Hierarchical Gaussian Processes for Spatially Dependent Model Selection

James T. Fry

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Scotland C. Leman, Chair
Robert B. Gramacy
Eric P. Smith
Lynn M. Resler

June 11, 2018
Blacksburg, Virginia

Keywords: spatial statistics, Gaussian process, model selection, nonstationary processes, Ising distribution, conditional autoregressive model

Hierarchical Gaussian Processes for Spatially Dependent Model Selection

James T. Fry

(ABSTRACT)

In this dissertation, we develop a model selection and estimation methodology for nonstationary spatial fields. Large, spatially correlated data often cover a vast geographical area. However, local spatial regions may have different mean and covariance structures. Our methodology accomplishes three goals: (1) cluster locations into small regions with distinct, stationary models, (2) perform Bayesian model selection within each cluster, and (3) correlate the model selection and estimation in nearby clusters. We utilize the Conditional Autoregressive (CAR) model and Ising distribution to provide intra-cluster correlation on the linear effects and model inclusion indicators, while modeling inter-cluster correlation with separate Gaussian processes. We apply our model selection methodology to a dataset involving the prediction of Brook trout presence in subwatersheds across Pennsylvania. We find that our methodology outperforms the stationary spatial model and that different regions in Pennsylvania are governed by separate Gaussian process regression models.

Hierarchical Gaussian Processes for Spatially Dependent Model Selection

James T. Fry

(GENERAL AUDIENCE ABSTRACT)

In this dissertation, we develop a statistical methodology for analyzing data where observations are related to each other due to spatial proximity. Our overall goal is to determine which attributes are important when predicting the response of interest. However, the effect and importance of an attribute may differ depending on the spatial location of the observation. Our methodology accomplishes three goals: (1) partition the observations into small spatial regions, (2) determine which attributes are important within each region, and (3) enforce that the importance of variables should be similar in regions that are near each other. We apply our technique to a dataset involving the prediction of Brook trout presence in subwatersheds across Pennsylvania.

Dedication

For Lauren Marie, who permitted me to upend our entire existence to pursue my calling.

Acknowledgments

I do not think it is possible to completely enumerate those who have helped me over these last five years. First, I must thank my advisor, Scotland Leman, for his guidance, optimistic outlook, and infinite flexibility. Upon my move to Harrisonburg after my third year, Scotland continued his support and direction as I worked remotely. I would also like to thank my committee, Bobby Gramacy, Eric Smith, and Lynn Resler for their part in my success and development as a statistician. I must also thank Jeff Birch, who took a chance on admitting me into the program, while other universities passed.

Countless friends, both within and outside of the statistics department, have been instrumental to my achievements and happiness. Matt Slifko, Justin Loda, and Tom Metzger for our "research" meetings, frequently held at the Cellar. Nathan Wycoff and Sierra Merkes for keeping the office atmosphere light. And one additional "thank you" to Matt Slifko, for giving me a place to stay while in Blacksburg.

To my family in both Pittsburgh and Richmond for their unwavering support, without which I would not have made the decision to return to graduate school. Finally, to my supportive wife, Lauren Marie, whose enduring patience allowed us to relocate from Hartford to Blacksburg to pursue my dream. I cannot wait to begin the next phase of our life together in Lewisburg, Pennsylvania.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Statistical modeling of all types require the consideration of levels of dependency. Even simple linear regression explicitly models the response as a linear function of the covariate. As data collection mechanisms and scientific questions become increasingly complex, additional modeling dependencies need to be considered. There is a vast literature of approaches for data that are temporally (Prado & West 2010) and spatially (Banerjee et al. 2014) dependent. Seemingly disparate models that rely on similar data sources induce a different dependency that can be handled through model fusion (Hoegh et al. 2015). In this thesis, we build a methodology that enforces spatial dependence both among observations, as well as spatial dependence in model selection and estimation that occurs in different spatial regions. We first begin by discussing the scientific application that motivates the methodology, as well as review some previous attempts the statistical modeling.

The East Brook Trout Joint Venture (EBTJV) is a partnership including wildlife agencies, federal resource agencies, and academic institutions. The EBTJV is one of the first Fish Habitat Partnerships to assist with implementing the National Fish Habitat Action Plan, which strives to address the loss and degradation of fish habitats (*East Brook Trout Joint Venture* 2018). Originally formed in 2004, the EBTJV's purpose is to develop and implement strategies for the conservation of brook trout. Hudy et al. (2008) analyzed and classified

11, 754 subwatersheds based on a 1969 map of the brook trouts' native distribution in the eastern United States (Figure 1.2). Using historical evidence and quantitative databases for these locations, each subwatershed was classified according the brook trout population descriptions in Table 1.1.

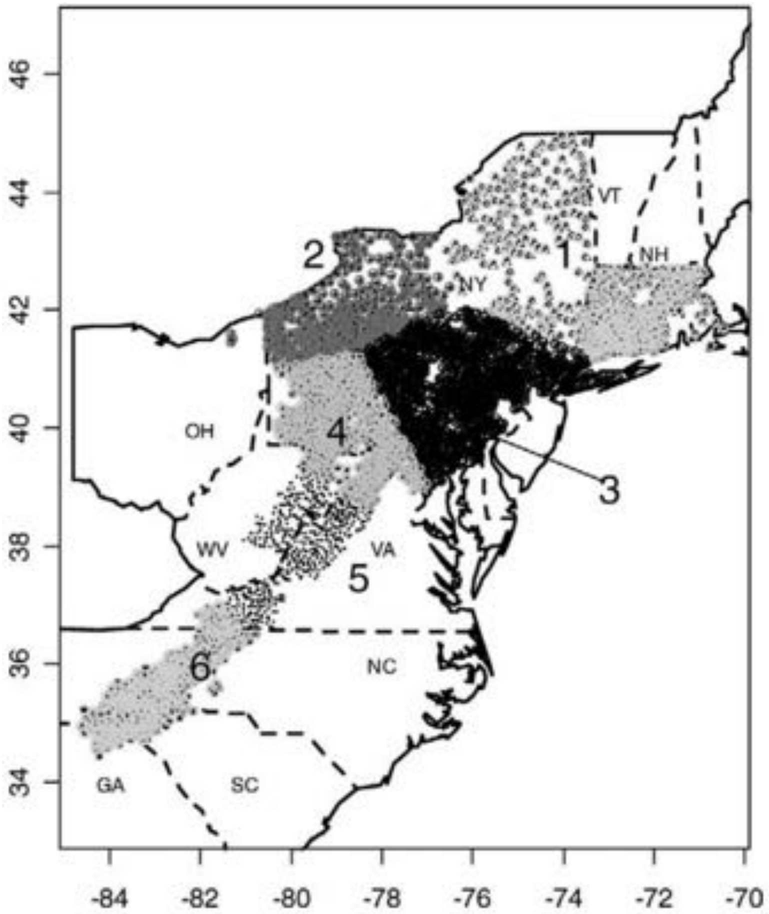


Figure 1.1: Voronoi tessellation of the spatial regions created by the EBTJV subwatershed locations. The methodology partitioned the space into 6 regions, each with its own logistic regression model.

Table 1.1: Classifications for each subwatershed and the associated description.

Classification	**Description**
Extirpated	All self-sustaining populations no longer exist.
Predicted extirpated	All self-sustaining populations are predicted to be extirpated.
Reduced	Of the historical habitat, $50-99\%$ no longer support self-sustaining populations.
Predicted reduced	Of the historical habitat, $50-99\%$ are predicted to no longer support self-sustaining populations.
Intact	Of the historical habitat, over 50% currently support self-sustaining populations.
Predicted intact	Of the historical habitat, over 50% are predicted to currently support self-sustaining populations.
Absent-unknown	Brook trout currently absent; historical status is unknown.
Present-qualitative	No quantitative data exist, or quantitative data are older than 10 years; available qualitative data show self-sustaining populations are present.
Unknown	No data are available, or there are not enough data to classify the subwatershed into any other category.

To classify each subwatershed as a binary response (presence/absence of brook trout), the categories are combined as: Present = Reduced + Intact, Absent = Extirpated. All other categories are not considered. Along with each classification, subwatersheds also have associated covariates, such as minimum elevation, spatial coordinates, road density, and percentage of various land coverages. Thieling (2006) thoroughly screened the covariates and tested them for completeness, range, redundancy, and importance to the response, using methods such as regression trees and logistic regression. To model the presence/absence of trout with spatially varying models, Zhang et al. (2008) partitioned the 2-dimension space using a Voronoi tessellation and fit local logistic regression models. This allows for each covariate to have a different effect on the mean response depending on the region. However, the tessellated regions were not jointly inferred with the parameters. These tessellated regions can be seen in Figure 1.1.

