM model is based on MCMC, and we run the chain for $20,000$ iterations, discard the first half, and collect posterior samples every 10 iterations from the second half. MAP estimators are also used for the DPM model. The performance of coefficient estimation is quantified by the mean squared error (MSE)

$$MSE_\beta = \frac{1}{np} \sum_{i=1}^{n} \sum_{m=1}^{p} \{\hat{\beta}_m(\mathbf{s}_i) - \beta_m(\mathbf{s}_i)\}^2.$$

We assess the performance of partition recovery by the Rand index, which is the proportion of agreements of the estimated partitions and the true ones. A Rand index that is closer to 1 indicates a better recovery of the true partition.

We implement the BSCC method in `R` using the `deldir` package for the Delaunay triangulation, the `igraph` package for graph operations, and the `ramcmc` package for the Cholesky update/downdate. The code will be made publicly available upon publication. The implementation of the SCC method is adapted from the `R` package `glmnet`. The DPM model is implemented in `R` using the `nimble` code provided in Ma et al. (2020). All computations were performed on a Linux server with two 2.4GHz 14-core processors and 64GB of memory.

	$\alpha = 0.0075$ ($c = 0.05$)	$\alpha = 0.0150$ ($c = 0.10$)	$\alpha = 0.1000$ ($c = 0.50$)	$\alpha = 0.3333$ ($c = 0.90$)
WAIC_1	49	37	13	1
WAIC_2	53	38	8	1

Table 1: Number of data sets (out of 100) with moderate spatial correlation in which WAIC prefers a certain value of α

	Rand index								
	β_1			β_2			β_3		
Spatial correlation	BSCC	SCC	DPM	BSCC	SCC	DPM	BSCC	SCC	DPM
Weak	0.986	0.716	0.686	0.990	0.819	0.781	0.997	0.852	0.822
Moderate	0.983	0.722	0.681	0.987	0.825	0.773	0.994	0.853	0.812
Strong	0.964	0.726	0.680	0.972	0.830	0.770	0.970	0.849	0.809

Table 2: The average Rand indices for BSCC, SCC, and DPM methods over 100 simulations

5.2 Simulation Results

We first consider selecting the hyperparameter α (or equivalently, c) using WAIC. Table 1 shows the number of data sets with moderate spatial correlation in which WAIC prefers each candidate value of α. The value $\alpha = 0.0075$, which leads to $c = 0.05$, is preferred in most of the data sets by both criteria. As a result, the rest results of the simulation studies are all based on $c = 0.05$. The sensitivity analysis of α is shown in Appendix C.1.

We then assess the performance of BSCC based on 100 repeated experiments. The boxplots of MSEs of BSCC, SCC, and DPM under three different settings of spatial correlation for predictors are shown in Figure 3. We can see that as the spatial correlation for predictors increases, all methods give higher MSEs. Under all settings, the MSE of BSCC is substantially lower than those of SCC and DPM. For instance, when the spatial range parameter of predictors is $\phi = 0.3$ (moderate correlation), the average MSE of BSCC is nearly 1/6 and 1/35 of the counterparts of SCC and DPM, respectively. Even when the spatial correlation is strong ($\phi = 1$), a less favorable case for parameter estimation, BSCC still provides a much more accurate coefficient estimation than SCC and DPM.

In terms of the performance in partition recovery, we compare the average Rand indices of BSCC, SCC, and DPM, over 100 simulations under each setting of spatial correlation. The results are presented in Table 2. BSCC considerably outperforms SCC and DPM in estimating the cluster patterns. Under weak or moderate spatial correlation, BSCC almost perfectly recovers the true partition, suggested by the high Rand indices close to 1. When the covariates are strongly correlated over the spatial domain, the Rand index of BSCC degenerates slightly, but overall still indicates remarkably accurate partition recovery.

Next, we analyze the result from one simulated data set under the setting with a moderate spatial correlation ($\phi = 0.3$) in covariates. The data set that has a median MSE among 100 data sets is chosen for illustration.

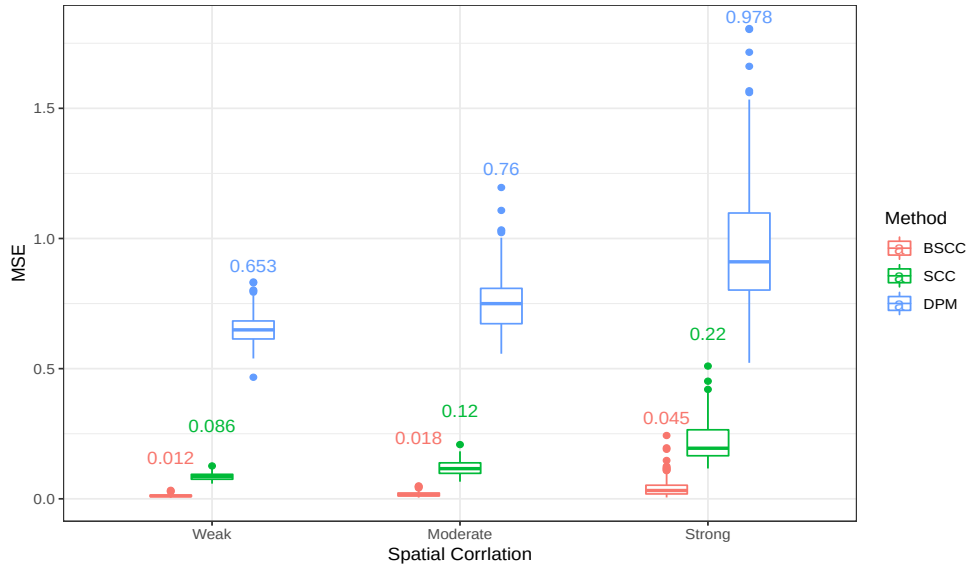

Figure 3: Boxplots of MSEs for BSCC, SCC, and DPM methods under 3 different settings of spatial correlation for predictors. 100 simulations are run for each setting. The average MSE_β over 100 simulations is shown above each box.

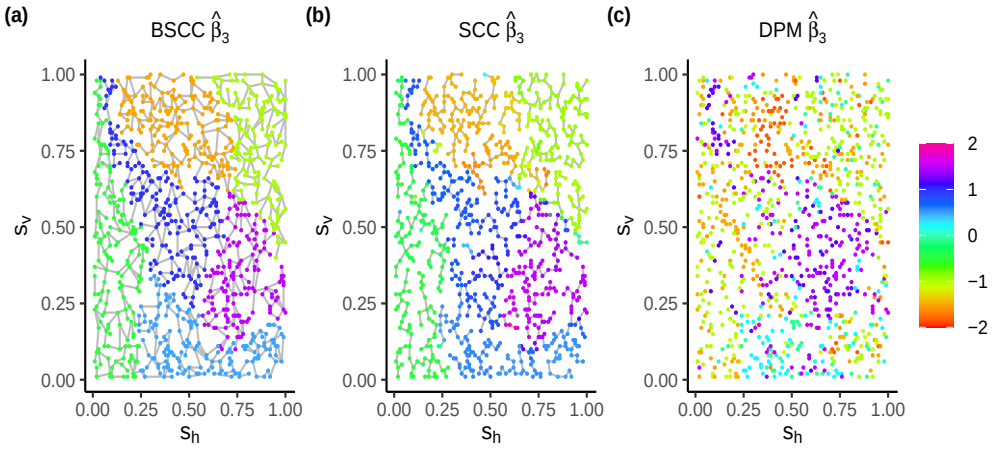

Figure 4: The estimated $\hat{\beta}_3(\mathbf{s}_i)$ from (a) BSCC, (b) SCC, and (c) DPM in one simulated data set with moderate spatially correlated predictors ($\phi = 0.3$). The MAP estimate of the spanning tree is shown in (a), and the minimum spanning tree used by SCC is shown in (b). Points with absolute values greater than 2 are marked in gray.

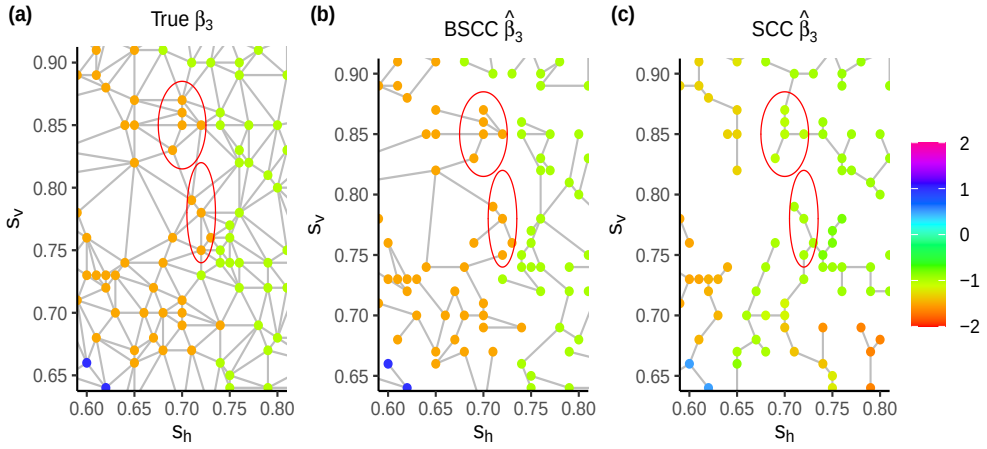

Figure 5: Zoomed version of Figure 2(c) and Figure 4(a, b) into the region $[0.6, 0.8] \times [0.65, 0.9]$. Some of the points mis-classified by SCC but correctly classified by BSCC are marked by red circles.

Figure 4 shows the estimated $\hat{\beta}_3(\mathbf{s}_i)$ from BSCC, SCC, and DPM. While all methods can approximately capture the true patterns shown in Figure 2(c), BSCC gives a much more consistent result in terms of both partition recovery and parameter estimation. In contrast, the result from SCC has more mis-classified points and gives larger estimation errors. The result from DPM is noisier, and the clusters it identifies are not spatially contiguous. The results for $\hat{\beta}_1(\mathbf{s}_i)$ and $\hat{\beta}_2(\mathbf{s}_i)$ are similar and thus omitted. The numbers of clusters given by BSCC are 5 for β_1, 5 for β_2, and 6 for β_3, while the ones given by SCC are 92, 69, and 132, respectively. DPM results in 23 clusters for each coefficient. The results suggest that BSCC can recover the true partitions in a highly accurate way, including the irregularly shaped partition of β_3.

The improvement of BSCC over SCC is largely attributed to the fact that BSCC allows the spanning tree to be updated so that it has a consistent ordering with the true partitions. To illustrate, we show an example in Figure 5, which is a zoom-in version of Figure 2(c) and Figure 4(a, b) on the selected window $[0.6, 0.8] \times [0.65, 0.9]$. The points within the red circles are mis-classified by SCC but correctly classified by BSCC. The reason is that the MST in Panel (c) used in SCC is not able to induce the true partition; the mis-classified points are only connected to the neighboring cluster (marked by green points) instead of the true cluster (marked by orange ones), as they should be. As a result, there is no hope for SCC to recover the true partition due to the use of an inconsistent fixed ordering spanning tree. In contrast, the MCMC procedure in BSCC can fix this issue by updating the spanning tree such that it connects points in a more desirable way, as is shown in Figure 5(b).

Another advantage of BSCC over SCC is that the Bayesian inference procedure naturally comes with an uncertainty measure. Distributions of posterior samples of β_2 at four representative locations are shown in Figure 6, where 95% HPD intervals are marked by red segments. For a location in the interior of a cluster (i.e., far away from the true boundaries), which is shown in Panel (a), the posterior distribution is unimodal, and the HPD interval

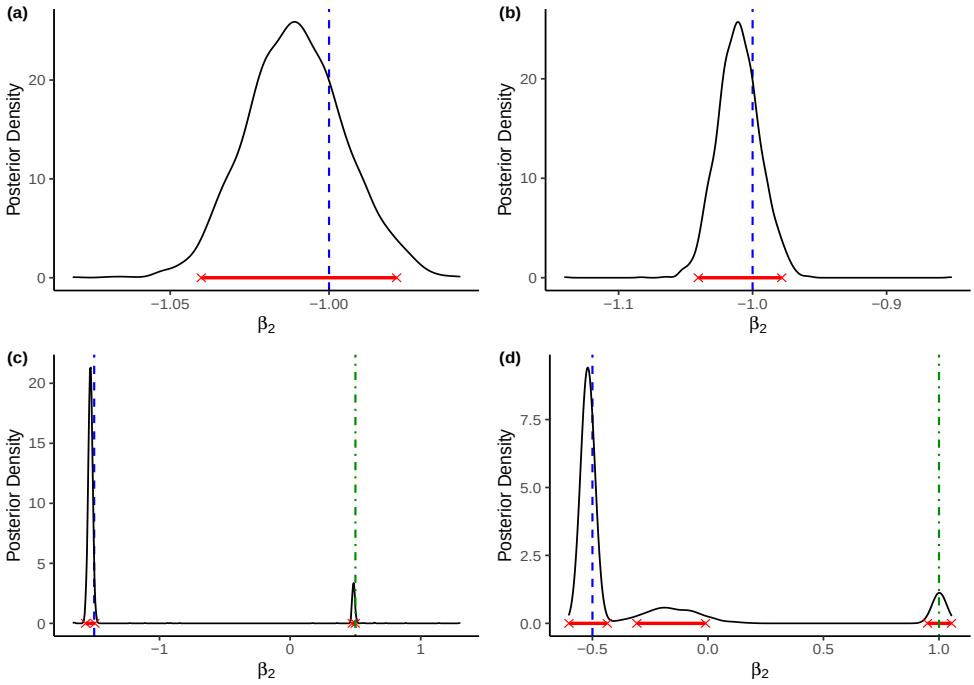

Figure 6: Distributions of posterior samples of β_2 at four locations (see the text for details). Red segments indicate 95% HPD intervals. True coefficient values are marked by blue dashed lines and true values of β_2 in neighboring clusters are marked by green dash-dotted lines. Note the scales of horizontal axes are different.

is narrow and covers its true coefficient (marked by the blue dashed line). The parameter estimation is accurate in this situation. Panels (b - d) show locations close to a true boundary of β_2. The posterior distribution in Panel (b) displays a similar pattern as Panel (a). A different pattern is shown in Panels (c) and (d), where the distributions are multimodal and have wider HPD intervals. Notice that lower modes in Panels (c) and (d) appear near the true values of β_2 in the neighboring clusters (indicated by the green dash-dotted line), and the HDP intervals also contain these values. In Panel (d) there is also a third mode between -0.5 and 0, probably because this location is assigned to some small-sized clusters in some of the MCMC samples. Overall, the posterior distributions assign a substantial amount of mass around the true coefficients. The multimodality reflects the uncertainty that a point near a boundary may be classified into either cluster around it. Posterior distributions of other locations display similar patterns.

Finally, we remark that the computational expense of BSCC is in general reasonable, thanks to the use of multiple computation strategies carefully designed for the collapsed RJ-MCMC algorithm in Section 3.4. With a moderate spatial correlation for covariates, the average time over 100 simulations to run $100,000$ iterations with 8 parallel chains is 20 minutes. As a comparison, DPM takes 56.3 minutes to finish $20,000$ MCMC iterations on average. Increasing spatial correlation has no impact on the running time.

6. Real Data Analysis

6.1 Data Set

We apply our BSCC method to analyze the temperature-salinity (T-S) relationship of seawater in the Atlantic Ocean. Our goal is to identify the Antarctic Intermediate Water (AAIW) characterized by a negative T-S relationship (Talley, 2011). The identification of the AAIW could provide valuable information about Earth's climate change and thus is an important research question in geoscience. It is known that the T-S relationship is relatively homogeneous within certain regions but could change abruptly across the borders of individual water masses. Therefore, the T-S relationship is often assumed to be a spatially piecewise constant in oceanography.

The data of temperature and salinity is downloaded from National Oceanographic Data Center (`https://www.nodc.noaa.gov/OC5/woa13/`). We chose a random sample of $5,130$ spatial locations from the observations in the segment of the Atlantic basin along 25°W between 60°S and the equator. The distributions of both temperature and salinity have strong anisotropic spatial patterns as a result of the Ocean's geometry, which has a width of around 20,000 km and a thickness of about 4 km. To eliminate the anisotropy, we follow a rescaling method commonly used in oceanic studies (Vallis, 2017) by letting $(s_h, s_v) = (s_h^0/L, s_v^0/H)$, where s_h^0 (s_v^0) is the original latitude (depth) and L (H) is the horizontal (vertical) length of the ocean.

6.2 Analysis Results

The relationship of temperature and salinity is modeled by

$$Sal(\mathbf{s}_i) = \beta_0(\mathbf{s}_i) + \beta_1(\mathbf{s}_i)Temp(\mathbf{s}_i) + \epsilon(\mathbf{s}_i),$$

where $Sal(\mathbf{s}_i)$ and $Temp(\mathbf{s}_i)$ are the salinity and temperature at location $\mathbf{s}_i = (s_{h,i}, s_{v,i})$, respectively, $\beta_0(\mathbf{s}_i)$ is the intercept, and $\beta_1(\mathbf{s}_i)$ denotes the T-S relationship of interest. Both β_1 and β_0 are assumed to be spatially piecewise constant. We adopt the same prior as the simulation studies described in Section 5.1 except that we only consider a candidate set of $\alpha \in \{0.0075, 0.015, 0.1\}$ due to computational expense. The optimal model selected by WAIC corresponds to $\alpha = 0.1$, which gives $c = 0.574$. We run $d = 20$ chains with lowest temperature $t_d = 0.1$. Each chain is run for $1,500,000$ iterations with the first $1,000,000$ as burn-in period. The swap and the thinning intervals are set to be 100 and 50, respectively, giving $10,000$ posterior samples in total. Typically such a long chain is needed for large data sets in Bayesian high dimensional regression models to get reliable uncertainty estimates (e.g., Zhou and Guan, 2019; Guan and Stephens, 2011).

The traceplot of posterior samples of σ^2 displays satisfactory convergence and mixing performance. The slope estimates from BSCC as well as SCC are shown in Figure 7, and the estimated boundaries of the AAIW regions (points with negative slope estimates) are marked by black dashed lines in Figure 7. BSCC gives 68 clusters for the slope β_1. In contrast, SCC gives 1141, which is too large for interpretation. A band-shaped AAIW region located near the sea surface from $s_h = -0.30$ to $s_h = -0.50$ is identified by BSCC. Its encompassing region covers the well-recognized generation site of AAIW and the low-salinity tongue which is believed to be associated with AAIW (Talley, 2011). We also notice that BSCC gives a spatially contiguous region of AAIW, while SCC does not.

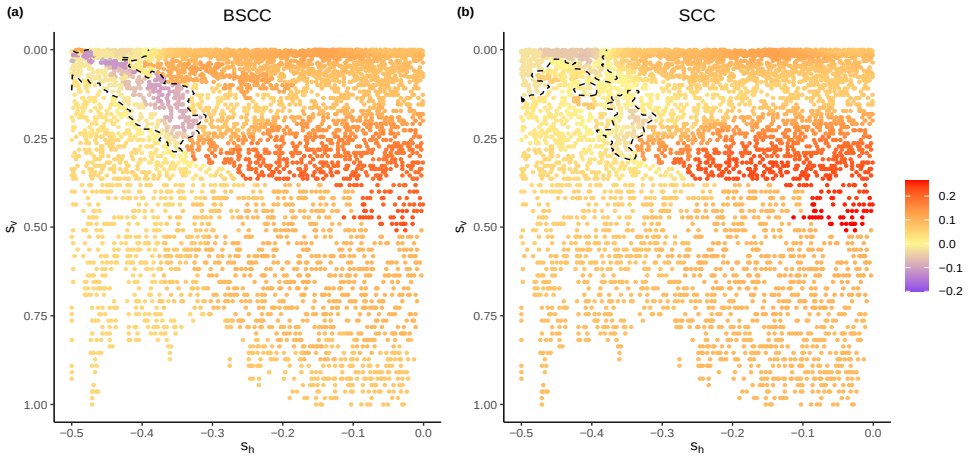

Figure 7: The T-S relationship β_1 estimated from (a) BSCC and (b) SCC. The contour of $\beta_1 = 0$ given by interpolation is shown as the black dashed line.

As suggested by geophysical theory, the T-S relationship may change dramatically across the boundary of AAIW (Talley, 2011). We quantify the change of the estimated T-S relationship by the magnitude of spatial difference quotient (Simmonds, 2012), which is given by

$$D(\mathbf{s}_i) = \left[\frac{\{\beta_1(\mathbf{s}_i) - \beta_1(\mathbf{s}_{i_1})\}^2}{d_1^2 \sin^2\gamma} + \frac{\{\beta_1(\mathbf{s}_i) - \beta_1(\mathbf{s}_{i_2})\}^2}{d_2^2 \sin^2\gamma} - \frac{2\{\beta_1(\mathbf{s}_i) - \beta_1(\mathbf{s}_{i_1})\}\{\beta_1(\mathbf{s}_i) - \beta_1(\mathbf{s}_{i_2})\}\cos\gamma}{d_1 d_2 \sin^2\gamma} \right]^{\frac{1}{2}},$$

where $\mathbf{s}_{i_1}$ and $\mathbf{s}_{i_2}$ are two nearest location of $\mathbf{s}_i$, d_j is the distance between $\mathbf{s}_i$ and $\mathbf{s}_{i_j}$, $j = 1, 2$, and γ is the angle between vectors $(s_{h,i_1} - s_{h,i}, s_{v,i_1} - s_{v,i})$ and $(s_{h,i_2} - s_{h,i}, s_{v,i_2} - s_{v,i})$. Figure 8 shows the results from BSCC and SCC. Consistent with the theoretical results in geophysics, the change of β_1 given by BSCC is abrupt around the boundary. For the results from SCC, the change has much smaller magnitude, partly due to the shrinkage effect of the L_1 penalty on the differences between neighboring regression coefficients.

Finally we illustrate the uncertainty of the T-S relationship estimation in Figure 9. The T-S relationship of purple points are estimated to be negative with high certainty. We find 3 locations along the boundary of the AAIW region whose 95% HPD intervals of β_1 include 0, and they can be viewed as part of the potential boundary of AAIW.

7. Conclusions and Discussion

In this article, a novel spatial regression method, called Bayesian Spatially Clustered Coefficient regression, is developed to estimate the clustered relationship among spatial variables. Our BSCC method is based on a model-based spatially contiguous clustering method defined via connected components of an undirected graph, which we prove can be induced by a spanning tree and a suitable subset of its edge set. A prior for spatial partitions is therefore developed hierarchically by assigning priors to spanning trees as well as their edge sets. We prove that the BSCC model achieves posterior consistency for point estimation

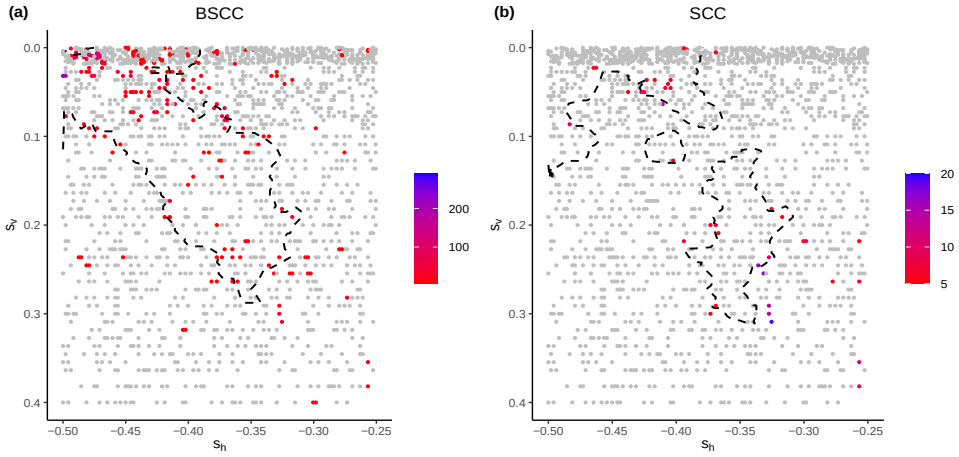

Figure 8: The magnitude of spatial difference quotient of the T-S relationship estimated by (a) BSCC and (b) SCC. Note the color scales in two panels are different. Points with magnitudes less than 5 are marked in gray.

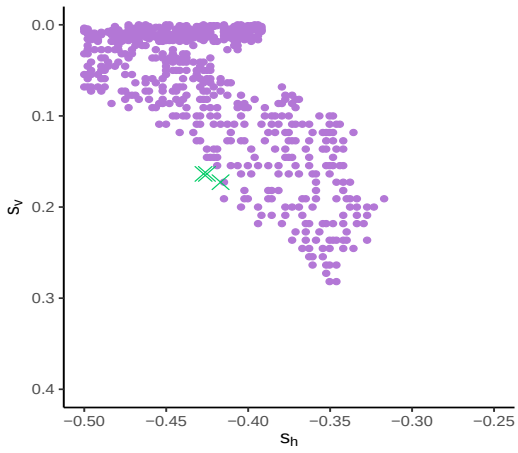

Figure 9: Potential AAIW regions estimated from BSCC. Points with negative $\hat{\beta}_1$ from MAP estimation are shown in purple. Locations where 95% HPD intervals of β_1 include 0 are marked by green crosses. Note that only the region $[-0.5, -0.25] \times [0, 0.4]$ is shown.

under this prior. However, results for posterior selection consistency (i.e., the property that the posterior distribution of partitions concentrates at the true partition) are non-trivial to prove, and we leave this for future research.

For computation, we propose an RJ-MCMC algorithm to sample spanning trees and partitions from their posterior distributions. Various computation methods such as parallel tempering are utilized to facilitate convergence. Our simulation studies demonstrate that BSCC remarkably outperforms its competitors SCC and DPM. In particular, BSCC achieves

nearly 80% reduction in MSE in our simulation studies when compared with its frequentist counterpart, SCC, partially for the reason that the MCMC procedure can effectively fix the mis-classification in SCC by proposing a more desired spanning tree. We also present an application of BSCC to the detection of water masses by estimating the spatial clustering patterns of T-S relationship in the Atlantic basin.

One potential research direction is to further improve the convergence and mixing of the BSCC algorithm. A long burn-in period is typically needed before the chain converges for our simulated and real data. For binary tree based methods, efficient proposals for new partitions have been well-studied in literature (Chipman et al., 1998, 2010; Wu et al., 2007). For the proposed spanning tree based model, we have tried to propose new partitions adaptively by splitting an existing cluster at boundaries. However, we did not observe substantial improvement in terms of mixing and convergence (see Appendix D.2 for details). Modifications of proposals in the current RJ-MCMC algorithm are currently under investigation. Nevertheless, we remark that based on our numerical experiments, even when the chain does not fully converge, one can often still get reasonably accurate point estimations of partitions and coefficient values, though the reliability of uncertainty measures such as HPD intervals and Bayesian model averaging might be a concern. Hyperparameter selection is another remaining challenge in the model. Despite the utility of the proposed hyperparameter selection method in Section 3.5, a careful choice of the candidate set for α is still required to achieve better performance when one has little information about the number of clusters *a priori*.

Our current model in (1) assumes that the intercept and other regression coefficients are spatially piecewise constant. It is straightforward to generalize (1) to be $y(\mathbf{s}_i) = \mathbf{x}_1(\mathbf{s}_i)^{\mathrm{T}}\boldsymbol{\beta}(\mathbf{s}_i) + \mathbf{x}_2(\mathbf{s}_i)^{\mathrm{T}}\boldsymbol{\alpha}(\mathbf{s}_i) + \epsilon(\mathbf{s}_i)$, where $\boldsymbol{\beta}$ has clustering patterns and $\boldsymbol{\alpha}$ is smoothly varying. Incorporating a spatial Gaussian process random effect into the BSCC model is a special case of it.

Acknowledgments

The research of Huiyan Sang was partially supported by NSF grant no. NSF DMS-1854655. The research of Bani Mallick was partially supported by NSF grant no. NSF CCF-1934904 and national cancer Institute of the National Institutes of Health grant under award number R01CA194391. The authors thank the referees and the editor for valuable comments. The authors also thank Drs. Renato Assunção, Zhengyuan Zhu, Furong Li and Guanyu Hu for several useful discussions.

Appendix A. Proofs

A.1 Proof of Proposition 2

To prove Proposition 2, we first introduce a lemma.

Lemma 8 *(Proposition 8.1.1 of Diestel 2016) Every connected graph contains at least one spanning tree.*

Now we prove Proposition 2.

Proof [of Proposition 2] We first construct a subgraph of $\mathcal{G}_0$ and then show that it is a spanning tree that induces π. Consider the following procedure with initial values $t = 1$ and $\mathcal{T}^0 = (\mathcal{T}^0, \mathcal{E}^0) = (\emptyset, \emptyset)$:

1. If $t = 1$, pick an arbitrary vertex $v \in \mathcal{V}_0$; otherwise, pick a vertex $v \in \mathcal{V}_0 \backslash \mathcal{V}^{t-1}$ that is connected to a vertex in $\mathcal{T}^{t-1}$ by an edge e (the existence of v is guaranteed since $\mathcal{G}_0$ is connected). Without loss of generality suppose v belongs to $\mathcal{C}_t$.