Motivated by the work of Zhang et al. (2008), Velasco-Cruz (2012) extended the method-

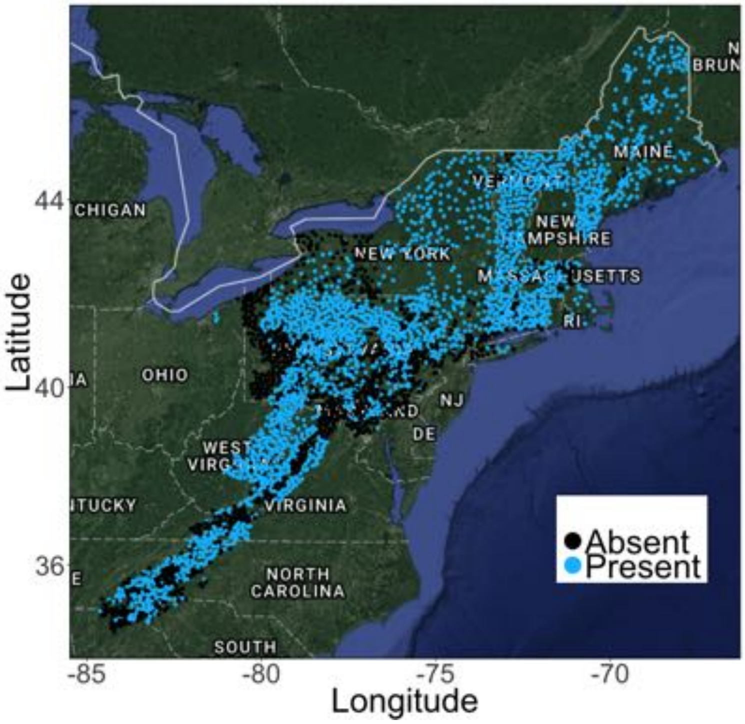


Figure 1.2: Map of the native regions for brook trout, stretching from Georgia to Maine. Points represent each subwatershed (Blue = trout present, Black = trout absent).

ology by not only using a Voronoi tessellation on the spatial coordinates but by building in inter-regional correlation. Using a Bayesian model selection framework, Velasco-Cruz used the Ising distribution (Section 2.4.1)to allow model selection in one region to affect the model selection in an adjacent region. The resulting tessellated regions look similar to that of Zhang et al. (2008). However, the regions are jointly inferred with the other parameters and the model selection within each cluster is performed differently. More details on Bayesian model selection and the Ising distribution is discussed in Sections 2.1.3 and 2.4.1. In Chapter 3, we give a thorough discussion of the Hierarchical Gaussian Process for Spatially Dependent Model Selection (HGP) methodology, which extends the methodology of Velasco-Cruz (2012).

1.1 Overview of method

In this work, we develop a methodology for spatial data where the correlation among observations and importance/effect of each covariate depends on location. To accomplish this, we form spatially contiguous clusters of locations to represent different spatial regions. Within these clusters, we fit a spatial model and perform variable selection to determine the importance of each covariate. Finally, we provide two mechanisms to allow the sharing of information between the spatial clusters: 1) if a covariate is included in the model of one cluster, it is more likely to be included in the model of a nearby cluster, 2) if a covariate is included in two nearby clusters, the effect should be similar in each. Upon application of our methodology, we learn the importance and effect of each covariate and how each changes throughout the spatial region.

1.2 Thesis overview

We present a methodology for variable selection in nonstationary (mean and covariance) spatial settings. The layout of this thesis is as follows. Chapter 2 contains a literature review that addresses popular techniques for variable selection, spatial modeling, and binary random fields. In Chapter 3, we fully address the methodology with the model statement and the specification of prior distributions for parameters. Chapter 4 addresses the computational concerns and implementation of the method. Chapter 5 contains simulation studies to analyze the performance of our methodology in a controlled setting. In Chapter 6, we apply our methodology to the EBTJV data and discuss the resulting inference. Chapter 7 contains final thoughts, conclusions, and future research.

Chapter 2

Literature review

The HGP methodology focuses on model selection and estimation in the presence of a nonstationary spatial field. Before addressing the details of the methodology, we review the literature associated with the three primary components: 1) model selection, 2) modeling spatially correlated data (both stationary and nonstationary), and 3) modeling spatially correlated binary data. Each section includes both common approaches, as well as techniques that are incorporated within the HGP methodology.

2.1 Model selection methods

Variable selection is adopted to choose which covariates best explain the variability in the response. Model selection methods help prevent the selection of models that overfit the data. An overfit model has too many inappropriate covariates and leads to more uncertainty in parameter estimation and poor out-of-sample prediction. Adding increasingly more parameters to a model causes the fitted line to more closely follow the fitted data, however at the cost of predicting new data (Figures 2.1a and 2.1b). Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots \boldsymbol{x}_n)'$ be an $n \times p$ matrix of n observations of p covariates and $\boldsymbol{y} = (y_1, \ldots, y_n)'$ be the corresponding $n \times 1$ vector of responses. We center and scale $\boldsymbol{X}$ and center $\boldsymbol{y}$ such that $\sum_{i=1}^{n} x_{ij} = 0$, $\sum_{i=1}^{n} x_{ij}^2 = 1$, and $\sum_{i=1}^{n} y_i = 0$. We denote the mean response

$$\mathbb{E}\big[\boldsymbol{y}|\boldsymbol{X}\big] = g(\mu\mathbf{1} + \tilde{\boldsymbol{X}}\boldsymbol{\beta}),$$

where μ is a scalar intercept, $\tilde{\boldsymbol{X}}$ is an $n \times q$ matrix that includes a subset of the columns from $\boldsymbol{X}$, $\boldsymbol{\beta}$ is a vector of linear effects, and $g(\cdot)$ is a link function that maps $\mathbb{R}$ to the domain of $\mathbb{E}\big[\boldsymbol{y}|\boldsymbol{X}\big]$. While the model selection techniques discussed within this section can be used for many link functions, we will assume that $g(\cdot)$ is the identity function for ease of notation. Searching over all possible models to find the optimum q covariates requires fitting 2^p models, which is infeasible for even small values of p. Many statistical techniques exist to find the best (depending on various criterion) subset of covariates without overfitting the data. We group the techniques into three categories:

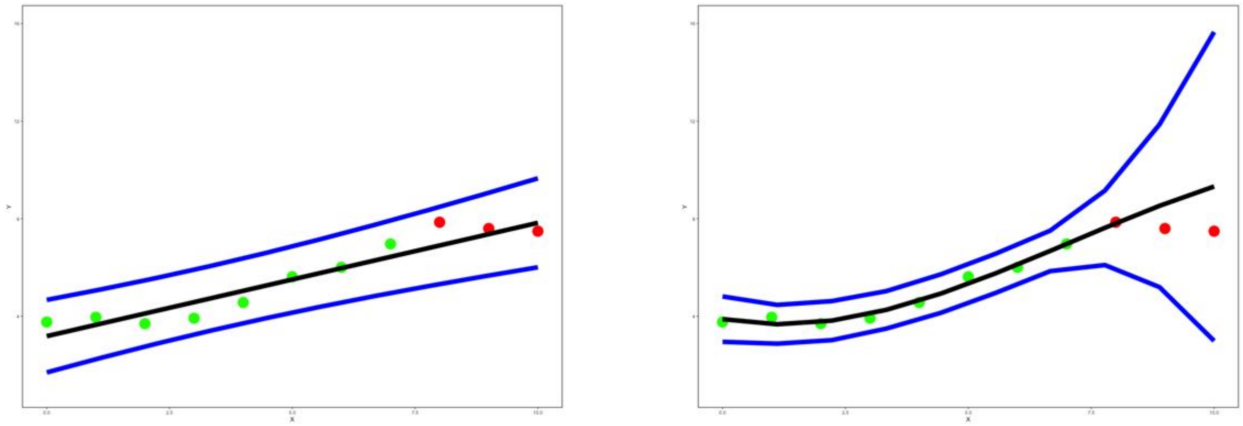
1. *Fixed-penalty likelihood methods* fit a model using maximum likelihood estimation and then evaluate the quality as a function of how well the model fits the data and model complexity.

2. *Regularization methods* constrain the parameter estimation to drive certain parameter estimates towards 0, which provides variable selection.

3. *Bayesian model selection methods* use specific prior distributions to infer which covariates should be included in a model.

2.1.1 Fixed-penalty likelihood methods

We will supplement the notation in Section 2.1 by defining the sampling distribution of $\boldsymbol{y}$, $f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta})$. Conditional on values for the unknown vector $\boldsymbol{\beta}$, we calculate the density for $\boldsymbol{y}$ (or probability, if $\boldsymbol{y}$ is discrete). Upon observing $\boldsymbol{y}$, the likelihood function (Casella & Berger



(a) First-order polynomial

(b) Third-order polynomial

Figure 2.1: A first-order polynomial ($R^2 = .85$) and third-order polynomial are fit to the green points ($R^2 = .97$). While the third-order model fits better, the prediction on the out of sample points (red) is poor. Blue lines represent 95% prediction intervals.

2002) for $\boldsymbol{\beta}$ is

$$L(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}).$$

That is, the likelihood of $\boldsymbol{\beta}$ is written as the sampling distribution of $\boldsymbol{y}$, but we treat it as a function of $\boldsymbol{\beta}$. The following procedures find estimates for $\boldsymbol{\beta}$ that maximizes $L(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{y})$, called maximum likelihood estimation (Casella & Berger 2002), and then penalizes the model based upon the total number of parameters. The penalty is dependent upon the method; we discuss two methods, Akaike Information Criterion and Bayesian Information Criterion.