2. By Lemma 8 we know there is a spanning tree $\mathcal{T}^* = (\mathcal{V}^*, \mathcal{E}^*)$ of the subgraph $(\mathcal{C}_t, \mathcal{E}_{\mathcal{C}_t})$, where $\mathcal{E}_{\mathcal{C}_t} \subseteq \mathcal{E}_0$ is the set of edges whose endpoints belong to $\mathcal{C}_t$. If $t = 1$ let $\mathcal{T}^t = \mathcal{T}^*$; otherwise, let $\mathcal{T}^t = (\mathcal{V}^{t-1} \cup \mathcal{V}^*, \mathcal{E}^{t-1} \cup \mathcal{E}^* \cup \{e\})$, where $\mathcal{V}^{t-1}$ and $\mathcal{E}^{t-1}$ are the vertex set and edge set of $\mathcal{T}^{t-1}$, respectively.

3. If $\mathcal{T}^t$ contains all vertices in $\mathcal{G}_0$, then stop; otherwise, let $t := t + 1$ and go to step 1.

We show that each $\mathcal{T}^t, t \geq 1$ is a tree by induction arguments. By construction $\mathcal{T}^1$ is a tree. Suppose $\mathcal{T}^{t-1}$ is a tree, then $\mathcal{T}^t$ is also a tree since both $\mathcal{T}^{t-1}$ and $\mathcal{T}^*$ are trees.

Therefore, the final $\mathcal{T}^t$ that contains all vertices of $\mathcal{G}_0$ is a spanning tree and the collection of e's in each iteration is $\mathcal{E}_{k-1}$. This completes the proof of Proposition 2. ■

A.2 Proof of Theorem 3

To prove Theorem 3 we need some lemmas.

Lemma 9 *(Lemma 1 of Laurent and Massart 2000) Let χ^2_d be a chi-square distribution with degree of freedom d. Then the following concentration inequalities hold for any $x > 0$:*

$$\text{pr}\left(\chi^2_d > d + 2x + 2\sqrt{dx}\right) \leq \exp(-x)$$

and

$$\text{pr}\left(\chi^2_d < d - 2\sqrt{dx}\right) \leq \exp(-x).$$

Lemma 10 *(Lemma 6 of Barron 1998) Let f_θ be the likelihood function with parameter $\theta \in \Theta_n$, $f^* \equiv f_{\theta^*}$ be the true probability density of data generation with true data generation parameter θ^*, E_θ, E^* denote the expectations under θ and θ^* respectively,* pr* *denote the probability measure for data generation under θ^*, and Π, Π_n denote the prior distribution on Θ_n with density $\pi(\theta)$ and the posterior, respectively. Let B_n and C_n be two subsets of*

the parameter space Θ_n, and ϕ_n be a test function satisfying $\phi_n(D_n) \in \{0,1\}$ for any data D_n. If $\Pi(B_n) \le b_n, E^*\{\phi(D_n)\} \le b_n', \sup_{\theta \in C_n} E_\theta\{1-\phi(D_n)\} \le c_n$, and

$$\mathrm{pr}^*\left(\frac{m(D_n)}{f^*(D_n)} \ge a_n\right) \ge 1 - a_n'$$

where $m(D_n) = \int_{\Theta_n} \pi(\theta) f_\theta(D_n)\, d\theta$ is the marginal likelihood of D_n. Then for any $\Delta_n > 0$,

$$\mathrm{pr}^*\left(\Pi_n(C_n \cup B_n \mid D_n) \ge \frac{b_n + c_n}{a_n \Delta_n}\right) \le \Delta_n + a_n' + b_n'.$$

Next we give the proof of Theorem 3. With some abuse of notations we use $i \in \mathcal{C}_j$ to denote that the ith location belongs to the jth cluster and (i,j) to denote the edge connecting $\mathbf{s}_i$ and $\mathbf{s}_j$ throughout the proof. We also denote the L_1 and supremum norm by $\|\cdot\|_1$ and $\|\cdot\|_\infty$, respectively.

Proof [of Theorem 3]

Given an arbitrary partition π with k clusters, for the jth cluster, we define an estimator as

$$\hat{\beta}_{(j)} = \frac{\sum_{i \in \mathcal{C}_j} x_i y_i}{\sum_{i \in \mathcal{C}_j} x_i^2},$$

where $y_i = y(\mathbf{s}_i)$. Further define $\hat{\beta}_\pi(\mathbf{y}) \in \mathbb{R}^n$ such that the ith element $\hat{\beta}_{\pi,i}(\mathbf{y}) = \hat{\beta}_{(j)}$ if $i \in \mathcal{C}_j$ under π, and $\hat{\sigma}_\pi^2(\mathbf{y}) = \|\mathbf{y} - \hat{\boldsymbol{\mu}}_\pi(\mathbf{y})\|^2/(n-k)$, where $\hat{\boldsymbol{\mu}}_{\pi,i}(\mathbf{y}) = x_i \hat{\beta}_{\pi,i}(\mathbf{y})$.

Step 1: Inspired by Song and Cheng (2020), we define a test function

$$\phi(\mathbf{y}) = \mathbf{1}\{\|\hat{\boldsymbol{\mu}}_\pi(\mathbf{y}) - \boldsymbol{\mu}^*\| \ge \sqrt{n}\sigma^*\varepsilon_n \text{ and } \left|\hat{\sigma}_\pi^2(\mathbf{y}) - \sigma^{*2}\right| > \sigma^{*2}\varepsilon_n \text{ for some } \pi_k \text{ nested in } \pi^* \text{ with } k \le (1+\delta)g_n^*\}$$

for some fixed $\delta > 0$ chosen later. Let $\circ$ denote the Hadamard product of two vectors. We define

$$C_n = \left\{(\boldsymbol{\beta}, \sigma) : \|\mathbf{x} \circ \boldsymbol{\beta} - \boldsymbol{\mu}^*\| \le M_1 \sqrt{n}\sigma^*\varepsilon_n \text{ and } \frac{1-\varepsilon_n}{1+\varepsilon_n} < \sigma^2/\sigma^{*2} < \frac{1+\varepsilon_n}{1-\varepsilon_n}\right\}^c \setminus B_n,$$

and

$$B_n = \left\{(\boldsymbol{\beta}, \sigma) : \text{The partition underlying } \boldsymbol{\beta} \text{ has at least } \delta g_n^* \text{ clusters }\right\}.$$

For any π_k nested in π^* with $k \le (1+\delta)g_n^*$ and the jth cluster $\mathcal{C}_j$ in π_k, we have $\hat{\beta}_{(j)} \sim \mathrm{N}\left(\beta_{(j)}^*, \sigma^{*2}/\sum_{i \in \mathcal{C}_j} x_i^2\right)$, where $\beta_{(j)}^*$ is the true coefficient in $\mathcal{C}_j$, and thus $\sum_{i \in \mathcal{C}_j}(x_i\hat{\beta}_{(j)} - x_i\beta_{(j)}^*)^2 \sim \sigma^{*2}\chi_1^2$. Hence, $\|\hat{\boldsymbol{\mu}}_\pi(\mathbf{y}) - \boldsymbol{\mu}^*\|^2/\sigma^{*2} \sim \chi_k^2$.

We now bound the type-I error of the test function. Since $k = O(g_n^*) \prec n\varepsilon_n^2$ by Assumption (C3) and $\varepsilon_n \asymp (g_n^* \log n/n)^{1/2}$, from the concentration inequality for χ^2 distribution in Lemma 9, we have

$$\begin{aligned}&\mathrm{pr}_{(\beta^*,\sigma^*)}\left(\|\hat{\boldsymbol{\mu}}_\pi(\mathbf{y}) - \boldsymbol{\mu}^*\| \ge \sqrt{n}\sigma^*\varepsilon_n, \left|\hat{\sigma}_\pi^2(\mathbf{y}) - \sigma^{*2}\right| > \sigma^{*2}\varepsilon_n\right)\\ \le\ &\mathrm{pr}(\chi_k^2 \ge n\varepsilon_n^2) \le \exp\left(-c_1' n\varepsilon_n^2\right),\end{aligned}$$

for some constant $c_1' > 0$. Therefore, using a union bound and the second part of Assumption (C3),

$$E_{(\beta^*,\sigma^*)}\{\phi(\mathbf{y})\} \leq P_n \cdot \exp\left(-c_1' n\varepsilon_n^2\right) \leq \exp\left(-c_1 n\varepsilon_n^2\right), \tag{13}$$

for some constant $c_1 > 0$ and large $n\varepsilon_n^2/(g_n^* \log n)$.

Next we bound the type-II error. We rewrite

$$C_n = C_n^{(1)} \cup C_n^{(2)}$$

where

$$C_n^{(1)} = \left\{(\boldsymbol{\beta}, \sigma) : \|\mathbf{x} \circ \boldsymbol{\beta} - \boldsymbol{\mu}^*\| > M_1\sqrt{n}\sigma^*\varepsilon_n, \frac{\sigma^2}{\sigma^{*2}} < \frac{1+\varepsilon_n}{1-\varepsilon_n}\right\} \cap B_n^c$$

and

$$C_n^{(2)} = \left\{\sigma : \frac{\sigma^2}{\sigma^{*2}} \leq \frac{1-\varepsilon_n}{1+\varepsilon_n} \text{ or } \frac{\sigma^2}{\sigma^{*2}} \geq \frac{1+\varepsilon_n}{1-\varepsilon_n}\right\} \cap B_n^c.$$

For any $(\boldsymbol{\beta}, \sigma) \in C_n$, let π be the corresponding partition of $\boldsymbol{\beta}$ and $\mathcal{T}$ be a spanning tree inducing π. Define $\hat{\pi}$ to be the partition formed by removing the edges $\left\{(i,j) \in \mathcal{E}_{\mathcal{T}} : |\beta_i - \beta_j| > 0 \text{ or } |\beta_i^* - \beta_j^*| > 0\right\}$ from $\mathcal{T}$. Then $\hat{\pi}$ is nested in both π and π^*, and has no more than $(1+\delta)g_n^*$ clusters (this is due to the construction of B_n^c and g_n^*). For any $\boldsymbol{\beta} \in C_n^{(1)}$, we have

$$\begin{aligned}
&\mathrm{pr}_{(\beta,\sigma)}\left(\|\hat{\boldsymbol{\mu}}_{\hat{\pi}}(\mathbf{y}) - \boldsymbol{\mu}^*\| \leq \sqrt{n}\sigma^*\varepsilon_n\right) \\
&= \mathrm{pr}_{(\beta,\sigma)}\left(\|\hat{\boldsymbol{\mu}}_{\hat{\pi}}(\mathbf{y}) - \mathbf{x} \circ \boldsymbol{\beta}) + \mathbf{x} \circ \boldsymbol{\beta} - \boldsymbol{\mu}^*\| \leq \sqrt{n}\sigma^*\varepsilon_n\right) \\
&\leq \mathrm{pr}_{(\beta,\sigma)}\left(\|\hat{\boldsymbol{\mu}}_{\hat{\pi}}(\mathbf{y}) - \mathbf{x} \circ \boldsymbol{\beta}\| \geq \|\boldsymbol{\mu}^* - \mathbf{x} \circ \boldsymbol{\beta}\| - \sqrt{n}\sigma^*\varepsilon_n\right) \\
&\leq \mathrm{pr}_{(\beta,\sigma)}\left(\|\hat{\boldsymbol{\mu}}_{\hat{\pi}}(\mathbf{y}) - \mathbf{x} \circ \boldsymbol{\beta}\| \geq (M_1 - 1)\sqrt{n}\sigma^*\varepsilon_n\right),
\end{aligned}$$

where the last inequality is due to the fact that when $\boldsymbol{\beta} \in C_n^{(1)}$, $\|\boldsymbol{\mu}^* - \mathbf{x} \circ \boldsymbol{\beta}\| > M_1\sqrt{n}\sigma^*\varepsilon_n$. Note also that within each cluster $\mathcal{C}_j$ under $\hat{\pi}$, $\sum_{i\in\mathcal{C}_j}\left(\hat{\boldsymbol{\mu}}_{\hat{\pi},i}(\mathbf{y}) - x_i\beta_{(j)}\right)^2 = \frac{\left(\sum_{i\in\mathcal{C}_j} x_i\epsilon_i\right)^2}{\sum_{i\in\mathcal{C}_j} x_i^2} \sim \sigma^2\chi_1^2$, where $\beta_{(j)}$ is the value of $\boldsymbol{\beta}$ in $\mathcal{C}_j$, and hence $\|\hat{\boldsymbol{\mu}}_{\hat{\pi}}(\mathbf{y}) - \mathbf{x} \circ \boldsymbol{\beta}\|^2/\sigma^2 \sim \chi_{\hat{k}}^2$ under the true parameters $(\boldsymbol{\beta}, \sigma)$, where $\hat{k}$ is the number of clusters in $\hat{\pi}$. Therefore,

$$\begin{aligned}
\mathrm{pr}_{(\beta,\sigma)}\left(\|\hat{\boldsymbol{\mu}}_{\hat{\pi}}(\mathbf{y}) - \boldsymbol{\mu}^*\| \leq \sqrt{n}\sigma^*\varepsilon_n\right) &\leq \mathrm{pr}\left(\chi_{\hat{k}}^2 \geq \frac{1-\varepsilon_n}{1+\varepsilon_n}(M_1-1)^2 n\varepsilon_n^2\right) \\
&\leq \exp\left(-c_2'(M_1-1)^2 n\varepsilon_n^2\right)
\end{aligned} \tag{14}$$

for large M_1 and some constant $c_2' > 0$.

Now consider $(\boldsymbol{\beta}, \sigma) \in C_n^{(2)}$. Under the true parameters $(\boldsymbol{\beta}, \sigma)$, by the normality of $\mathbf{y}$ we have $\|\mathbf{y} - \hat{\boldsymbol{\mu}}_{\hat{\pi}}(\mathbf{y})\|^2 \sim \sigma^2 \chi^2_{n-\hat{k}}$. Therefore, since $\sigma \in C_n^{(2)}$,

$$\begin{aligned}
&\mathrm{pr}_{(\beta,\sigma)}\left(\left|\hat{\sigma}^2_{\hat{\pi}}(\mathbf{y}) - \sigma^{*2}\right| < \sigma^{*2}\varepsilon_n\right) \\
&= \mathrm{pr}_{(\beta,\sigma)}\left(\left|\frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}_{\hat{\pi}}(\mathbf{y})\|^2}{\sigma^{*2}(n-\hat{k})} - 1\right| < \varepsilon_n\right) \\
&= \mathrm{pr}_{(\beta,\sigma)}\left((1-\varepsilon_n)\frac{\sigma^{*2}}{\sigma^2} < \frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}_{\hat{\pi}}(\mathbf{y})\|^2}{\sigma^2(n-\hat{k})} < (1+\varepsilon_n)\frac{\sigma^{*2}}{\sigma^2}\right) \\
&\leq \mathrm{pr}_{(\beta,\sigma)}\left(\left|\frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}_{\hat{\pi}}(\mathbf{y})\|^2}{\sigma^2} - (n-\hat{k})\right| > (n-\hat{k})\varepsilon_n\right) \\
&\leq \mathrm{pr}\left(\left|\chi^2_{n-\hat{k}} - (n-\hat{k})\right| > (n-\hat{k})\varepsilon_n\right) \\
&\leq \exp\left(-c_2 n\varepsilon_n^2\right), \qquad (15)
\end{aligned}$$

for some constant $c_2 > 0$ and large n.