Akaike Information Criterion

Akaike (1974) introduced the Akaike Information Criterion (AIC) to measure the relative quality of a regression model. Let $\tilde{\boldsymbol{X}}_i$ be the matrix of covariates for model i with q covariates and let $\boldsymbol{\beta}_i$ denote the associated unknown vector of regression coefficients. First, we estimate

$\boldsymbol{\beta}_i$ with $\hat{\boldsymbol{\beta}}_i = \underset{\boldsymbol{\beta}_i}{\text{ArgMax}}\ L(\boldsymbol{\beta}_i|\tilde{\boldsymbol{X}}_i, \boldsymbol{y})$. Then we compute AIC for model i as

$$AIC(\boldsymbol{\beta}_i) = -2Log(L(\hat{\boldsymbol{\beta}}_i|\tilde{\boldsymbol{X}}_i, \boldsymbol{y})) + 2q.$$

AIC combines the likelihood function with a penalty for model size. The penalty is a linear function of the number of parameters included. Upon fitting a collection of different models, we select those with the smallest values for AIC. Since $-2Log(L(\hat{\boldsymbol{\beta}}_i|\tilde{\boldsymbol{X}}_i, \boldsymbol{y}))$ never increases as the dimensionality of $\tilde{\boldsymbol{X}}_i$ increases (McCullagh & Nelder 1989), AIC penalizes complex models by adding $2q$ to discourage overfitting the data.

AIC is derived from determining the information lost by estimating a true generating process $\boldsymbol{y} \sim g(\boldsymbol{y})$ by a parametric model $f(\boldsymbol{y}|\boldsymbol{\theta})$. The information lost can be written as

$$D_{KL}\big(g(\boldsymbol{y})||f(\boldsymbol{y}|\boldsymbol{\theta})\big) = \int g(\boldsymbol{y})Log(f(\boldsymbol{y}|\boldsymbol{\theta}))d\boldsymbol{y},$$

where D_{KL} is the Kullback-Leibler divergence (Kullback & Leibler 1951). If the true model is of the form $g(\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta}_0)$ for some $\boldsymbol{\theta}_0$, then choosing the model with the smallest value of AIC is equivalent to finding the model that minimizes the Kullback-Leiber divergence (Claeskens et al. 2008).

Bayesian Information Criterion

The Bayesian Information Criterion (BIC), or Schwarz Criterion (Schwarz et al. 1978), is another metric for measuring the relative quality of a regression model. Like AIC, BIC finds parameter estimates, $\hat{\boldsymbol{\beta}}_i$ that maximizes the likelihood for a given set of covariates, $\tilde{\boldsymbol{X}}_i$. However, instead of introducing a penalty of $2q$ for a model with q covariates, BIC penalizes based on q and the number of observations n. It is calculated as

$$BIC(\boldsymbol{\beta}_i) = -2Log(L(\hat{\boldsymbol{\beta}}_i|\tilde{\boldsymbol{X}}_i, \boldsymbol{y})) + qLog(n).$$

Within a collection of models, we would select those with the smallest BIC values. BIC also has a connection to the Bayes factor, which will be discussed in Section 2.1.3. As with AIC, BIC only provides a selection device among the subset of models. When p is large, BIC is limited by the number of models the user is willing to separately estimate. To circumvent this limitation, regularization methods simultaneously fit the model to the data while shrinking some parameter estimates towards 0.

2.1.2 Regularization methods

AIC and BIC provide a method for model selection only after a sequence of models have already been estimated using maximum likelihood estimation. However, when p grows large, it is desirable to utilize methods that induce sparsity in $\boldsymbol{X}$ without having to specify which models to explore. We discuss three related approaches that regularize $\boldsymbol{\beta}$ by driving some elements towards 0.

Ridge regression

Originally developed to provide stability within the covariance matrix of the Ordinary Least-Squares (OLS) estimator for $\boldsymbol{\beta}$, Hoerl & Kennard (1970) proposed ridge regression. Before describing ridge regression, we briefly discuss OLS regression. Given our matrix of covariates $\boldsymbol{X}$ and vector of responses $\boldsymbol{y}$, to estimate the unknown parameter vector $\boldsymbol{\beta}$, OLS finds $\hat{\boldsymbol{\beta}}$ that minimizes the sum of squared errors (Schabenberger & Pierce 2001)

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{ArgMin}}\ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{2.1}$$

Equation 2.1 can be solved analytically and expressed in closed-form as $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$.

Ridge regression also minimizes the sum of squared errors but adds a constraint on the possible solutions for $\hat{\boldsymbol{\beta}}$. The new optimization problem becomes

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{ArgMin}}\ (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2,$$

where $\lambda > 0$ is a pre-specified penalty. Once λ is fixed, the constraint prevents parameter estimates from growing too large, thus stabilizing $Var(\hat{\beta}_j)$. Consequently, by increasing λ, the estimates for β_j are also driven towards 0, providing information as to which covariates have minimal effects on $\boldsymbol{y}$. As with OLS, ridge regression results in a closed-form solution: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$. Various procedures exist to choose an appropriate value for λ, but cross-validation is a standard choice (Golub et al. 1979). The ridge estimate also has a connection with Bayesian regression analysis when using a Normal likelihood and conjugate prior distribution. Under the model specification

$$\begin{aligned} \boldsymbol{y} &\sim N(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}), \\ \boldsymbol{\beta} &\sim N(\boldsymbol{0}, \sigma^2/\lambda \boldsymbol{I}), \end{aligned}$$

the posterior mean (and mode) of $\boldsymbol{\beta}$ is $\mathbb{E}\big[\boldsymbol{\beta}|\boldsymbol{y}\big] = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$. The quadratic constraint on β_j in the ridge optimization is enforced using the conjugate Normal prior distribution for each β_j.

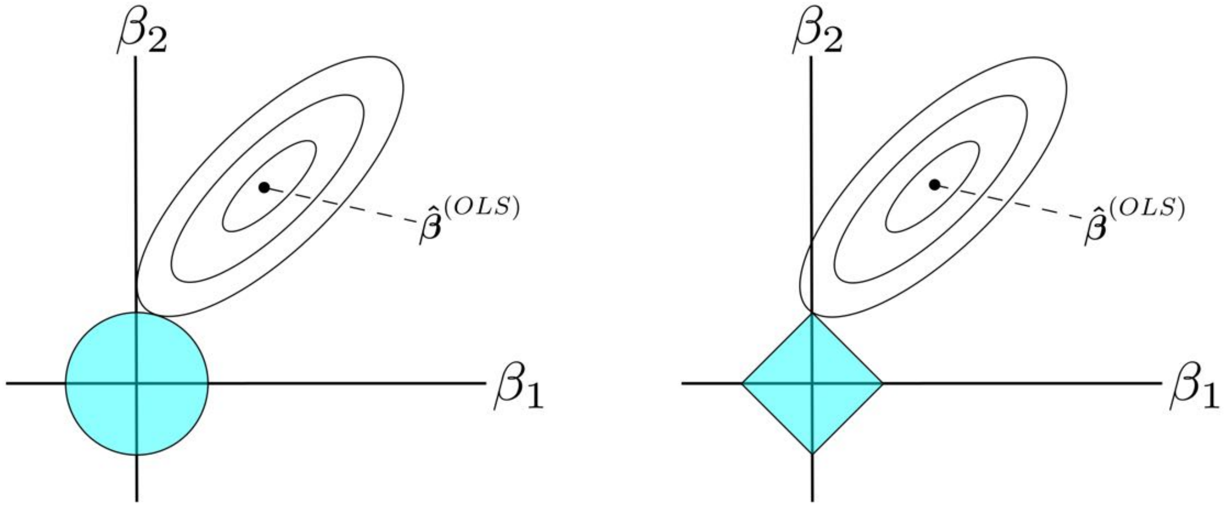
While the ridge penalty drives $\hat{\beta}_j$ towards 0, it does not provide actual estimates of 0. Rather, ridge regression will often result in estimates such that $\hat{\beta}_j \approx 0$. This behavior follows from the squared nature of the penalty term. In order to obtain actual coefficient estimates of 0, another penalty is necessary.

Least Absolute Shrinkage and Selection Operator

The Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani 1996) modifies ridge regression by replacing the squared-penalty on β_j with an L_1-norm penalty. LASSO solves the optimization

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{ArgMin}}\ (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|.$$

As with ridge regression, λ needs to be pre-specified and is typically calibrated using cross-validation. Unlike ridge regression, however, as λ increases, some of the estimated values $\hat{\beta}_j$ will be exactly 0, implying that these covariates should not be included in the model. This behavior is easily seen using a two-dimensional visualization of the optimization surface (Figures 2.2a and 2.2b). The interpretability of certain $\beta_j = 0$, as opposed to $\beta_j \approx 0$, is useful for variable selection. Thus, LASSO provides a selected list of covariates that should be removed from the model.


β_2
$\hat{\boldsymbol{\beta}}^{(OLS)}$
β_1
β_2
$\hat{\boldsymbol{\beta}}^{(OLS)}$
β_1

(a) Ridge

(b) LASSO

Figure 2.2: The optimization surface and constraint region (blue) for $\boldsymbol{\beta} = (\beta_1, \beta_2)'$. The diamond-shape of the LASSO penalty allows for estimates to be exactly 0.

Much like ridge regression, LASSO also has an equivalent Bayesian representation. Park & Casella (2008) prove that the LASSO estimates can be replicated by using a Normal likelihood and an independent Laplace prior distribution on each β_j. That is,

$$\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}),$$
$$P(\boldsymbol{\beta}) = \prod_{j=1}^{p} \lambda e^{-\lambda|\beta_j|/\sigma}/\sigma.$$

Under this model, the posterior model for $\boldsymbol{\beta}$ will match the LASSO estimates of Tibshirani (1996). Figure 2.3 shows the priors distributions for both ridge regression and LASSO. LASSO has a sharp, non-differentiable point at 0, while the ridge prior is much smoother.

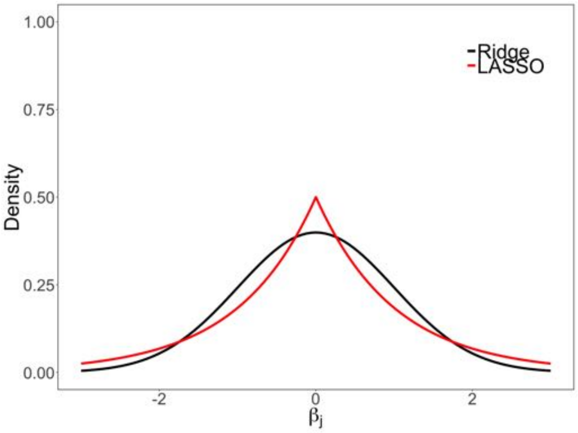


Figure 2.3: Prior distributions for LASSO (red) and ridge regression (black).

Elastic Net

Although LASSO provides variable selection by assigning estimated coefficient values to exactly 0, it has its limitations. For instance, when there is high correlation between covariates, the predictive performance of ridge regression is better than LASSO (Zou & Hastie 2005). Consequently, Zou & Hastie (2005) developed a hybrid approach, called the Elastic Net, that utilizes both L_1 and L_2 penalties simultaneously by solving

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{ArgMin}}\ (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j|.$$

Letting $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$, this is equivalent to the optimization problem

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{ArgMin}}\ (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) \text{ subject to } (1-\alpha)\sum_{j=1}^{p} |\beta_j| + \alpha \sum_{j=1}^{p} \beta_j^2 \leq t$$

for some t. The resulting constraint region is a hybrid of both ridge regression and LASSO, with rounder sides than the LASSO but sharper corner than ridge regression (Figure 2.4). Both ridge regression and LASSO are special cases where $\alpha = 1$ and $\alpha = 0$, respectively. The Bayesian counterpart to Elastic Net uses a Normal likelihood with a prior distribution that uses both L_1 and L_2 norms (Li et al. 2010). It is written as

$$\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}),$$
$$P(\boldsymbol{\beta}) \propto e^{-(\lambda_1 ||\boldsymbol{\beta}||_1 + \lambda_2 ||\boldsymbol{\beta}||_2)/(2\sigma^2)}.$$

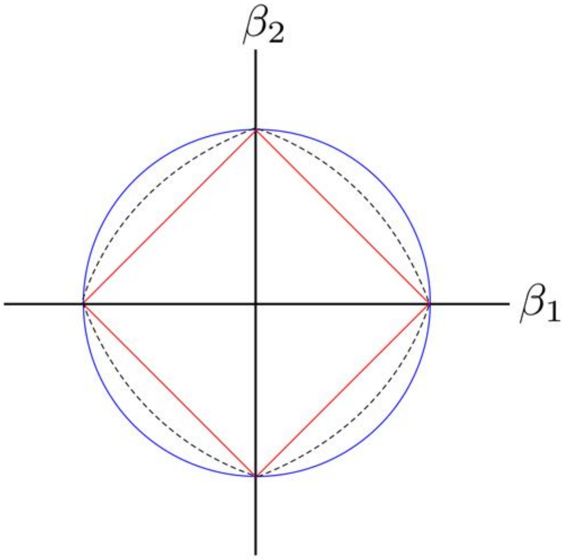


Figure 2.4: The constraint region for ridge regression (blue), LASSO (red), and Elastic Net (black).

2.1.3 Bayesian model selection

While ridge regression and LASSO were originally proposed outside of a Bayesian framework, the results for each could be replicated through specification of certain prior distributions. The following methods, however, are specific to the Bayesian paradigm and require prior distributions to allow posterior inference of which models best explain the variability in $\boldsymbol{y}$.

Stochastic Search Variable Selection (SSVS)

George & McCulloch (1993) built a scale mixture of Normal distributions as a prior distribution for β_j. That is,

$$\begin{aligned} \boldsymbol{y} &\sim Normal(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}), \\ \beta_j|\gamma_j &\sim (1-\gamma_j)Normal(0,\tau_j^2) + \gamma_j Normal(0, c_j^2\tau_j^2), \\ \boldsymbol{\gamma} &\sim \pi(\boldsymbol{\gamma}), \end{aligned}$$

where τ_j^2 set to be small, that if $\gamma_j = 0$, we could be quite sure that β_j is small enough that it could be safely removed from the model. In contrast, c_j is set to be very large, so that if $\gamma_j = 1$, β_j should likely be included in the model (Figure 2.5a). Of utmost interest is $\mathbb{E}\big[\gamma_j|\boldsymbol{y}\big]$, the posterior marginal probability that β_j should be included in the model. Unlike the fixed penalty and regularization methods, SSVS provides a probability, given the data, that each covariate should be included in the model.

One challenge arises from choosing the prior distribution for $\boldsymbol{\gamma}$ and values for the hyperparameters τ_j^2 and c_j^2. Often, each γ_j is treated as an independent variable, implying $P(\gamma_j = 1) = .5$ for $j = 1, \ldots, p$, or $P(\boldsymbol{\gamma}) = 1/2^p$. This choice encourages models that have $\approx p/2$ parameters. Other options involve penalizing the model complexity such as $P(\boldsymbol{\gamma}) \propto 1/(\sum_j^p \gamma_j)$. For the hyperparameters, the authors recommend setting $3\tau_j$ to the maximum size at which β_j could practically be 0, although this would involve understanding the possible effect of β_j. George & McCulloch (1993) suggest some semi-automatic methods for tuning c_j and τ_j^2, but these will not be discussed further.

Geweke (1996) replaces the Normal mixture prior distribution with a mixture of a Normal distribution with a point-mass at 0. The prior distribution can be written as

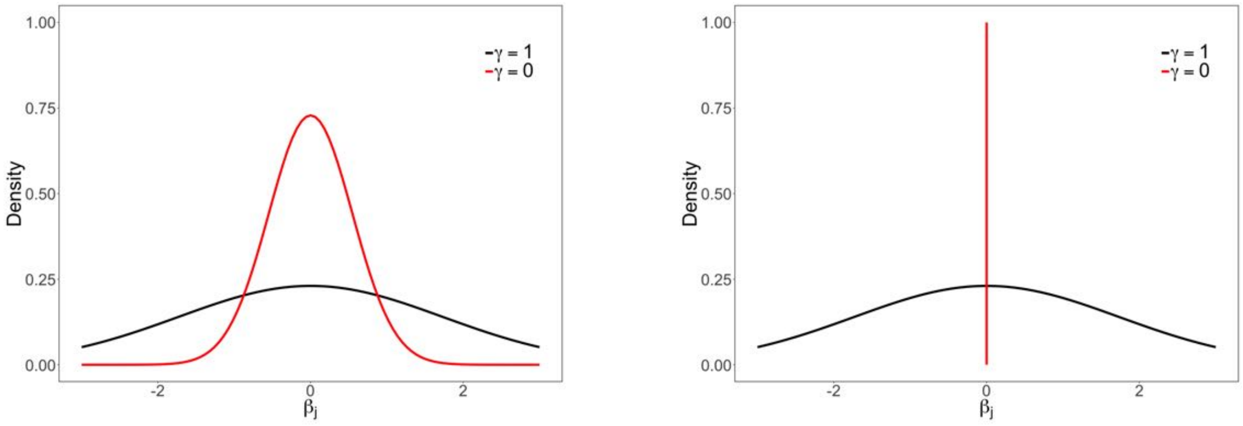
$$\beta_j|\gamma_j \sim (1-\gamma_j)\delta(\beta_j = 0) + \gamma_j Normal(0, \sigma_\beta^2),$$

where $\delta(\beta_j = 0)$ is the Dirac delta function, which places density of 1 at $\beta_j = 0$ and a density of 0 for $\beta_j \neq 0$ (Figure 2.5b). As a result, there is only one hyperparameter, σ_β^2, that needs to be specified. This parameter controls how large the effect of β_j needs to be in order to be considered nonzero. However, this parameter needs to be selected carefully, as it can greatly influence the posterior distribution.

A major advantage to SSVS is the simplicity of the implementation. The posterior distribution of the variable inclusion indicators and linear effects, (γ_j, β_j), can be sampled from the joint full conditional distribution using a Gibbs sampler. This allows an easy way to analyze the posterior distribution of the model space. Reversible jump Markov chain Monte Carlo (RJMCMC) (Green 1995) is required in many other settings to explore the posterior distribution of a model space, but we omit the details here.

Horseshoe estimator

Carvalho et al. (2010) proposed a prior distribution for situations where $\boldsymbol{\beta}$ is expected to be sparse. The horseshoe estimator arises from multivariate-Normal scale mixtures. The full



(a) George & McCulloch (1993)

(b) Geweke (1996)

Figure 2.5: The prior distribution for β_j under the models of George & McCulloch (1993) and Geweke (1996). Hyperparameters were selected to be $c_j^2 = 0.1$, $\tau_j^2 = 3$, and $\sigma_\beta^2 = 3$.