Combining (14) and (15), we obtain

$$\begin{aligned}
\sup_{(\beta,\sigma)\in C_n} E_{(\beta,\sigma)}\{1-\phi(y)\} &\leq \max\left\{\exp\left(-c_2'(M_1-1)^2 n\varepsilon_n^2\right), \exp\left(-c_2 n\varepsilon_n^2\right)\right\} \\
&\leq \exp\left(-c_2 n\varepsilon_n^2\right), \qquad (16)
\end{aligned}$$

if M_1 is chosen to be large.

Step 2: Let $m(\mathbf{y})$ be the marginal likelihood, $f^*(\mathbf{y})$ be the true likelihood and $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{x} \circ \boldsymbol{\beta}^*$ be the vector of error terms.

We claim that, with probability $\mathrm{pr}\left(\|\boldsymbol{\epsilon}\| \leq 2\sqrt{n}\sigma^*\right)$,

$$\tilde{H}_n := \left\{(\boldsymbol{\beta}, \sigma) : \left\|\frac{1}{\sigma}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\right\|_\infty \leq g_n^* \frac{\log n}{n}, 0 \leq \sigma^2 - \sigma^{*2} \leq \sigma^{*2} g_n^* \frac{\log n}{n}\right\} \subset H_n,$$

where H_n is defined as

$$H_n = \left\{(\boldsymbol{\beta}, \sigma) : \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x} \circ \boldsymbol{\beta}^* - \mathbf{x} \circ \boldsymbol{\beta} + \boldsymbol{\epsilon}\|^2 + \frac{\|\boldsymbol{\epsilon}\|^2}{2\sigma^{*2}} - n\log\frac{\sigma}{\sigma^*}\right) \geq \exp(-c_3' g_n^* \log n)\right\}$$

for some constant $c_3' > 0$. Thus,

$$\begin{aligned}
\frac{m(\mathbf{y})}{f^*(\mathbf{y})} &\geq \int_{H_n} \exp\left(-\frac{1}{\sigma^2}\|\mathbf{x} \circ \boldsymbol{\beta}^* - \mathbf{x} \circ \boldsymbol{\beta} + \boldsymbol{\epsilon}\|^2 + \frac{\|\boldsymbol{\epsilon}\|^2}{\sigma^{*2}} - n\log\frac{\sigma}{\sigma^*}\right) p\left(\boldsymbol{\beta}, \sigma^2\right) d\boldsymbol{\beta} d\sigma^2 \\
&\geq \Pi(H_n) \cdot \exp\left(-c_3' g_n^* \log n\right) \geq \Pi(\tilde{H}_n) \cdot \exp\left(-c_3' g_n^* \log n\right). \qquad (17)
\end{aligned}$$

To see the claim, write

$$\begin{aligned}
&\frac{1}{\sigma^2}\left\|\mathbf{x}\circ\boldsymbol{\beta}^*-\mathbf{x}\circ\boldsymbol{\beta}+\boldsymbol{\epsilon}\right\|^2-\frac{\|\boldsymbol{\epsilon}\|^2}{\sigma^{*2}}+n\log\frac{\sigma}{\sigma^*}\\
=&\underbrace{\frac{\left\|\mathbf{x}\circ\boldsymbol{\beta}^*-\mathbf{x}\circ\boldsymbol{\beta}\right\|^2}{\sigma^2}}_{I}+\underbrace{\frac{2\left(\mathbf{x}\circ\boldsymbol{\beta}^*-\mathbf{x}\circ\boldsymbol{\beta}\right)^{\mathrm{T}}\boldsymbol{\epsilon}}{\sigma^2}}_{II}\underbrace{-\|\boldsymbol{\epsilon}\|^2\left(\frac{1}{\sigma^{*2}}-\frac{1}{\sigma^2}\right)}_{\leq 0 \text{ since } \sigma\geq\sigma^*}+\underbrace{\frac{n}{2}\log\frac{\sigma^2}{\sigma^{*2}}}_{III}.
\end{aligned}$$

Noticing that when $(\boldsymbol{\beta},\sigma)\in\tilde{H}_n$ and $\|\boldsymbol{\epsilon}\|\leq 2\sqrt{n}\sigma^*$, by Assumption (C1) and the first part of Assumption (C3) we have

$$I\leq\frac{M_0^2}{\sigma^2}\left\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\right\|^2\leq nM_0^2\left\|\frac{\boldsymbol{\beta}-\boldsymbol{\beta}^*}{\sigma}\right\|_\infty^2\leq M_0^2\frac{g_n^{*2}(\log n)^2}{n}=O\left(g_n^*\log n\right),$$

by Hölder's inequality and $\sigma^*\leq\sigma$ we have

$$\begin{aligned}
II&\leq 2\left\|\frac{\mathbf{x}\circ\boldsymbol{\beta}^*-\mathbf{x}\circ\boldsymbol{\beta}}{\sigma}\right\|_\infty\cdot\|\boldsymbol{\epsilon}\|_1\cdot\frac{1}{\sigma^*}\leq\frac{2M_0}{\sigma^*}\left\|\frac{\boldsymbol{\beta}^*-\boldsymbol{\beta}}{\sigma}\right\|_\infty\cdot\sqrt{n}\left\|\boldsymbol{\epsilon}\right\|\\
&\leq\frac{2M_0}{\sigma^*}\left\|\frac{\boldsymbol{\beta}^*-\boldsymbol{\beta}}{\sigma}\right\|_\infty\cdot\sqrt{n}\cdot 2\sqrt{n}\sigma^*=O\left(g_n^*\log n\right),
\end{aligned}$$

and

$$III\leq\frac{n}{2}\frac{g_n^*\log n}{n}=O\left(g_n^*\log n\right).$$

The claim then follows.

Next we show the prior assigns sufficient probability mass to $\tilde{H}_n$. Notice that $\Pi(\tilde{H}_n)=\sum_{\mathcal{T}\in\mathbb{T}_n}\Pi(\tilde{H}_n\mid\mathcal{T})\Pi(\mathcal{T})\geq\min_{\mathcal{T}\in\mathbb{T}_n}\Pi(\tilde{H}_n\mid\mathcal{T})$, and for each $\mathcal{T}$, $\Pi(\tilde{H}_n|\mathcal{T})\geq\Pi(\pi^*_{\mathcal{T}}\mid\mathcal{T})\Pi(\tilde{H}_n\mid\pi^*_{\mathcal{T}})$, where $\pi^*_{\mathcal{T}}$ is the partition obtained by removing the edges in $G^*_{\mathcal{T}}$ from $\mathcal{T}$. The number of clusters in $\pi^*_{\mathcal{T}}$, denoted by $k^*_{\mathcal{T}}$, is upper bounded by g_n^*.

First consider $\Pi(\pi^{\mathcal{T}}\mid\mathcal{T})=\Pi(k=k^*_{\mathcal{T}})\binom{n-1}{k^*_{\mathcal{T}}-1}^{-1}$. By Assumption (C4),

$$\begin{aligned}
\log\Pi(k=k^*_{\mathcal{T}})&\geq\log\frac{(1-c)^{g_n^*}}{\sum_{k=1}^n(1-c)^k}\\
&=(g_n^*-1)\log(1-c)+\log c-\log\{1-(1-c)^n\}\\
&\geq-2\alpha g_n^*\log n.
\end{aligned}\tag{18}$$

In addition,

$$-\log\binom{n-1}{g_n^*-1}\geq-g_n^*\log n.\tag{19}$$

Now we consider

$$\begin{aligned}
\Pi(\tilde{H}_n|\pi^*_{\mathcal{T}})=\Pi\Big(&\frac{1}{\sigma}|\beta_{(j)}-\beta^*_{(j)}|\leq\frac{g_n^*\log n}{n}\text{ for }j=1,2,\ldots,k^*_{\mathcal{T}},\\
&0\leq\sigma^2-\sigma^{*2}\leq\sigma^{*2}g_n^*\frac{\log n}{n}\Big).
\end{aligned}$$

Since the prior for $\beta_{(j)}$ is given by

$$\beta_{(j)} \mid \lambda, \sigma \overset{iid}{\sim} \mathrm{N}(0, \lambda^{-1}\sigma^2), \quad \lambda \sim \mathrm{Gamma}(c_0/2, d_0/2),$$

by Assumption (C2) and (C3) we have, conditional on $0 \le \sigma^2 - \sigma^{*2} \le \sigma^{*2} g_n^* \frac{\log n}{n}$,

$$\begin{aligned}
&\Pi\left(\frac{1}{\sigma}|\beta_{(j)} - \beta^*_{(j)}| \le \frac{g_n^* \log n}{n} \text{ for all } j = 1, 2, \ldots, k^*_{\mathcal{T}} \;\middle|\; \sigma\right) \\
&= \int_0^\infty \prod_{j=1}^{k^*_{\mathcal{T}}} \Pi\left(\frac{1}{\sigma}|\beta_{(j)} - \beta^*_{(j)}| \le \frac{g_n^* \log n}{n} \;\middle|\; \lambda, \sigma\right) p(\lambda) d\lambda \\
&\ge \int_0^\infty \left(\frac{g_n^* \log n}{n}\right)^{k^*_{\mathcal{T}}} \left(\frac{\lambda}{2\pi}\right)^{k^*_{\mathcal{T}}/2} \exp\left(-\frac{k^*_{\mathcal{T}}}{2}\lambda Z^2\right) \cdot p(\lambda) d\lambda, \\
&\quad \text{where } Z = \max_{1 \le j \le k^*_{\mathcal{T}}} \frac{|\beta^*_{(j)}|}{\sigma^*} + 1 = \max_{1 \le i \le n} \frac{|\beta^*_i|}{\sigma^*} + 1, \\
&\ge \tilde{c}_3 \cdot \left(\frac{g_n^* \log n}{n}\right)^{g_n^*} \Gamma\left(\frac{k^*_{\mathcal{T}} + c_0}{2}\right) \cdot \left[\frac{1}{2}\{d_0 + g_n^* Z^2\}\right]^{-(g_n^* + c_0)/2}, \\
&\quad \text{where } \tilde{c}_3 \text{ is a constant not involving } n, \\
&\ge \exp\left(-c_3'' g_n^* \log n\right)
\end{aligned} \tag{20}$$

for some constant $c_3'' > 0$ when n is sufficiently large.

Finally, for some constant $c_3''' > 0$,

$$\begin{aligned}
&\Pi\left(0 \le \sigma^2 - \sigma^{*2} \le \sigma^{*2} g_n^* \frac{\log n}{n}\right) \\
&\ge \sigma^{*2} g_n^* \frac{\log n}{n} \cdot \min_{\sigma^2 \in [\sigma^{*2},\, \sigma^{*2}(1 + g_n^* \log n/n)]} p(\sigma^2) \\
&\ge \exp\left(-c_3''' g_n^* \log n\right).
\end{aligned} \tag{21}$$

Combining (18), (19), (20) and (21) we obtain $\Pi(\tilde{H}_n \mid \mathcal{T}) \ge \exp\left(-c_3 g_n^* \log n\right)$ and thus $\Pi(\tilde{H}_n) \ge \exp\left(-c_3 g_n^* \log n\right)$, for some constant $c_3 > 0$ not depending on $\mathcal{T}$. Hence, with probability

$$\mathrm{pr}\left(\|\boldsymbol{\epsilon}\| \le 2\sqrt{n}\sigma^*\right) \ge \mathrm{pr}(\chi^2_n \le 4n) \ge 1 - \exp(-c_4 n), \tag{22}$$

for some constant $c_4 > 0$, we have

$$\frac{m(\mathbf{y})}{f^*(\mathbf{y})} \ge \exp\left(-(c_3 + c_3')g_n^* \log n\right). \tag{23}$$

Step 3: By Assumption (C4), for any $\mathcal{T}$ and some constant $c_5 > 0$ not depending on $\mathcal{T}$,

$$\begin{aligned}\Pi\left(B_n \mid \mathcal{T}\right) &\leq \Pi(k \geq \delta g_n^*) \\ &= \frac{\sum_{k=\delta g_n^*}^{n}(1-c)^k}{\sum_{k=1}^{n}(1-c)^k} \\ &= \frac{(1-c)^{\delta g_n^*-1}\{1-(1-c)^{n-\delta g_n^*+1}\}}{1-(1-c)^n} \\ &= O(1)\cdot(1-c)^{\delta g_n^*-1} \\ &\leq \exp\{-c_5\delta g_n^*\log n\}.\end{aligned}$$

We therefore have

$$\Pi\left(B_n\right) = \sum_{\mathcal{T}\in\mathbb{T}_n}\Pi\left(B_n \mid \mathcal{T}\right)\Pi(\mathcal{T}) \leq \exp\{-c_5\delta g_n^*\log n\}. \tag{24}$$

Combining parts: By Lemma 10, (13), (16), (23), (24) and (22), it follows that for sufficiently large δ and $n\varepsilon_n^2/(g_n^*\log n)$,

$$\begin{aligned}&\mathrm{pr}^*\left\{\Pi_n\left(\frac{1}{\sqrt{n}}\left\|\boldsymbol{\mu}-\boldsymbol{\mu}^*\right\| \geq M_1\sigma^*\varepsilon_n \mid \mathbf{y}\right) \geq \rho_n\right\} \\ &\leq \mathrm{pr}^*\left\{\Pi_n\left(C_n \cup B_n \mid \mathbf{y}\right) \geq \rho_n\right\} \\ &\leq \exp(-g_n^*\log n) + \exp(-c_4 n) + \exp(-c_1 n\varepsilon_n^2),\end{aligned} \tag{25}$$

with

$$\rho_n = \frac{\exp\left(-c_2 n\varepsilon_n^2\right) + \exp\left(-c_5\delta g_n^*\log n\right)}{\exp(-g_n^*\log n)\exp\{-(c_3+c_3')g_n^*\log n\}} \to 0. \tag{26}$$

The result then follows from Borel-Cantelli lemma as the right-hand-side of (25) is summable. ■

A.3 Proof of Propositon 5

We begin with the following lemmas.

Lemma 11 *(Chernoff Bounds for Sum of Bernoulli Trials). Let $z = \sum_{i=1}^n Z_i$, where $Z_i = 1$ with probability p_i and $Z_i = 0$ with probability $1-p_i$, and all Z_i are independent. Then $\mathrm{pr}\big(z \geq (1+\delta_2)E(z)\big) \leq \exp\left(-\frac{\delta_2^2}{2+\delta_2}E(z)\right) = \exp\left(-\frac{\delta_2^2}{2+\delta_2}\sum_{i=1}^n p_i\right)$, for all $\delta_2 > 0$.*

Lemma 12 *Under Assumption (C6), both the R-NN graph and the restricted Delaunay triangulation graph are connected graphs with probability 1 as n tends to infinity.*

Proof [of Lemma 12] By Theorem 1.1 in Penrose (1999), it is readily to check that the minimum value of the radius γ_1 such that R-NN is connected equals the maximum edge length of the MST on $\mathcal{S}_n$, and it scales with $\{(\pi p_s^{\min})^{-1}\log n/n\}^{1/2}$ with probability 1 as

n tends to infinity. Notice that the MST is a subgraph of the Delaunay triangulation. By letting $\gamma_2 \asymp (\log n/n)^{1/2}$ and be larger than the maximum edge length of the minimum spanning tree, the restricted Delaunay triangulation contains all edges in the MST and hence is still a connected graph. ■

Then we prove Proposition 5.

Proof [of Proposition 5] Let $d(\mathbf{s}, \mathcal{B}) = \min_{\mathbf{s}_b \in \mathcal{B}} \|\mathbf{s} - \mathbf{s}_b\|$ denote the distance from a point $\mathbf{s} \in \mathbb{R}^2$ to a closed set $\mathcal{B} \subset \mathbb{R}^2$. For the boundary set $\mathcal{B}_{\beta^*}$, given $\upsilon_n > 0$, we define the υ_n-neighborhood of $\mathcal{B}_{\beta^*}$ as

$$\mathcal{N}(\mathcal{B}_{\beta^*}, \upsilon_n) = \left\{\mathbf{s} \in \mathbb{R}^2 : d(\mathbf{s}, \mathcal{B}_{\beta^*}) < \upsilon_n\right\}.$$

When Assumption (C6) holds, the maximum edge lengths in the R-NN graph and the restricted Delaunay triangulation graph scale with $(\log n/n)^{1/2}$. Therefore, by letting $\upsilon_n \asymp (\log n/n)^{1/2}$ and $\upsilon_n \geq \max(\gamma_1, \gamma_2)$, we can show that for any edge crossing $\mathcal{B}_{\beta^*}$, both of its endpoints must fall within $\mathcal{N}(\mathcal{B}_{\beta^*}, \upsilon_n)$.

We then define a set of edges from the original graph that have both endpoints within υ_n distance to the boundary set $\mathcal{B}_{\beta^*}$ as follows

$$\mathcal{E}_{\mathcal{B}}(\upsilon_n) := \left\{(i,j) : (i,j) \in \mathcal{E}_0 \text{ and } \max\left\{d\left(\mathbf{s}_i, \mathcal{B}_{\beta^*}\right), d\left(\mathbf{s}_j, \mathcal{B}_{\beta^*}\right)\right\} \leq \upsilon_n\right\}.$$

From the Definition 4, it is readily to check that the edge differences are all zero when $(i,j) \in \mathcal{E}_0 \backslash \mathcal{E}_{\mathcal{B}}(\upsilon_n)$, i.e., $\sum_{(i,j) \in \{\mathcal{E}_0 \backslash \mathcal{E}_{\mathcal{B}}(\upsilon_n)\}} |\beta_i^* - \beta_j^*|_0 = 0$.

For any given spanning tree $\mathcal{T}$, $|\mathcal{E}_{\mathcal{B}}(\upsilon_n) \cap \mathcal{E}_T| < z$, where $z = |\mathcal{S}_n \cap \mathcal{N}(\mathcal{B}_{\beta^*}, \upsilon_n)|$ denotes the number vertices falling within $\mathcal{N}(\mathcal{B}_{\beta^*}, \upsilon_n)$. The last inequality holds because $\mathcal{E}_{\mathcal{B}}(\upsilon_n) \cap \mathcal{E}_T$ is a spanning forest and hence its total number of edges is less than z.

Recall the boundary set $\mathcal{B}_{\beta^*}$ has a υ_n-covering number $N(\mathcal{B}_{\beta^*}, \upsilon_n, \|\cdot\|) \leq M_2 \upsilon_n^{-1}$, it follows that the υ_n-packing number $M(\mathcal{B}_{\beta^*}, \upsilon_n, \|\cdot\|) \leq M_2 \upsilon_n^{-1}$. From triangular inequality, there exists a maximal υ_n-packing for $\mathcal{B}_{\beta^*}$, denoted as $\mathbf{s}_{c,1}, \cdots, \mathbf{s}_{c,k}$ with the packing number $k \leq M_2 \upsilon_n^{-1}$ such that

$$\bigcup_{j=1,\ldots,k} B\left(\mathbf{s}_{c,j}, \upsilon_n/2\right) \subset \mathcal{N}(\mathcal{B}_{\beta^*}, \upsilon_n) \subset \bigcup_{j=1,\ldots,k} B\left(\mathbf{s}_{c,j}, 2\upsilon_n\right) \tag{27}$$

where $B\left(\mathbf{s}_c, \upsilon_n\right)$ denotes a ball centered at $\mathbf{s}_c$ with radius υ_n.

Therefore, z follows a binomial distribution with size n and

$$\begin{aligned} E(z) &\leq E\big(|\mathcal{S}_n \cap \{\bigcup_{j=1,\ldots,k} B\left(\mathbf{s}_{c,j}, 2\upsilon_n\right)\}|\big) \\ &\leq nkE\big(|\mathbf{s}_i \cap B\left(\mathbf{s}_{c,j}, 2\upsilon_n\right)|\big) \\ &= nk \int_{B(\mathbf{s}_{c,j}, 2\upsilon_n) \cap [0,1]^2} p_s(\mathbf{s}) d\mathbf{s} \\ &\leq 4\pi nk \upsilon_n^2 p_s^{\max} := E_{max} = O(nk\upsilon_n^2). \end{aligned}$$

Let $\tilde{z}$ be another binomial distribution that is independent from z with size n and $E(\tilde{z}) = E_{max}$. From Lemma 11,

$$\mathrm{pr}\{z \geq (1+\delta_2)E_{max}\} \leq \mathrm{pr}\{\tilde{z} \geq (1+\delta_2)E_{max}\} \leq \exp\left(-\frac{\delta_2^2}{2+\delta_2}E_{max}\right) \tag{28}$$

for all $\delta_2 > 0$. When $\upsilon_n \asymp (\log n/n)^{1/2}$, $E_{max} = O(nk\upsilon_n^2) = O(n\upsilon_n) = O\{(n\log n)^{1/2}\}$. Let $\delta_2 = 1$, then $P(z \geq 2E_{max}) \leq \exp(-E_{\max}/3) = \exp\{-M_4(n\log n)^{1/2}\}$ for some constant $M_4 > 0$. It implies with probability going to 1, the number of vertices falling within $\mathcal{N}(\mathcal{B}_{\beta^*}, \upsilon_n)$ is $O\{(n\log n)^{1/2}\}$.

Finally we have

$$\begin{aligned} |G^*_{\mathcal{T}}| = \sum_{(i,j)\in\mathcal{E}_T} |\beta_i^* - \beta_j^*|_0 &= \sum_{(i,j)\in\mathcal{E}_{\mathcal{B}}(\upsilon_n)\cap\mathcal{E}_T} |\beta_i^* - \beta_j^*|_0 + \sum_{(i,j)\in\{\mathcal{E}_0\setminus\mathcal{E}_{\mathcal{B}}(\upsilon_n)\}\cap\mathcal{E}_T} |\beta_i^* - \beta_j^*|_0 \\ &\leq |\mathcal{E}_{\mathcal{B}}(\upsilon_n)\cap\mathcal{E}_T| + \sum_{(i,j)\in\mathcal{E}_0\setminus\mathcal{E}_{\mathcal{B}}(\upsilon_n)} |\beta_i^* - \beta_j^*|_0 < z. \end{aligned}$$

Since z does not depend on the choice of $\mathcal{T}$, we have $g_n^* = \max_{\mathcal{T}\in\mathbb{T}_n} |G^*_{\mathcal{T}}| < z$. Combining with the result in (28), we complete the proof. ■

A.4 Proof of Corollary 6

Proof For $\mathcal{S}_n$ satisfying $g_n^* \leq M_3(n\log n)^{1/2}$ and $\log\tilde{P}_n \leq M_5 n^{1/2}\log^{3/2} n$, following the same proof of Theorem 3 with g_n^*, P_n and ε_n replaced by $M_3(n\log n)^{1/2}$, $\tilde{P}_n$ and $\tilde{\varepsilon}_n$ respectively, we have

$$\begin{aligned} &\mathrm{pr}^*\left\{\Pi_n\left(\frac{1}{\sqrt{n}}\|\boldsymbol{\mu}-\boldsymbol{\mu}^*\| \geq M_6\sigma^*\tilde{\varepsilon}_n \mid \mathbf{y}, \mathcal{S}_n\right) \geq \rho_n \mid \mathcal{S}_n\right\} \\ &\leq \exp(-M_3 n^{1/2}\log^{3/2} n) + \exp(-c_4 n) + \exp(-c_1 n\tilde{\varepsilon}_n^2), \end{aligned}$$

where ρ_n has the same form as (26), with possibly different constants that do not depend on $\mathcal{S}_n$. Let Q_n be the event that $g_n^* \leq M_3(n\log n)^{1/2}$ and $\log\tilde{P}_n \leq M_5 n^{1/2}\log^{3/2} n$ hold. Then

$$\begin{aligned} &\mathrm{pr}^*\left\{\Pi_n\left(\frac{1}{\sqrt{n}}\|\boldsymbol{\mu}-\boldsymbol{\mu}^*\| \geq M_6\sigma^*\tilde{\varepsilon}_n \mid \mathbf{y}, \mathcal{S}_n\right) \leq \rho_n\right\} \\ &\geq \int_{Q_n} \mathrm{pr}^*\left\{\Pi_n\left(\frac{1}{\sqrt{n}}\|\boldsymbol{\mu}-\boldsymbol{\mu}^*\| \geq M_6\sigma^*\tilde{\varepsilon}_n \mid \mathbf{y}, \mathcal{S}_n\right) \leq \rho_n \mid \mathcal{S}_n\right\} p_s(\mathcal{S}_n)\mathrm{d}\mathcal{S}_n \\ &\geq \left\{1 - \exp(-M_3 n^{1/2}\log^{3/2} n) - \exp(-c_4 n) - \exp(-c_1 n\tilde{\varepsilon}_n^2)\right\}\cdot \mathrm{pr}(Q_n). \end{aligned}$$

The result then follows since $\mathrm{pr}(Q_n) \to 1$ and $\rho_n \to 0$ as n tends to infinity. ■

A.5 Proof of Propositon 7

We begin with a brief review of Prim's algorithm for finding the MST and set up some notations. Prim's algorithm starts with an arbitrary vertex $\mathbf{s}_0$ of $\mathcal{G}_0$. In the t-th iteration, let $\mathcal{T}^t = (\mathcal{V}^t, \mathcal{E}^t)$ be a connected subgraph of the MST and $\tilde{\mathcal{E}}(\mathcal{V}^t) \subset \mathcal{E}_0$ be the set of all edges in $\mathcal{E}_0$ that has *one and only one* endpoint in $\mathcal{V}^t$ (for $t = 0$, we define $\mathcal{T}^0 = (\{\mathbf{s}_0\}, \emptyset)$). $\mathcal{T}^t$ is constructed by picking the edge in $\tilde{\mathcal{E}}(\mathcal{V}^{t-1})$ with the least edge weight and adding this edge and its endpoint that is not in $\mathcal{V}^{t-1}$ into $\mathcal{T}^{t-1}$. The algorithm stops when $\mathcal{V}^t$ includes all the vertices in $\mathcal{G}_0$.

Proof [of Propositon 7] Let A^t be the event that $\mathcal{T}^t$ is a connected subgraph of $\mathcal{T}$. It suffices to show that A^t happens with nonzero probability for all t. Notice that by Prim's algorithm, $A^t \subset A^{t-1}$ and thus

$$\text{pr}(A^t) = \text{pr}(A^t|A^{t-1})\text{pr}(A^{t-1}). \tag{29}$$

Consider two cases: (i) all vertices in $\mathcal{V}_0 \setminus \mathcal{V}^{t-1}$ have different cluster memberships than the ones in $\mathcal{V}^{t-1}$, and (ii) otherwise. For (i), let e be an arbitrary edge in $\tilde{\mathcal{E}}(\mathcal{V}^{t-1})$. Then

$$\text{pr}(A^t|A^{t-1}) \geq \text{pr}(\{e \text{ has the minimal weight among } \tilde{\mathcal{E}}(\mathcal{V}^{t-1})\}) > 0. \tag{30}$$

The strict inequality is due to the i.i.d. Uniform$(1/2, 1)$ on the weights of $\tilde{\mathcal{E}}(\mathcal{V}^{t-1})$. For (ii), let e be an edge in $\tilde{\mathcal{E}}(\mathcal{V}^{t-1})$ connecting two endpoints in the same cluster. Then (30) still holds due to the way that we sample edge weights. The proposition then follows by induction arguments on t using (29).

■

Appendix B. RJ-MCMC Algorithm

In this appendix we provide details of our RJ-MCMC algorithm.

Recall from Section 3.4 that in each iteration of RJ-MCMC, we further iterate through each covariate from $m = 1$ to p. In each inner iteration one of the following four moves, birth, death, change, and hyper, is performed with probabilities $r_B(k_m), r_D(k_m), r_C(k_m)$ and $r_H(k_m)$, respectively. We set $r_B(k) = r_D(k) = 0.425$ for $k \in \{2, 3, \ldots, n-1\}$, $r_B(k) = 0.85$ for $k = 1$, $r_D(k) = 0.85$ for $k = n$, $r_C(k) = 0.1$ and $r_H(k) = 0.05$ for $k \in \{1, \ldots, n\}$.

Detailed implementation as well as acceptance probability of each move are given as follows.

(a) *Birth* ($k_m \to k_m+1$): Randomly choose one edge from $n-k_m$ edges in the spanning tree $\mathcal{T}^{(m)}$ that connect vertices belonging to a same cluster with equal probability. Suppose the chosen edge connects two endpoints $\mathbf{s}_i, \mathbf{s}_{i'} \in \mathcal{C}_j^{(m)}$ with $i < i'$. By removing this edge we split $\mathcal{C}_j^{(m)}$ into two connected components, one containing $\mathbf{s}_i$ and other containing $\mathbf{s}_{i'}$. We set the component containing $\mathbf{s}_{i'}$ to be a new cluster $\mathcal{C}_{k_m+1}^{(m)\star}$ and set the other one to be $\mathcal{C}_j^{(m)\star}$. We let $\mathcal{C}_l^{(m)\star} = \mathcal{C}_l^{(m)}$ for $l = 1, \ldots, j-1, j+1, \ldots, k_m$. By doing so we propose a new partition $\pi^{(m)\star}$.