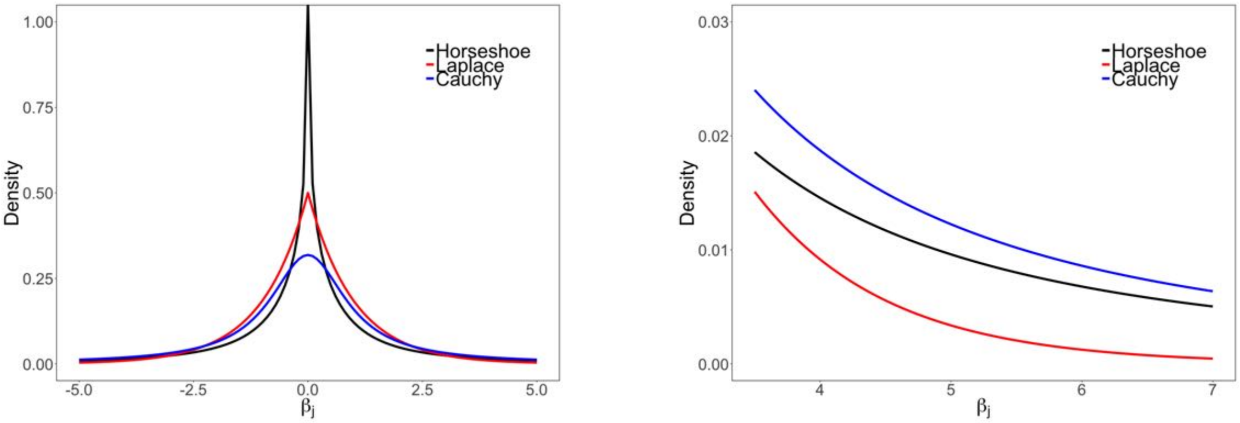
model is specified as

$$\begin{aligned} \boldsymbol{y} &\sim Normal(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}), \\ \beta_j|\lambda_j &\sim Normal(0, \lambda_j^2) \\ \lambda_j|\tau &\sim Cauchy^+(0, \tau) \\ \tau|\sigma &\sim Cauchy^+(0, \sigma), \end{aligned}$$

where $Cauchy^+(0, a)$ is a half-Cauchy distribution with scale parameter a. A priori, each β_j comes from Normal distribution, much like ridge regression. However, instead of tuning λ_j based on cross validation, it has its own prior distribution: a half-Cauchy distribution with a common scale parameter τ. Further, the scale parameter τ also comes from a half-Cauchy distribution. While the analytic form of the horseshoe prior, $P(\beta_j)$, is unavailable, the density is bounded such that

$$\begin{cases} \beta_j = 0, & P(\beta_j) \to \infty, \\ \beta_j \neq 0, & \frac{K}{2} Log(1 + 4\beta_j^{-2}) < P(\beta_j) < K Log(1 + 2\beta_j^{-2}), \end{cases}$$

where $K = (2\pi^3)^{-1/2}$. The horseshoe prior has a pole at $\beta_j = 0$ and heavy tails. This prior distribution places more mass at 0 than both the Laplace (used in Bayesian LASSO) and Cauchy distributions (Figure 2.6a). However, the tail mass is between that of the Laplace and Cauchy distributions (Figure 2.6b). While this prior distribution shrinks $\boldsymbol{\beta}$, a transdimensional proposal (and RJMCMC) is necessary to obtain actual variable selection.



(a) Center of distribution

(b) Tail of distribution

Figure 2.6: Comparison of the Horseshoe (black), Laplace (red), and Cauchy (blue) distributions.

Bayes factor

The Bayes factor was developed by Harold Jeffreys (Jeffreys 1935, 1961), and independently by Alan Turing in 1940 (Good 1983) to quantify the evidence in favor of a hypothesis. Consider some data $\boldsymbol{D}$ that is assumed to have been generated under one of two competing hypotheses, H_1 and H_2, by probability distributions $P(\boldsymbol{D}|H_1)$ and $P(\boldsymbol{D}|H_2)$, respectively. Then by way of Bayes theorem,

$$P(H_k|\boldsymbol{D}) = \frac{P(\boldsymbol{D}|H_k)P(H_k)}{P(\boldsymbol{D}|H_1)P(H_1) + P(\boldsymbol{D}|H_2)P(H_2)}$$

for $k = 1, 2$ and therefore,

$$\frac{P(H_1|\boldsymbol{D})}{P(H_2|\boldsymbol{D})} = \frac{P(\boldsymbol{D}|H_1)}{P(\boldsymbol{D}|H_2)} \frac{P(H_1)}{P(H_2)},$$

where $P(H_1|\boldsymbol{D})/P(H_2|\boldsymbol{D})$ is the posterior odds, $P(H_1)/P(H_2)$ is the prior odds, and $P(\boldsymbol{D}|H_1)/P(\boldsymbol{D}|H_2)$ is the Bayes factor. The Bayes factor is the multiplicative adjustment to the prior odds of H_1 to obtain the posterior odds of H_1 and, thus represents the evidence provided by the data (Kass & Raftery 1995).

Using the same logic for model selection, we can treat the model specification as a hypothesis in order to use a Bayes factor to compare models. Consider two models for the data $\boldsymbol{y}$, $\mathcal{M}_1$ and $\mathcal{M}_2$ with parameters vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively. Then the Bayes factor in support of $\mathcal{M}_1$ is

$$BF_{1,2} = \frac{\int_{\boldsymbol{\theta}_1} P(\boldsymbol{y}|\mathcal{M}_1, \boldsymbol{\theta}_1) P(\boldsymbol{\theta_1}|\mathcal{M}_1) d\boldsymbol{\theta}_1}{\int_{\boldsymbol{\theta}_2} P(\boldsymbol{y}|\mathcal{M}_2, \boldsymbol{\theta}_2) P(\boldsymbol{\theta_2}|\mathcal{M}_2) d\boldsymbol{\theta}_2}.$$

Since the Bayes factor is a Bayesian method of addressing hypothesis testing, one must specify a proper prior distribution $P(\boldsymbol{\theta}_k|\mathcal{M}_k)$. That is,

$$\int_{\boldsymbol{\theta}_k} P(\boldsymbol{\theta}_k|\mathcal{M}_k) d\boldsymbol{\theta}_k = 1.$$

Establishing a proper prior distribution may be difficult depending on the context. Choosing an improper prior distribution for $\boldsymbol{\theta}_k$ means that its distribution is only known up to an undefined multiplicative constant. Since the constant would be arbitrary, the resulting Bayes factor would also be arbitrary. Therefore, improper prior distributions should not be used with Bayes factors.

The Bayes factor has a special relationship to BIC. Let $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ be the maximum likelihood estimates under $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. Also, let $\boldsymbol{I}(\boldsymbol{\beta}_1)$ and $\boldsymbol{I}(\boldsymbol{\beta}_2)$ denote the expected information matrix for one observation:

$$\boldsymbol{I}(\boldsymbol{\beta}_k) = \mathbb{E}\left[\left(\frac{\partial}{\partial\boldsymbol{\beta}_k} Log\big(P(y|\boldsymbol{\beta}_k, \mathcal{M}_k)\big)\right)^2 \middle| \boldsymbol{\beta}_k\right]$$

If we specify the prior distribution $\boldsymbol{\beta}_k \sim N(\hat{\boldsymbol{\beta}}_k, I(\boldsymbol{\beta}_k))$ for $k = 1, 2$, then $-2Log(BF_{1,2}) \approx BIC_1 - BIC_2$ (Raftery 1995).

2.2 Random fields

A random field is a stochastic process that is indexed by a spatial variable. As a result, there is a dependent covariance structure between spatial locations. Within this framework, we denote the response vector as $\boldsymbol{y} = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))'$, where $\boldsymbol{s}_l$ indexes the coordinates for y_l. The covariance for $\boldsymbol{y}$, $\boldsymbol{\Sigma}$, is typically of the form

$$\boldsymbol{\Sigma} = \sigma^2\boldsymbol{\rho}(\boldsymbol{D}|\boldsymbol{\theta}) + \tau^2\boldsymbol{I},$$

where $\boldsymbol{D}$ is a symmetric matrix of pairwise distances $d_{l,l'}$, between $\boldsymbol{s}_l$ and $\boldsymbol{s}_{l'}$, and $\boldsymbol{\rho}(\cdot)$ is a correlation function that depends on $\boldsymbol{D}$ and hyperparameters $\boldsymbol{\theta}$. Common choices for $d_{l,l'}$ are Euclidean distance and Geodesic distance, which are very similar when spatial sites are in a local area (Figure 2.7). $\boldsymbol{\theta}$ varies depending on the choice of correlation function. σ^2 and and τ^2 are often referred to as the partial sill and nugget, respectively. Since $\boldsymbol{\rho}$ is a correlation matrix, each element is between -1 and 1. σ^2 converts $\boldsymbol{\rho}$ into a covariance matrix with unbounded elements and τ^2 then further adjusts the diagonal elements. Each element within $\boldsymbol{\Sigma}$ can be summarized as:

$$\boldsymbol{\Sigma}_{l,l'} = \begin{cases} \sigma^2\rho(d_{l,l'}|\boldsymbol{\theta}), & d_{l,l'} > 0, \\ \sigma^2 + \tau^2, & d_{l,l'} = 0. \end{cases}$$

The selection of the correlation function $\boldsymbol{\rho}(\boldsymbol{D}|\boldsymbol{\theta})$ depends on the application and many

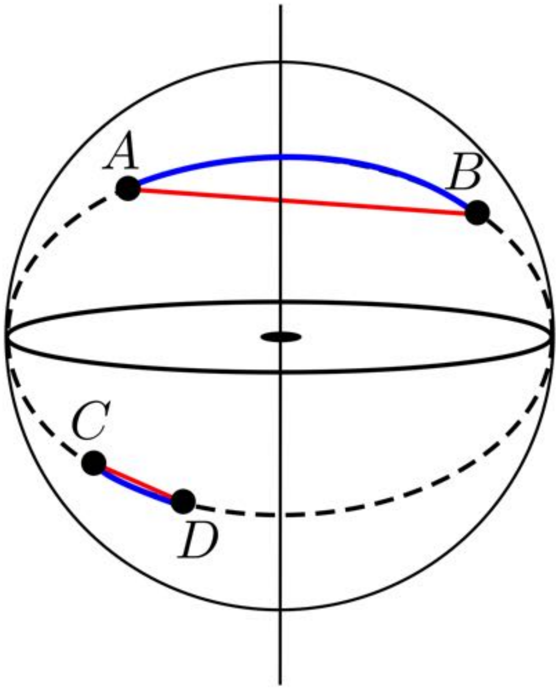


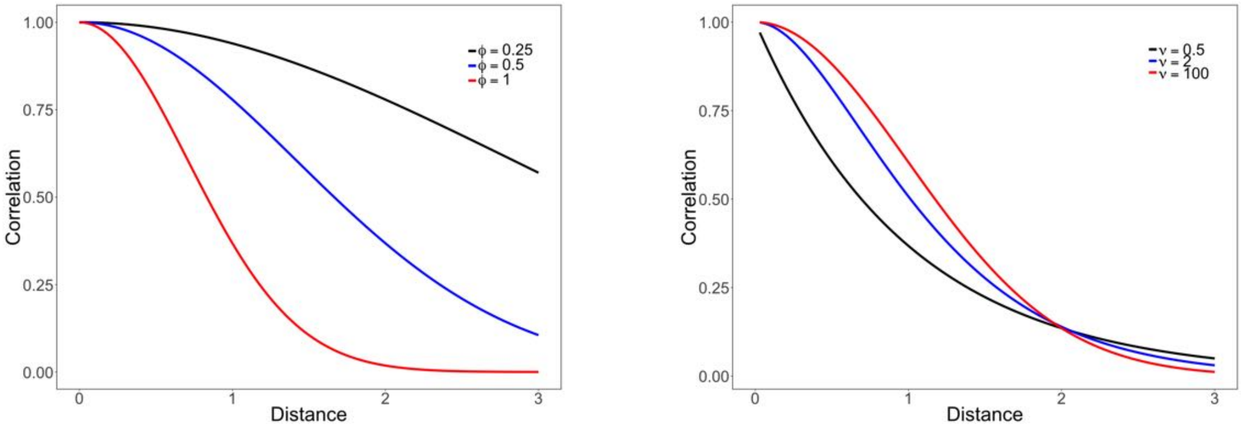
Figure 2.7: Examples of Geodesic (blue) and Euclidean (red) distances. The distances are noticeably difference between points A and B, since they are far apart on the surface of sphere. However, the two distances are almost equal for points C and D, which are near each other.

options exist. Several popular functions are listed in Table 2.1 (De Oliveira et al. 1997, Matérn 2013, Yaglom 2012).

Table 2.1: Correlation functions and the associated hyperparameters.

Function	$\boldsymbol{\theta}$	$\rho(d_{l,l'}\|\boldsymbol{\theta})$
Gaussian	$\{\phi\}$	$exp(-\phi^2 d_{l,l'}^2)$
Matérn	$\{\phi, \nu\}$	$(2^{\nu-1}\Gamma(\nu))^{-1}(\phi\sqrt{2\nu}d_{l,l'})^{\nu}K_{\nu}(\phi\sqrt{2\nu}d_{l,l'})$
Powered Exponential	$\{\phi, p\}$	$exp(-\|\phi d_{l,l'}\|^p)$
Rational Quadratic	$\{\phi\}$	$1 - d_{l,l'}/(\phi + d_{l,l'}^2)$

Each correlation function has its own benefits and fallbacks. For instance, the Matérn function allows the user to control the smoothness of the space via the hyperparameter ν; however the evaluation of the Bessel function K_ν can be time consuming. The Gaussian function is infinitely differentiable, which may not be ideal for some real applications but is fairly robust to misspecification. The powered exponential function is a generalization of the Gaussian function that allows users to control the smoothness by tuning p. When $v \to \infty$ for the Matérn function and $p = 2$ for the powered exponential function, each becomes a member of the Gaussian family. Figure 2.8 shows the theoretical correlation for $\rho(d_{l,l'}|\boldsymbol{\theta})$ as a function of distance for the Gaussian and Matérn correlation functions with various choices for hyperparameters. As ϕ increases, the Gaussian correlation decreases more quickly as distance increases. When $\phi = 0$, there is no spatial structure and each observation is independent. For the Matérn function, as ν approaches ∞, the curve shape more closely matches that of the Gaussian function.



(a) Gaussian function

(b) Matérn function

Figure 2.8: Theoretical correlation as a function of distance for various correlation functions.

2.2.1 Stationarity

In many applications, the spatial process is assumed to be stationary. A process is said to be strictly stationary if, for any n sites $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ and any $\boldsymbol{h} \in \mathbb{R}^r$ (where $r = \dim(\boldsymbol{s}_l)$), the joint distribution of $(y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))'$ is the same as the joint distribution of $(y(\boldsymbol{s}_1 + \boldsymbol{h}), \ldots, y(\boldsymbol{s}_n + \boldsymbol{h})))'$ (Banerjee et al. 2014). In other words, the joint distribution of $\boldsymbol{y}$ depends only on the relative distances between the observations and not on the actual locations. A less restrictive form of a stationary process is one that is said to be weakly stationary.

A spatial process is weakly stationary if the following conditions hold:

1. $\mathbb{E}\big[y(\boldsymbol{s})\big] = \mu$,
2. $Cov(y(\boldsymbol{s}), y(\boldsymbol{s}+\boldsymbol{h})) = C(\boldsymbol{h})$.

That is, the expected value of an observation at a given site is constant throughout the space and the covariance between observations in two locations is a function of only the separation vector $\boldsymbol{h}$. Strong stationarity always implies weak stationarity, but the converse is not always true. If the spatial process is Gaussian (i.e. $\boldsymbol{y}$ follows a multivariate Gaussian distribution), weak stationarity does imply strong stationarity. While the definition of weak stationarity requires two conditions, in spatial analyses, we are often only concerned with the stationarity of the covariance structure.

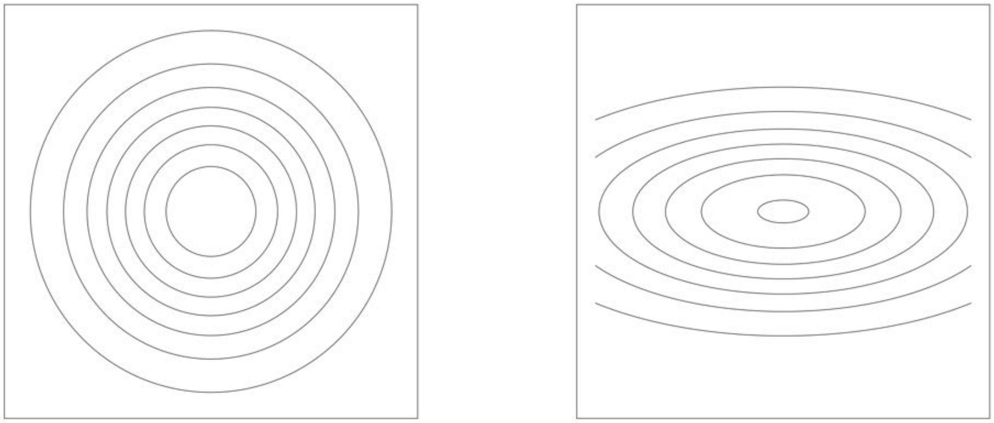
2.2.2 Isotropy

Another commonly assumed condition of a spatial process is that of isotropy. If the covariance function only depends on the Euclidean distance between observations and does not vary across space, it is said to be isotropic. Otherwise, it is called *anisotropic* (Cressie 2015). In practice, a correlation function that decays radially is often unrealistic. The association between sites may also depend on the direction. One example is the separable correlation function.

A separable correlation function can be written as

$$\rho(y(\boldsymbol{s}), y(\boldsymbol{s}+\boldsymbol{h})) = \prod_{i=1}^{r} \rho_i(h_i), \tag{2.2}$$

where each dimension in $\boldsymbol{h}$ has its own valid correlation function $\rho_k(\cdot)$ in $\mathbb{R}^1$. Since the separable correlation does not vary throughout space, it is stationary but not isotropic. Visually, the correlation decay resembles elongated contours with orientation that lies parallel to the axis of each dimension (Figure 2.9). As a further extension of the separable correlation function, geometric anisotropy is used when the correlation structure does not decay in a parallel manner to each axis (Figure 2.10). $\rho(\cdot)$ is defined as a function distance along some direction, given by



(a) Isotropic correlation

(b) Separable correlation

Figure 2.9: Correlation contours for two-dimensional isotropic and separable correlation functions.

$$\rho(y(\boldsymbol{s}), y(\boldsymbol{s}+\boldsymbol{h})) = \rho(\boldsymbol{h}'\boldsymbol{B}\boldsymbol{h}),$$

where $\boldsymbol{B}$ is a positive definite matrix. The elements in $\boldsymbol{B}$ determine the angle of rotation and the strength of the relationship between the axes. Methodology for the estimation of $\boldsymbol{B}$, by using a Wishart prior distribution, has been developed by Ecker & Gelfand (1999). The separable correlation function (Equation 2.2) can be thought as a geometric correlation function where $\boldsymbol{B}$ has a diagonal structure.