The acceptance probability is

$$\alpha_1 = \min\{1,\ \mathcal{A} \cdot \mathcal{P} \cdot \mathcal{L}\}, \tag{31}$$

where

$$\mathcal{A} = \frac{k_m}{n - k_m} \cdot (1 - c)$$

is the prior ratio,

$$\mathcal{P} = \frac{r_D(k_m + 1)}{r_B(k_m)} \cdot \frac{n - k_m}{k_m}$$

is the proposal ratio,

$$\mathcal{L} = \frac{p\left[\mathbf{y} \ \middle|\ \pi^{(m)\star}, k_m + 1, \mathcal{T}^{(m)}, \left\{\pi^{(l)}, k_l, \mathcal{T}^{(l)}\right\}_{l \neq m}, \sigma^2, \lambda\right]}{p\left[\mathbf{y} \ \middle|\ \left\{\pi^{(m)}, k_m, \mathcal{T}^{(m)}\right\}_{m=1}^{p}, \sigma^2, \lambda\right]}$$

is the likelihood ratio whose numerator and denominator are given by (6).

(b) *Death* ($k_m + 1 \to k_m$): Randomly choose one edge from k_m edges in the spanning tree $\mathcal{T}^{(m)}$ that connect different clusters with equal probability. Suppose the chosen edge connects two endpoints $\mathbf{s}_i \in \mathcal{C}_j^{(m)}$ and $\mathbf{s}_{i'} \in \mathcal{C}_{j'}^{(m)}$ with $i < i'$. We merge these two clusters to be $C_j^{(m)\star}$ and remove $\mathcal{C}_{j'}^{(m)}$. We set $\mathcal{C}_l^{(m)\star} = \mathcal{C}_l^{(m)}$ for $l < j'$, and $\mathcal{C}_l^{(m)\star} = \mathcal{C}_{l+1}^{(m)}$ for $l \geq j'$. Then we propose $\pi^{(m)\star}$. The acceptance probability is the reciprocal of the one in birth step, i.e., $1/\alpha_1$, where α_1 is given by (31).

(c) *Change* ($k_m \to k_m$): First perform a death step by merging $\mathcal{C}_{j_1}^{(m)}$ and $\mathcal{C}_{j_2}^{(m)}$ to be $\mathcal{C}_{j_1'}^{(m)\star}$, and then perform a birth step by splitting $\mathcal{C}_{j_3}^{(m)\star}$ to be $\mathcal{C}_{j_3}^{(m)\star\star}$ and $\mathcal{C}_k^{(m)\star\star}$. The acceptance probability is $\alpha_1 = \min\{1,\ \mathcal{A} \cdot \mathcal{P} \cdot \mathcal{L}\}$, where $\mathcal{A} = 1$, $\mathcal{P} = 1$, and

$$\mathcal{L} = \frac{p\left[\mathbf{y} \ \middle|\ \pi^{(m)\star\star}, k_m, \mathcal{T}^{(m)}, \left\{\pi^{(l)}, k_l, T^{(l)}\right\}_{l \neq m}, \sigma^2, \lambda\right]}{p\left[\mathbf{y} \ \middle|\ \left\{\pi^{(m)}, k_m, T^{(m)}\right\}_{m=1}^{p}, \sigma^2, \lambda\right]}.$$

(d) *Hyper*: In this step $\mathcal{T}^{(m)}, \sigma^2$ and λ are updated. We first update σ^2 by a Gibbs step:

$$\sigma^2 \sim \mathrm{IG}\left(\frac{n + a_0}{2},\ \frac{1}{2}[b_0 + \mathbf{y}^{\mathrm{T}} \mathbf{P}_\lambda^{-1} \mathbf{y}]\right).$$

To update $\mathbf{w}^{(m)}$ (and hence $\mathcal{T}^{(m)}$), a Metropolis-Hastings procedure is utilized. We first sample edge weights of the cross-cluster edges from i.i.d. Uniform$(1/2, 1)$ and edge weights of those within-cluster edges from i.i.d. Uniform$(0, 1/2)$. Then we propose a new spanning tree using Prim's algorithm based on the new weights. The proposed spanning tree is guaranteed to induce the current partition $\pi^{(m)}$ (Teixeira et al., 2015). Since the full conditional of $\mathbf{w}^{(m)}$ remains the same for the proposed weights, the acceptance probability is always 1.

Finally we update λ using a Metropolis-Hastings step with a symmetric random walk proposal. We propose $\lambda^\star$ by

$$\log \lambda^\star \sim \mathrm{N}(\log \lambda, \sigma^2_{MH}),$$

and the acceptance probability is $\alpha_1 = \min\{1,\ \mathcal{A} \cdot \mathcal{P} \cdot \mathcal{L} \cdot \lambda^\star/\lambda\}$, where

$$\mathcal{A} = \left(\frac{\lambda^\star}{\lambda}\right)^{c_0/2-1} \exp\{-d_0(\lambda^\star - \lambda)/2\}$$

is the prior ratio, $\mathcal{P} = 1$ is the proposal ratio, and

$$\mathcal{L} = \frac{p\left[\mathbf{y} \mid \left\{\pi^{(m)}, k_m, \mathcal{T}^{(m)}\right\}_{m=1}^p, \sigma^2, \lambda^\star\right]}{p\left[\mathbf{y} \mid \left\{\pi^{(m)}, k_m, \mathcal{T}^{(m)}\right\}_{m=1}^p, \sigma^2, \lambda\right]}$$

is the likelihood ratio.

Appendix C. Additional Simulation Results

In this appendix we provide results on additional simulation settings.

C.1 Sensitivity Analysis of c

We first examine how sensitive the results from BSCC model to α. We reconsider the 100 data sets with moderate spatial correlation that are used in the Simulation Studies section. We fit BSCC models with four candidates $\alpha \in \{0.0075, 0.0150, 0.1000, 0.3333\}$, which give $c = 0.05, 0.1, 0.5, 0.9$, respectively.

Figure 10 shows MSEs for BSCC models under different candidate values of α (or equivalently, c). We can see in all settings BSCC outperforms SCC in terms of MSEs, and overall the MSEs for BSCC are not sensitive to α (or c). However, careful choice of α does lead to improvements in MSEs.

Recall that Table 1 in the main text shows the number of data sets in which WAIC prefers a candidate value of α. In most of the data sets $\alpha = 0.0075$ or 0.0150 is preferred, which are two models with least MSE (see Figure 10). Also notice that $\alpha = 0.3333$ that leads to higher MSE is rarely chosen by WAIC.

In summary, our simulation results suggest that the MSE performance is fairly robust to the choice of α (and thus c), as long as the value of α is within a reasonable range (e.g., $\alpha \leq 0.1$ in this example). We hence recommend using WAIC to determine the desired range of α.

C.2 Simulations under Different σ

In this subsection we evaluate the performance of BSCC under different settings of signal-to-noise ratio (SNR). We regenerate data sets from (12) with $\sigma \in \{0.1, 0.5, 0.75, 1\}$, and 100 data sets are generated for each value of σ. The rest data generating settings are the same as the ones for data sets with a moderate spatial correlation. The choices of σ correspond to different levels of SNR—as σ in increases, the variation in the residuals becomes larger

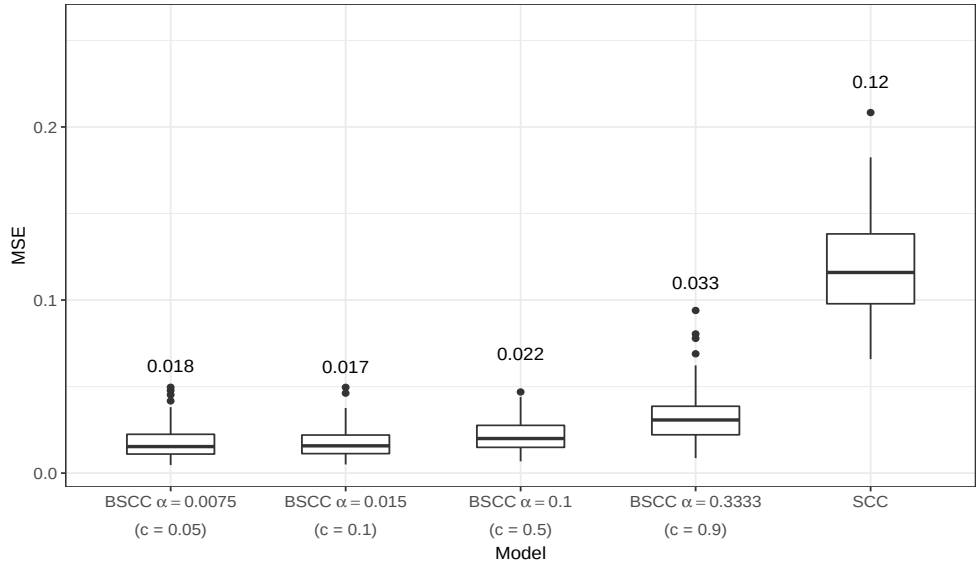

Figure 10: Boxplots of MSEs for BSCC method under 4 different choices of hyperparameter α (or equivalently, c). 100 simulations are run for each choice. The average MSE_β over 100 simulations is shown above each box. MSEs for SCC method is also shown for reference.

	Rand index					
	β_1		β_2		β_3	
σ	BSCC	SCC	BSCC	SCC	BSCC	SCC
0.1	0.983	0.722	0.987	0.825	0.994	0.853
0.5	0.902	0.737	0.904	0.830	0.931	0.852
0.75	0.816	0.736	0.825	0.822	0.869	0.849
1	0.751	0.734	0.763	0.822	0.818	0.846

Table 3: The average Rand indices for BSCC and SCC methods over 100 simulations under 4 different settings of SNR.

with respect to spatially varying effects in $\mathbf{x}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\beta}(\mathbf{s})$. We fit BSCC and SCC models to each data set using the same settings as in the main text.

Figure 11 presents boxplots of MSEs for both models under different choice of SNRs, and Table 3 shows average Rand indices. As expected, the MSE performance of both methods degenerates as SNRs decrease. In terms of partition recovery, the Rand indices for BSCC also decreases as σ becomes larger. When $\sigma \in \{0.1, 0.5, 0.75\}$, BSCC outperforms SCC in both coefficient estimation and partition recovery. In the extreme case where $\sigma = 1$, BSCC still has a better MSE but slightly lower Rand indices.

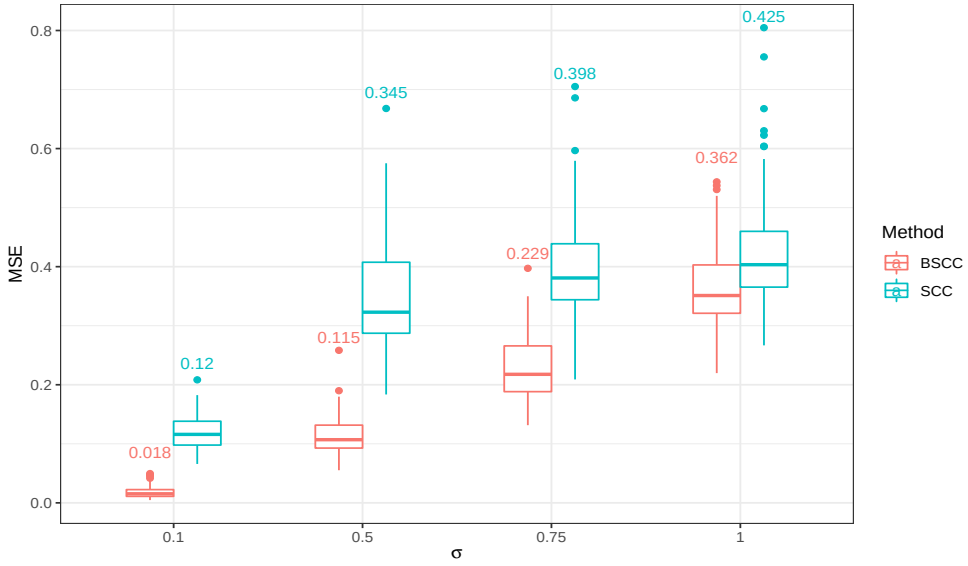

Figure 11: Boxplots of MSEs for BSCC and SCC methods under 4 different choices of noise standard deviation σ. 100 simulations are run for each choice. The average MSE_β over 100 simulations is shown above each box.

C.3 Simulations under Different Cross-Correlations

In many spatial applications, in addition to spatial dependence within each covariate, there may also be cross-dependence among covariates. In this subsection we investigate how BSCC performs under different settings of cross-dependence.

As discussed in Section 5.1, the two covariates in the simulation data are generated by a linear transformation of two independent Gaussian process realizations: $x_1(\mathbf{s}_i) = \zeta_1(\mathbf{s}_i), \quad x_2(\mathbf{s}_i) = r\zeta_1(\mathbf{s}_i) + \sqrt{1-r^2}\zeta_2(\mathbf{s}_i)$, where ζ_m $(m = 1, 2)$ is the realization of a Gaussian process and r controls the strength of cross-correlation between x_1 and x_2.

We consider $r \in \{0, 0.375, 0.75, 0.9\}$, which corresponds to zero, weak, moderate, and strong cross-correlation cases, respectively. For each value of r, we regenerate 100 data sets using the same true clustering patterns as Figure 2 in the main text shows. In practice, however, one may expect highly correlated covariates to have similar clustering configurations in their coefficients. As a result, we further consider a scenario where $r = 0.9$ and β_1 shares the same true partition as β_2 (Figure 12). We refer to this scenario as "correlated partitions" in what follows. We fit BSCC and SCC models to each of them using the same settings as in the main text.