2.2.3 Nonstationary spatial processes

While the geometric correlation function provides additional flexibility beyond its isotropic counterpart, many natural settings are not only anisotropic but also nonstationary. For

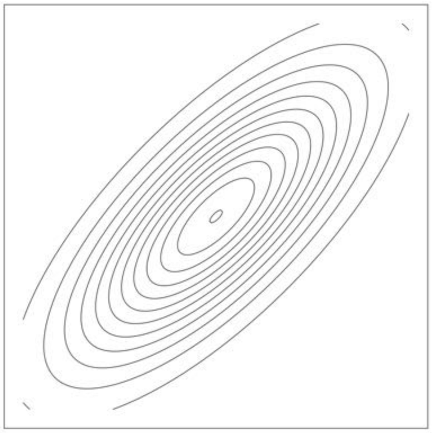


Figure 2.10: Contours for a geometric correlation function.

instance, natural boundaries such as mountains or valleys can create sharp breakpoints in the covariance structure. Or, the hyperparameters governing correlation may be dependent on location within the space. As nonstationarity plays an important role in the methodology of this thesis, the following subsections will describe other techniques for modeling nonstationary spatial processes.

Treed Gaussian process

The Treed Gaussian process (TGP) (Gramacy & Lee 2008) was originally developed for computer experiments applications. Although not explicitly spatial, the method can be used in nonstationary spatial settings. The TGP first fits a regression tree to partition the input space. Then, within each section, an independent Gaussian process is fit to the observations within the section. Using a Bayesian approach, Markov chain Monte Carlo (MCMC) is used to sample the parameters of each Gaussian process and the tree structure. Each iteration provides a nonstationary Gaussian process for the entire input space with sharp breaks between regions. However, a smoothed surface can be obtained by averaging over the tree space. Figure 2.11 shows the posterior mean of the process along with the posterior mode from the tree space. Since most of the input space is relatively flat, the TGP partitions in such a way that the nonlinear bottom-left is separate from the remaining space.

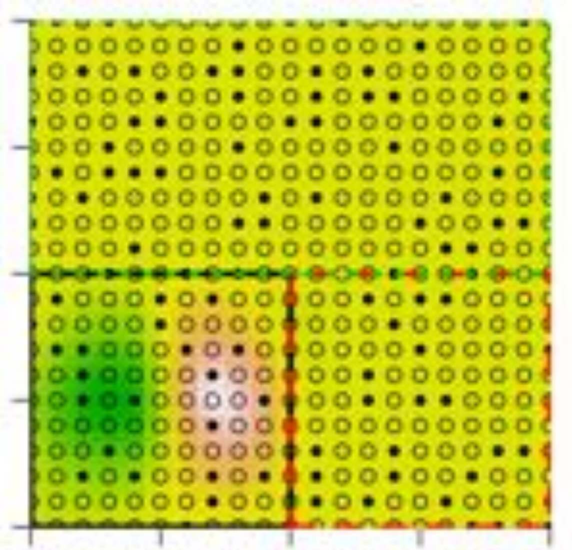


Figure 2.11: Posterior mean of the surface with an overlay of the posterior mode of the tree space.

Kernel convolution

A stationary Gaussian process $f(\boldsymbol{x})$ over a region $\boldsymbol{x} \in \mathcal{X}$ can be built by the convolution of a white noise process, $\epsilon(x)$ with a smoothing kernel, $K(\boldsymbol{x})$ (Higdon et al. 2002, Paciorek & Schervish 2006). For a typical spatial setting where $\boldsymbol{s} \in \mathbb{R}^2$, we write

$$y(\boldsymbol{s}) = \int_{\mathbb{R}^2} K(\boldsymbol{s} - \boldsymbol{u})\epsilon(\boldsymbol{u})d\boldsymbol{u}.$$

Further, the function K may vary for different locations $\boldsymbol{s}$. Consequently, the resulting convolution provides a nonstationary Gaussian process (Higdon 1998). By choosing a nonstationary Kernel function K, the convolution provides smoothing while also allowing different spatial dependence in various parts of the space.

Local approximate Gaussian process

Gramacy & Apley (2015) proposed the local approximation to the Gaussian process by conditioning only local input configurations. Rather than simply using the nearest neighbors to predict, the nearby points, called satellite points, are selected in a greedy, stepwise manner. If a prediction is to be made at location x, the n satellite points are selected by sequentially minimizing the empirical Bayes mean-squared prediction error at x. While originally developed for fast prediction, the local approximate Gaussian process also provides nonstationary modeling by allowing for different dependency at each prediction site. Figure 2.12 shows a two-dimensional example of the sequential design with $n = 50$ reference points.

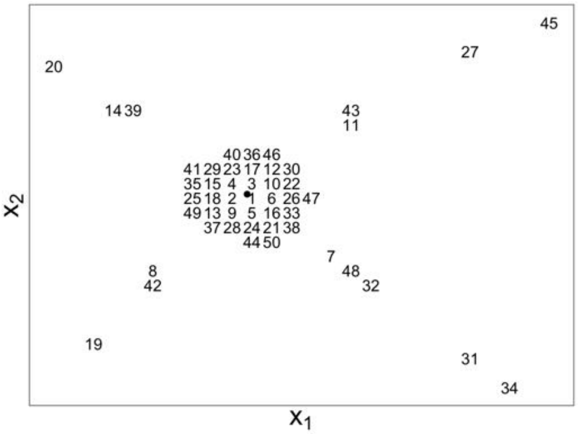


Figure 2.12: An example of order in which configurations, (x_1, x_2), were selected with a total of $n = 50$ satellite points to predict at the black point.

Piecewise Gaussian process

The piecewise Gaussian process (Kim et al. 2005) provides a framework for modeling nonstationary spatial processes. The method first partitions the entire two-dimensional space into R disjoint and independent regions. Within each individual region, a stationary and isotropic Gaussian process is fit using the observations contained within the boundary of the region. The partitioning is accomplished by using a Voronoi tessellation (Figure 2.13). Utilizing a Bayesian framework, the Gaussian process parameters are sampled via MCMC. Further, the number of regions R is also treated as unknown and is sampled using RJMCMC (Green 1995).

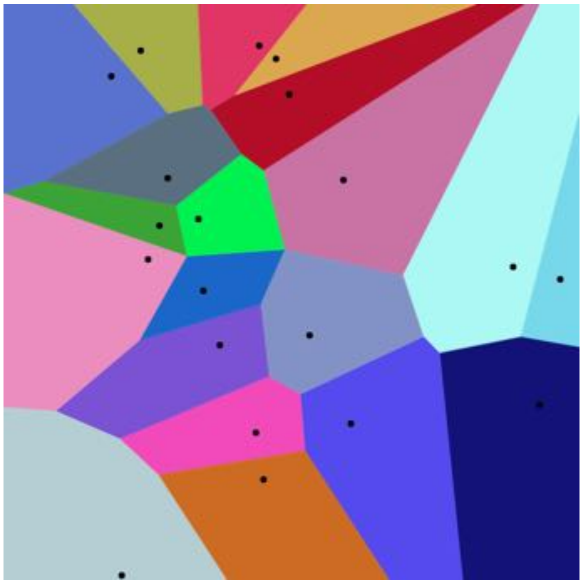


Figure 2.13: Example of a Voronoi tessellation with 20 regions.

2.3 Markov random fields

A more specific spatial process is the Markov random field (MRF). The unlike the random field discussed in Section 2.2, the MRF is based on an undirected graphical model. A random field is considered Markov if

$$P(y(\boldsymbol{s}_\ell)|y(\boldsymbol{s}_1),\ldots,y(\boldsymbol{s}_{\ell-1}),y(\boldsymbol{s}_{\ell+1}),\ldots,y(\boldsymbol{s}_n)) = P(y(\boldsymbol{s}_\ell)|y(\boldsymbol{s}_{\ell'}),\ell\sim\ell'),$$

where $\ell \sim \ell'$ denotes locations ℓ' that are first-order neighbors with location ℓ. If the response $y(\boldsymbol{s}_\ell)$ is Gaussian, the spatial process is called a Gaussian Markov random field (GMRF). Many techniques for analyzing GMRF data involve working with the inverse covariance matrix, rather than the covariance matrix. We briefly describe two commonly used models for these data.