Figure 13 shows MSEs under the five settings, and BSCC outperforms SCC in all of them. When β_1 and β_2 have different true clustering patterns, the MSE performance of BSCC is fairly robust to multicollinearity. This result is not surprising for two reasons. First, we assume a ridge regression type of prior on $\boldsymbol{\beta}$ conditional on the partitions that mitigates multicollinearity problems. Second, the matrix $\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}}$ is well-conditioned when the partitions of β_1 and β_2 are different, where $\tilde{\mathbf{X}}$ is the transformed design matrix. When β_1

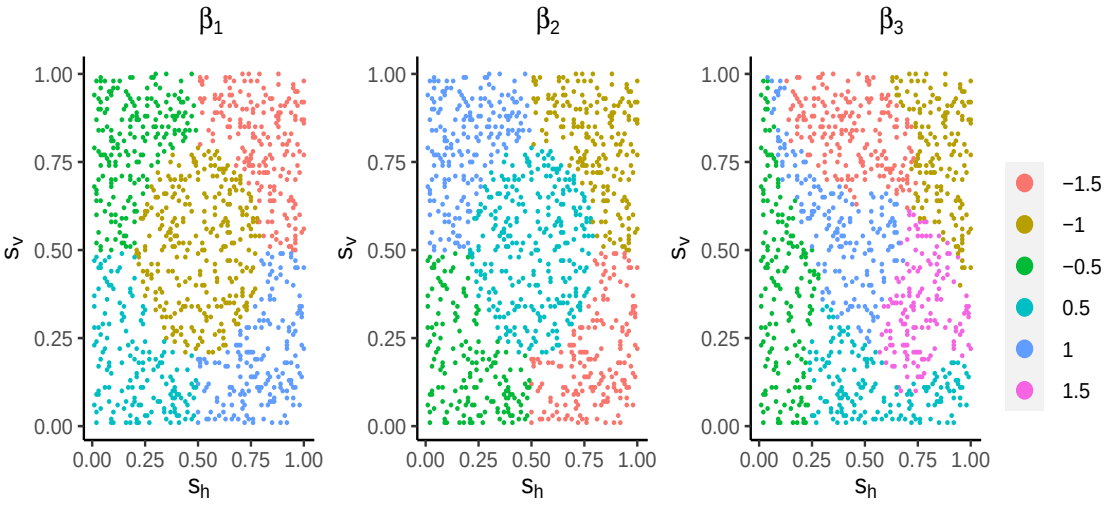

Figure 12: Spatial structures of true coefficients used in the correlated partitions scenario in Section C.3, where β_1 and β_2 have the same true partitions.

	Rand index					
	β_1		β_2		β_3	
Cross-covariate correlation	BSCC	SCC	BSCC	SCC	BSCC	SCC
$r = 0$	0.984	0.719	0.988	0.824	0.994	0.853
$r = 0.375$	0.985	0.719	0.988	0.824	0.994	0.853
$r = 0.75$	0.983	0.722	0.987	0.825	0.994	0.853
$r = 0.9$	0.980	0.722	0.985	0.826	0.994	0.852
$r = 0.9$ with correlated partitions	0.961	0.830	0.963	0.829	0.989	0.853

Table 4: The average Rand indices for BSCC and SCC methods over 100 simulations under 5 different settings of cross-covariate correlation.

and β_2 share the same true partitions, the multicollinearity problem becomes more severe in $\tilde{\mathbf{X}}$ and we observe a drop in the accuracy of coefficient estimation.

The Rand indices under fives scenarios are shown in Table 4. Similar to the findings in terms of MSEs, the partition estimation performance of BSCC is robust when β_1 and β_2 have different true partitions. On the other hand, when they have an identical partition, partition recovery for both coefficients become worse, probably due to the interference of the posterior distributions of the two partitions, as pointed out by an anonymous reviewer.

C.4 Comparisons with DPM Models with Spatial Random Effects

In this subsection we compare our method to the original version of the DPM model proposed by Ma et al. (2020), which includes a spatially varying intercept term modelled by a Gaussian process (referred to as DPM-GP model). We adopt the same hyperparameter

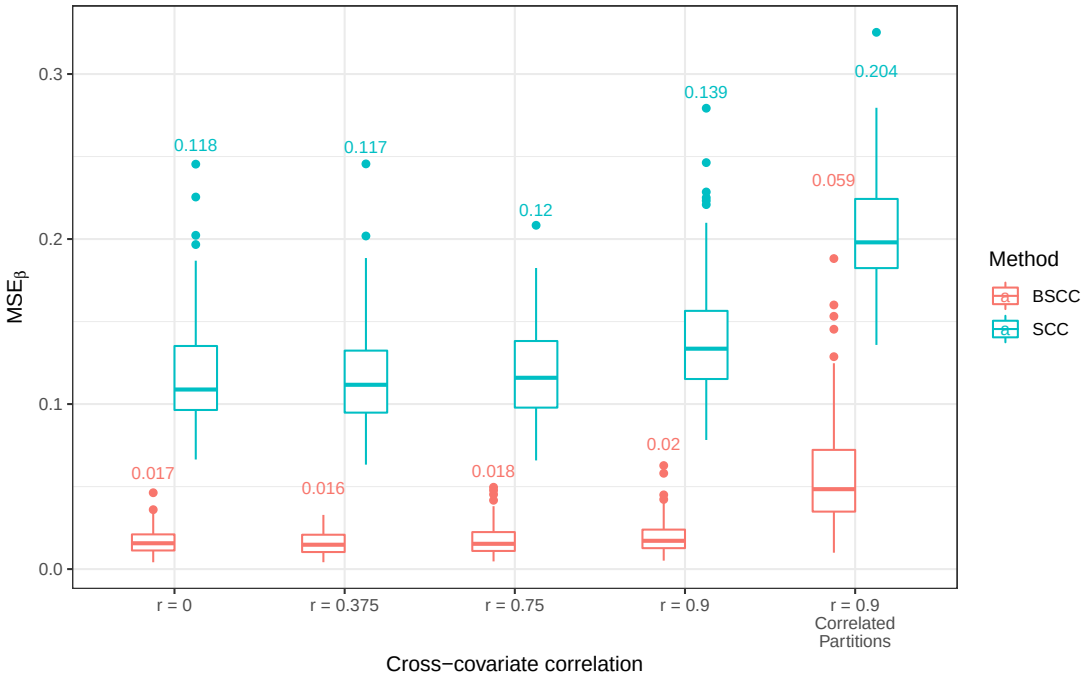

Figure 13: Boxplots of MSEs for BSCC and SCC methods under 5 settings of cross-covariate correlation. "Correlated partitions" refers to the scenario where β_1 shares same true partition as β_2. 100 simulations are run for each choice. The average MSE_β over 100 simulations is shown above each box.

	BSCC	SCC	DPM	DPM-GP
β_1	0.986	0.718	0.683	0.664
β_2	0.984	0.822	0.776	0.751
β_3	0.997	0.848	0.817	0.781

Table 5: The average Rand indices for BSCC, SCC, DPM, and DPM-GP methods over 10 simulations with moderate spatial correlation.

settings as in the code provided in their paper, except that we set the maximum possible number of clusters to 50. We run the chain for 20,000 iterations, discard the first half, and collect posterior samples every 10 iterations after burn-in. It takes on average 11 hours to run a DPM-GP model for one simulation data set used in the main text. Due to its computational expensiveness, we only run the model for the first 10 data sets with a moderate spatial correlation.

Figure 14 and Table 5 show the MSEs and Rand indices of BSCC, SCC, DPM, and DPM-GP models for the 10 data sets, respectively. BSCC model achieves the best performance among the four models in estimating coefficient values and partitions.

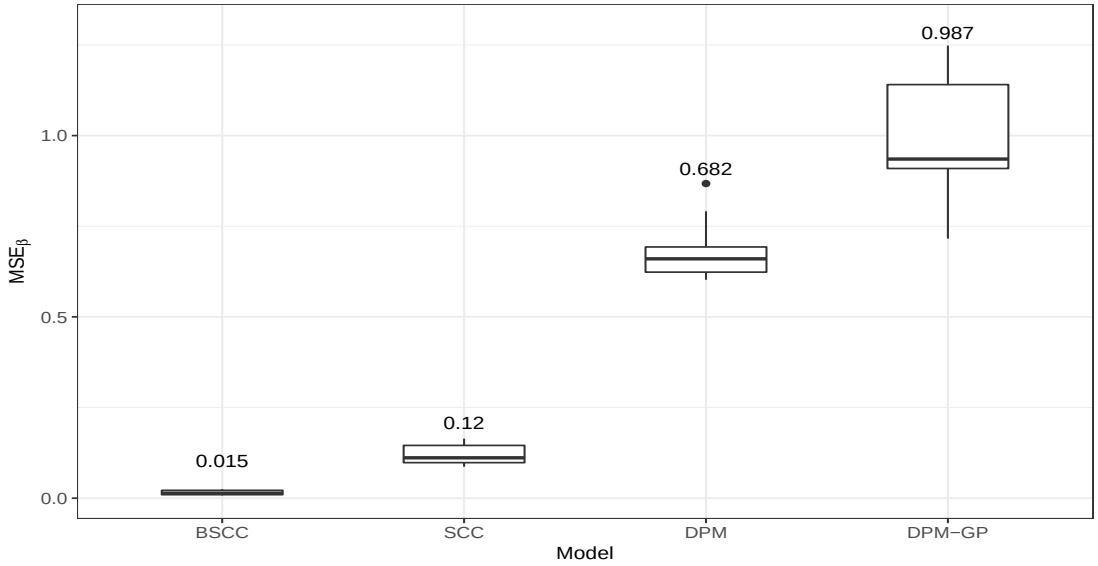

Figure 14: Boxplots of MSEs for BSCC, SCC, DPM, and DPM-GP methods for 10 data sets with moderate spatial correlation. The average MSE_β over 10 simulations is shown above each box.

Appendix D. Discussion on RJ-MCMC

D.1 Mixing of RJ-MCMC

In this subsection we discuss the mixing of tempered RJ-MCMC chains in more details. We consider the data set with a moderate spatial correlation that is analyzed in the Simulation Studies section of the main text, and compare the BSCC model fittings with and without parallel tempering (which are referred to as tempered and untempered models/chains, respectively, in what follows). Both chains are run for $50,000$ iterations after a burn-in period of the same length, and we thin the chains by taking samples every 20 iterations. For the tempered model, we adopt the sigmoidal temperature ladder (Gramacy and Taddy, 2010) with minimum inverse temperature $t_d = 0.35$ and run 8 parallel chains. See Section 5.1 in the main text for other settings of the RJ-MCMC algorithm.

Table 6(a) shows acceptance rates of each move in each of the tempered chains. The chains with inverse temperatures less than 1 have flatter target distributions than the posterior distribution, allowing for a more efficient exploration of the state space, as suggested by the fact that most of the chains with low inverse temperatures have higher Metropolis-Hastings acceptance rates. In particular, the acceptance rates for the Birth, Death, and Change moves of the hottest chain (i.e., with the lowest inverse temperature) are at least twice as high as their counterparts in the coolest chain.

Due to the higher acceptance rates, the hotter chains are able to visit the states that are hard to visit by conventional samplers. These states are passed to cooler chains via state swapping between chains. Acceptance rates of the swap attempts are shown in Table 7.

(a) Tempered model

Chain #	Inverse temperature	Birth	Death	Change	Hyper
1	1.000	0.177	0.179	0.090	0.495
2	0.989	0.211	0.211	0.116	0.543
3	0.967	0.276	0.277	0.184	0.554
4	0.922	0.184	0.187	0.096	0.486
5	0.841	0.169	0.172	0.084	0.489
6	0.708	0.174	0.176	0.083	0.508
7	0.532	0.239	0.241	0.140	0.526
8	0.350	0.364	0.364	0.264	0.548

(b) Untempered model

Birth	Death	Change	Hyper
0.154	0.156	0.067	0.481

Table 6: Acceptance rates of the four moves in (a) tempered model and (b) untempered model.

Chain #	1	2	3	4	5	6	7	8
Inverse temperature	1.000	0.989	0.967	0.922	0.841	0.708	0.532	0.350
Acceptance rate	0.620	0.526	0.581	0.566	0.453	0.367	0.144	0.055

Table 7: Swap acceptance rates of tempered chains.

The swap acceptance rates are lower for hotter chains, probably due to larger gaps between adjacent inverse temperatures.

As a comparison, the acceptance rates for Metropolis-Hastings moves of the untempered chain are lower (Table 6(b)), suggesting that the parallel tempering techniques can improve the efficiency for exploring the posterior space.

Traceplots of the thinned posterior densities after burn-in of the tempered and untempered models are shown in Figure 15, where the densities for the tempered model are computed based on the draws from the coolest chain. The chains from both models seem to converge, but the tempered chain exhibits better mixing and less autocorrelation. The tempered chain transits between high posterior regions and low posterior regions more quickly and it visits low posterior regions more frequently.

Finally, we look at posterior distributions of the number of clusters for each coefficient obtained from the tempered and untempered models, which are shown in Figure 16. The conventional untempered chain concentrates more on the regions near the posterior mode, while with the aid of parallel tempering, the tempered chain is able to visit some partitions that the untempered chain never does. For the coefficient β_3, for example, the tempered chain frequently visits partitions with 6 clusters, which are missed by the untempered chain. As indicated by the right tails, the untempered chain also underestimates the probability of getting partitions with large number of clusters.

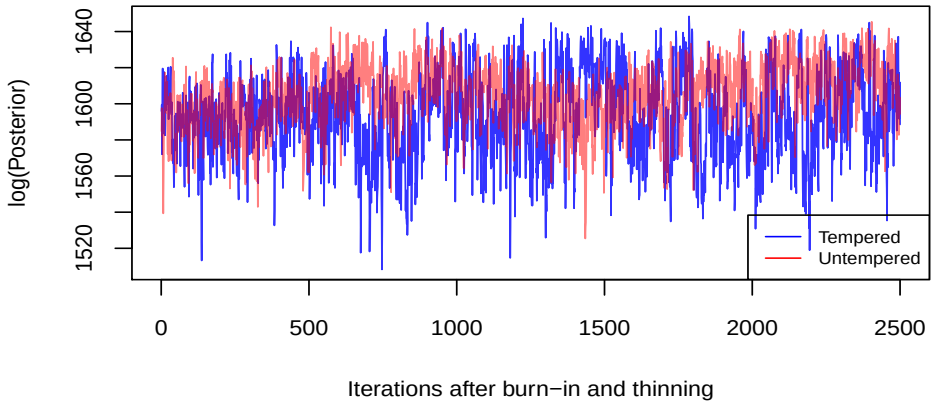

Figure 15: Traceplot of thinned log posterior densities from tempered and untempered model after burn-in period.

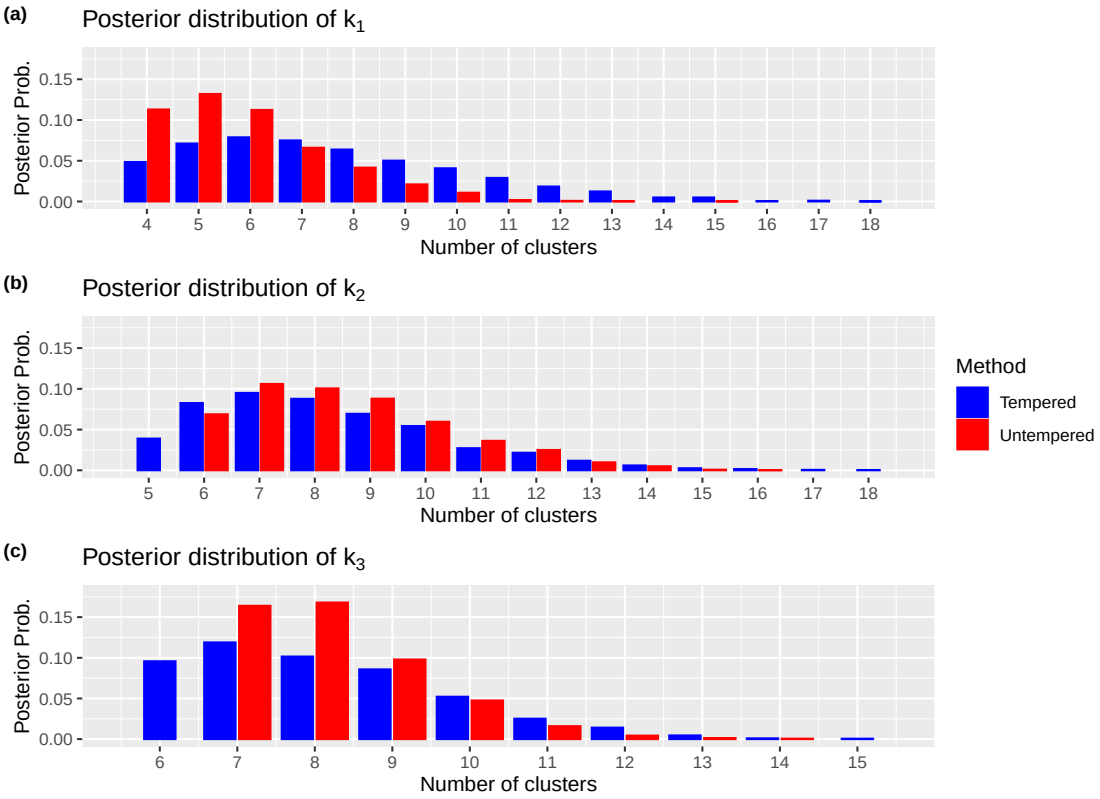

Figure 16: Posterior distributions of k_m, the number of clusters for coefficient β_m, estimated from MCMC samples of the tempered and untempered models.

	Birth	Death	Change	Hyper
With BAP	0.142	0.145	0.068	0.477
Without BAP	0.154	0.156	0.067	0.481

Table 8: Acceptance rates of the four moves with and without BAPs.

D.2 Boundary-Adjusted Proposals

In this subsection we include the results of applying boundary-adjusted proposals (BAPs) for splitting clusters. The idea is that proposals splitting a cluster near its boundary is more likely to be accepted, which might improve mixing. BAPs thus assign higher probability on removing edges near boundaries. However, we do not observe satisfying improvement in mixing for this proposal. We summarize our methods and numerical results below.

Given partitions of all covariates $\{\pi^{(m)}\}_{m=1}^{p}$, we divide the vertex set $\mathcal{V}$ into two subsets, namely, *internal* vertices and *boundary* vertices, using 3-nearest neighbors methods. Specifically, a vertex is an internal vertex if all of its 3 nearest neighbors have the same cluster memberships for all covariates; otherwise, we treat it as a boundary vertex. We further divide the edge set $\mathcal{E}$ into three subsets to distinguish which edges are on the boundaries of clusters that we should target at:

1. *Between-cluster* edges: We define an edge to be a *between-cluster* edge if it is connecting two vertices belonging to different clusters.

2. *Boundary* edges: We define an edge to be a *boundary* edge if it is not a between-cluster edge and at least one of its endpoints is a boundary vertex. BAPs place higher probability on removing this type of edges.

3. *Within-cluster* edges: We define an edge to be a *within-cluster* edge if it is not a between-cluster edge and both of its endpoints are internal vertices.

In BAPs, a cluster is uniformly chosen to be split. Then with probability p_w, a within-cluster edge that connects two vertices in this cluster is removed, and with probability $1 - p_w$, a boundary edge is chosen to remove.

In this following simulation, we apply BAPs to the data set analyzed in Section D.1. We set $p_w = 0.2$ and do not apply parallel tempering.

Figure 17 shows the thinned posterior densities after burn-in of the models with and without BAPs, and Table 8 shows the acceptance rates of each move for both models. It seems that applying BAPs does not improve our results in terms of mixing and acceptance rates. Further investigations on more efficient partition proposals, including combining BAPs with parallel tempering, are left as future works.

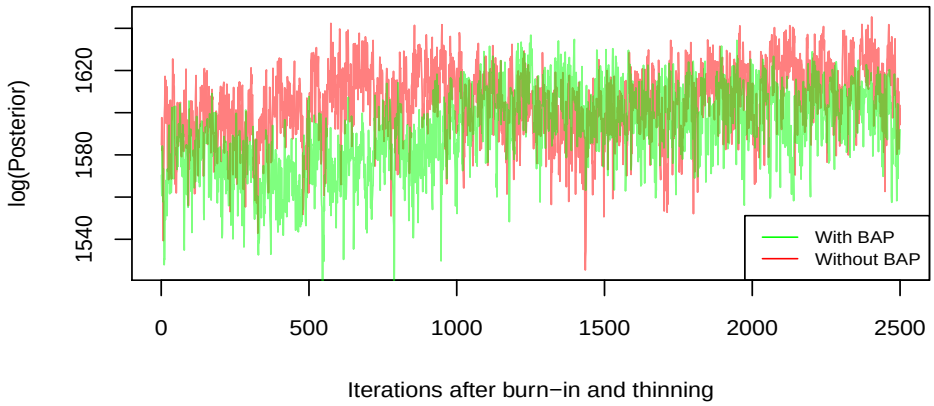

Figure 17: Traceplot of thinned log posterior densities after burn-in period from models with and without BAPs.

References

James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679, 1993.

Artin Armagan, David B Dunson, Jaeyong Lee, Waheed U Bajwa, and Nate Strawn. Posterior consistency in linear models under shrinkage priors. *Biometrika*, 100(4):1011–1018, 2013.

Renato M Assunção, Marcos Corrêa Neves, Gilberto Câmara, and Corina da Costa Freitas. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, 20(7):797–811, 2006.

Orhun Aydin, Mark V Janikas, Renato Assunção, and Ting-Hwan Lee. SKATER-CON: Unsupervised regionalization via stochastic tree partitioning within a consensus framework using random spanning trees. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 33–42, 2018.

Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, 2014.

Andrew R Barron. Information-theoretic characterization of bayes performance and the choice of priors in parametric and nonparametric problems. In JM Bernardo, J Burger, and ADEM Smith, editors, *Bayesian Statistics 6*, pages 27–52. Oxford University Press, 1998.

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

Leo Breiman, Jerome Friedman, Richard A Olshen, and Charles J Stone. *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.

Yanqing Chen, Timothy A Davis, William W Hager, and Sivasankaran Rajamanickam. Algorithm 887: CHOLMOD, supernodal sparse Cholesky factorization and update/downdate. *ACM Transactions on Mathematical Software (TOMS)*, 35(3):22, 2008.

Hugh A Chipman, Edward I George, and Robert E McCulloch. Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935–948, 1998.

Hugh A Chipman, Edward I George, and Robert E McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

Tingjin Chu, Jun Zhu, and Haonan Wang. Penalized maximum likelihood estimation and variable selection in geostatistics. *The Annals of Statistics*, 39(5):2607–2625, 2011.

David GT Denison and Christofer C Holmes. Bayesian partitioning for estimating disease risk. *Biometrics*, 57(1):143–149, 2001.

David GT Denison, Bani K Mallick, and Adrian FM Smith. A Bayesian CART algorithm. *Biometrika*, 85(2):363–377, 1998.

Reinhard Diestel. *Graph Theory*. Electronic library of mathematics. Springer, 5 edition, 2016.

Wenning Feng, Chae Young Lim, Tapabrata Maiti, and Zhen Zhang. Spatial regression and estimation of disease risks: A clustering-based approach. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 9(6):417–434, 2016.

Stewart Fotheringham, Chris Brunsdon, and Martin Charlton. *Geographically Weighted Regression*. John Wiley & Sons, Chichester, 2003.

Ronald E Gangnon and Murray K Clayton. Bayesian detection and modeling of spatial disease clustering. *Biometrics*, 56(3):922–935, 2000.

Alan E Gelfand, Hyon-Jung Kim, CF Sirmans, and Sudipto Banerjee. Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462):387–396, 2003.

Alan E Gelfand, Athanasios Kottas, and Steven N MacEachern. Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, 100(471):1021–1035, 2005.

Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and computing*, 24(6):997–1016, 2014.

Charles J Geyer. Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 156–163. Interface Foundation of North America, 1991.

Robert B Gramacy and Herbert K H Lee. Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103 (483):1119–1130, 2008.

Robert B Gramacy and Matthew Taddy. Categorical inputs, sensitivity analysis, optimization and importance tempering with tgp version 2, an R package for treed Gaussian process models. *Journal of Statistical Software*, 33(6):1–48, 2010.

Peter J Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

Peter J Green and Robin Sibson. Computing Dirichlet tessellations in the plane. *The Computer Journal*, 21(2):168–173, 1978.

Oleksandr Grygorash, Yan Zhou, and Zach Jorgensen. Minimum spanning tree based clustering algorithms. In *2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06)*, pages 73–81. IEEE, 2006.

Yongtao Guan and Matthew Stephens. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, pages 1780–1815, 2011.

Diansheng Guo. Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science*, 22 (7):801–823, 2008.

Avril Hegarty and Daniel Barry. Bayesian disease mapping using product partition models. *Statistics in Medicine*, 27(19):3868–3893, 2008.

Hyoung-Moon Kim, Bani K Mallick, and Chris C Holmes. Analyzing nonstationary spatial data using piecewise Gaussian processes. *Journal of the American Statistical Association*, 100(470):653–668, 2005.

Leonhard Knorr-Held and Günter Raßer. Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, 56(1):13–21, 2000.

Bledar A Konomi, Huiyan Sang, and Bani K Mallick. Adaptive Bayesian nonstationary modeling for large spatial datasets using covariance approximations. *Journal of Computational and Graphical Statistics*, 23(3):802–829, 2014.

Daniel R Kowal, David S Matteson, and David Ruppert. Dynamic shrinkage processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4):781–804, 2019.

Martin Kulldorff. A spatial scan statistic. *Communications in Statistics-Theory and Methods*, 26(6):1481–1496, 1997.

Martin Kulldorff and Neville Nagarwalla. Spatial disease clusters: detection and inference. *Statistics in Medicine*, 14(8):799–810, 1995.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Der-Tsai Lee and Bruce J Schachter. Two algorithms for constructing a delaunay triangulation. *International Journal of Computer & Information Sciences*, 9(3):219–242, 1980.

Junho Lee, Ronald E Gangnon, and Jun Zhu. Cluster detection of spatial regression coefficients. *Statistics in Medicine*, 36(7):1118–1133, 2017.

Furong Li and Huiyan Sang. Spatial homogeneity pursuit of regression coefficients for large datasets. *Journal of the American Statistical Association*, 114(527):1050–1062, 2019.

Pei-Sheng Lin. Generalized scan statistics for disease surveillance. *Scandinavian Journal of Statistics*, 41(3):791–808, 2014.

Pei-Sheng Lin, Yi-Hung Kung, and Murray Clayton. Spatial scan statistics for detection of multiple clusters with arbitrary shapes. *Biometrics*, 72(4):1226–1234, 2016.

Antonio R Linero and Yun Yang. Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):1087–1110, 2018.

Zhihua Ma, Yishu Xue, and Guanyu Hu. Heterogeneous regression models for clusters of spatial dependent data. *Spatial Economic Analysis*, 15(4):459–475, 2020.

Jingru Mu, Guannan Wang, and Li Wang. Estimation and inference in spatially varying coefficient models. *Environmetrics*, 29(1):e2485, 2018.

Michael A Osborne. *Bayesian Gaussian processes for sequential prediction, optimisation and quadrature*. PhD thesis, Oxford University, UK, 2010.

Garritt L Page and Fernando A Quintana. Spatial product partition models. *Bayesian Analysis*, 11(1):265–298, 2016.

Richard D Payne, Nilabja Guha, Yu Ding, and Bani K Mallick. A conditional density estimation partition model using logistic Gaussian processes. *Biometrika*, 107(1):173–190, 2020.

Mathew D Penrose. A strong law for the longest edge of the minimal spanning tree. *The Annals of Probability*, 27(1):246–260, 1999.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349, 2013.

Franco P Preparata and Michael I Shamos. *Computational Geometry: An Introduction*. Springer Science & Business Media, 2012.

Veronika Ročková and Enakshi Saha. On theory for BART. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89, pages 2839–2848. PMLR, 2019.

Veronika Ročková and Stéphanie van der Pas. Posterior concentration for Bayesian regression trees and forests. *Annals of Statistics*, 48(4):2108–2131, 2020.

James G Simmonds. *A Brief on Tensor Analysis*. Springer Science &Simmonds Business Media, 2012.

Qifan Song and Guang Cheng. Bayesian fusion estimation via t shrinkage. *Sankhya A*, 82 (2):353–385, 2020.

Lynne D Talley. *Descriptive Physical Oceanography: An Introduction*. Academic Press, London, 2011.

Leonardo V Teixeira, Renato M Assunção, and Rosangela Helena Loschi. A generative spatial clustering model for random data through spanning trees. In *2015 IEEE International Conference on Data Mining*, pages 997–1002. IEEE, 2015.

Leonardo V Teixeira, Renato M Assunção, and Rosangela H Loschi. Bayesian space-time partitioning by sampling and pruning spanning trees. *Journal of Machine Learning Research*, 20(85):1–35, 2019.

Geoffrey K Vallis. *Atmospheric and Oceanic Fluid Dynamics*. Cambridge University Press, 2017.

Sumio Watanabe. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(12), 2010.

Rebecca Willett, Robert Nowak, and Rui M Castro. Faster rates in regression via active learning. In *Advances in Neural Information Processing Systems*, pages 179–186, 2006.

Yuhong Wu, Håkon Tjelmeland, and Mike West. Bayesian CART: Prior specification and posterior simulation. *Journal of Computational and Graphical Statistics*, 16(1):44–66, 2007.

Yun Yang, Martin J Wainwright, and Michael Jordan. On the computational complexity of high-dimensional Bayesian variable selection. *The Annals of Statistics*, 44(6):2497–2532, 2016.

Charles T Zahn. Graph theoretical methods for detecting and describing gestalt clusters. *IEEE Trans. Comput.*, 20(SLAC-PUB-0672-REV):68, 1970.

Linlin Zhang, Michele Guindani, Francesco Versace, and Marina Vannucci. A spatio-temporal nonparametric Bayesian variable selection model of fMRI data for clustering correlated time courses. *NeuroImage*, 95:162–175, 2014.

Quan Zhou and Yongtao Guan. Fast model-fitting of Bayesian variable selection regression using the iterative complex factorization algorithm. *Bayesian Analysis*, 14(2):573, 2019.