2.3.1 Conditional autoregressive model

The conditional autoregressive (CAR) model was originally introduced by Besag (1974). While CAR models work for MRFs, Besag (1974) showed that it can be extended for higher-order relationships between locations. For the CAR model, the distribution is specified through the set of conditional distributions

$$(y(\boldsymbol{s}_\ell)|y(\boldsymbol{s}_{\ell'}), \ell \neq \ell') \sim Normal\Bigg(\mu_\ell + \sum_{\ell'=1}^{n} W_{\ell,\ell'}(\rho)(y(\boldsymbol{s}_{\ell'}) - \mu_{\ell'}), \sigma^2\Bigg),$$

where $W_{\ell,\ell'}$ are weights between locations ℓ and ℓ' with $W_{\ell,\ell} = 0$ for all ℓ. Combining the conditional distributions, we obtain the joint specification for $y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n)$,

$$\boldsymbol{y} \sim Normal(\boldsymbol{\mu}, \sigma^2(\boldsymbol{I} - \rho\boldsymbol{C})^{-1}).$$

$\boldsymbol{C}$ is a proximity matrix that is assumed to be known and fixed. Some common choices include an adjacency matrix ($C_{ij} = 1$ if regions i and j are neighbors) and a distance-based matrix (C_{ij} is inversely proportional to the distance between the centroids of i and j). ρ is often referred to as a spatial strength parameter. ρ needs to be carefully specified, however, to ensure that the covariance matrix is positive definite. Let $\lambda_1, \ldots, \lambda_n$ denote the sorted eigenvalues for $\boldsymbol{C}$. Since $Tr(\boldsymbol{C}) = 0$, $\lambda_1 < 0$ and $\lambda_n > 0$. For $\sigma^2(\boldsymbol{I} - \rho\boldsymbol{C})^{-1}$ to be positive definite, $\lambda_1^{-1} < \rho < \lambda_n^{-1}$ (Ren & Sun 2013).

CAR models have become increasingly popular in recent years. Parameter estimation for a Gaussian process model involves the inversion of the $n \times n$ covariance matrix, an $\mathcal{O}(n^3)$ operation. For the CAR model, the inverse covariance matrix is simply $(\boldsymbol{I} - \rho\boldsymbol{C})/\sigma^2$, thus avoiding a computationally demanding task.

2.3.2 Simultaneous autoregressive model

Rather than using the conditional distribution for $y(\boldsymbol{s}_\ell)$ to induce the joint distribution, the simultaneous autoregressive (SAR) model uses the error terms instead (Whittle 1954). The error for location ℓ is written as a function of the errors for adjacent locations:

$$e_\ell = \sum_{\ell'=1}^{n} b_{\ell,\ell'} e_{\ell'} + \epsilon_i$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)' \sim Normal(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ The resulting joint distribution for the responses becomes

$$\boldsymbol{y} \sim Normal(\boldsymbol{\mu}, \sigma^2 \left[(\boldsymbol{I} - \boldsymbol{B})(\boldsymbol{I} - \boldsymbol{B})'\right]^{-1}),$$

where $\boldsymbol{B}$ is a matrix of weights $b_{\ell,\ell'}$. A common choice for $\boldsymbol{B}$ is similar to that of the CAR model: $\boldsymbol{B} = \rho \boldsymbol{W}$, where ρ is again constrained by the eigenvalues of $\boldsymbol{W}$. As with the CAR model, computation of the likelihood function does not require inverting the $n \times n$ covariance matrix. The difference between the CAR and SAR models can be seen through the covariance. The two methods will be equivalent only if $\sigma^2(\boldsymbol{I} - \rho \boldsymbol{W})^{-1} = \sigma^2 \left[(\boldsymbol{I} - \boldsymbol{B})(\boldsymbol{I} - \boldsymbol{B})'\right]^{-1}$. Therefore, any SAR model can be rewritten as a CAR model, but the converse is not true.

2.4 Binary random fields

A more specific spatial process is the Binary Random Field (BRF). Unlike the spatial models of Section 2.2, which assume $y(\boldsymbol{s}) \in \mathbb{R}$, the response of the BRF, $z(\boldsymbol{s})$, can only take on two values. For convention, we will assume $z(\boldsymbol{s}) \in \{0, 1\}$. Unlike a Binomial distribution, the response at nearby sites are associated with each other. In many applications, the location $\boldsymbol{s}$ acts as the vertex on a lattice with edges connected to adjacent locations (Figure 2.14).

The BRF is called a Binary Markov Random Field (BMRF) when

$$P(z(\boldsymbol{s_l})|z(\boldsymbol{s_{l'}}), l \neq l') = P(z(\boldsymbol{s_l})|z(\boldsymbol{s_{l'}}, l \sim l'),$$

where $l \sim l'$ means that $\boldsymbol{s}_l$ is in the neighborhood of $\boldsymbol{s}_{l'}$. The distribution at location $\boldsymbol{s}_l$ is only dependent upon its neighbors. In the following subsections, we discuss some approaches to BRF modeling.

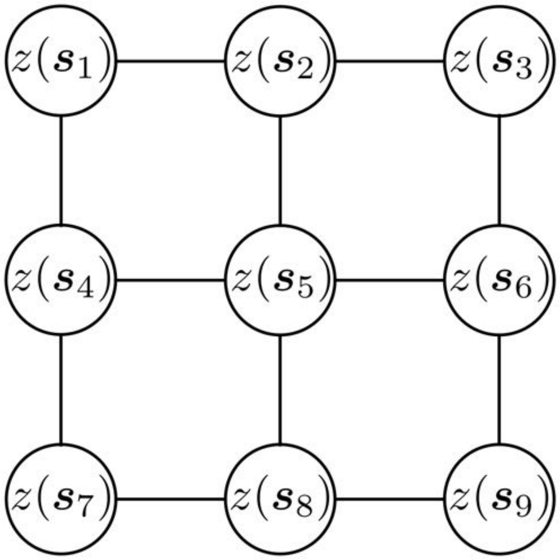


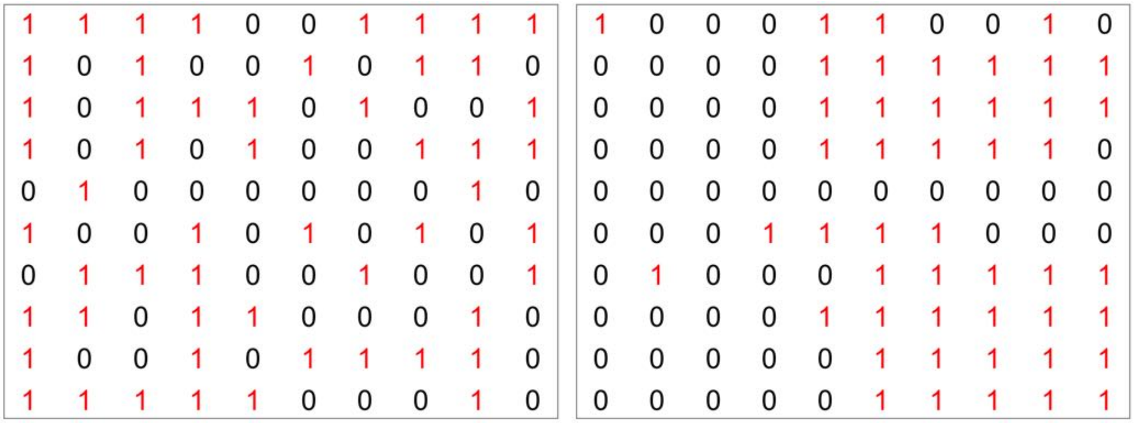
Figure 2.14: Example of a size $n = 9$ BMRF on a lattice. Vertices that are connected with an edge are associated.

2.4.1 Ising model

Let $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ be locations on a lattice where we observe binary outcomes $z(\boldsymbol{s}_1), \ldots, z(\boldsymbol{s}_n)$. Then $\boldsymbol{z} = (z(\boldsymbol{s}_1), \ldots, z(\boldsymbol{s}_n))'$ follows an Ising distribution if

$$P(\boldsymbol{z}) = \frac{1}{\Omega(\boldsymbol{\alpha}, \boldsymbol{\theta})} exp\bigg\{ \sum_{i=1}^{n} \alpha_i(z(\boldsymbol{s}_i)) + \sum_{i \sim j} \theta_{i,j} w_{i,j} \delta\big(z(\boldsymbol{s}_i) = z(\boldsymbol{s}_j)\big) \bigg\}, \tag{2.3}$$

where $i \sim j$ means that vertex $\boldsymbol{s}_i$ is adjacent to $\boldsymbol{s}_j$ and $w_{i,j}$ is a weight for the dependency between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. The parameters for the model are α_i for $i = 1, \ldots n$ and $\theta_{i,j}$ for $i \sim j$ (Higdon 1995). $\sum_{i=1}^{n} \alpha_i(z(\boldsymbol{s_i}))$ is called the external field and determines the probability distribution when no neighborhoods exist. The remaining portion of the exponent can be considered the interaction effect for neighboring sites. Often $\alpha_i(z(\boldsymbol{s}_i))$ is assumed to be linear and is fixed a priori, while $\theta_{i,j}$ is assumed to be θ for all i, j pairs. When $\theta = 0$ and $\alpha_i = \alpha$ $\forall i$, there is no spatial dependency and the joint distribution of the indicators is Binomial.



1	1	1	1	0	0	1	1	1	1
1	0	1	0	0	1	0	1	1	0
1	0	1	1	1	0	1	0	0	1
1	0	1	0	1	0	0	1	1	1
0	1	0	0	0	0	0	0	1	0
1	0	0	1	0	1	0	1	0	1
0	1	1	1	0	0	1	0	0	1
1	1	0	1	1	0	0	0	1	0
1	0	0	1	0	1	1	1	1	0
1	1	1	1	1	0	0	0	1	0

(a) Ising with $\theta = 0$

1	0	0	0	1	1	0	0	1	0
0	0	0	0	1	1	1	1	1	1
0	0	0	0	1	1	1	1	1	1
0	0	0	0	1	1	1	1	1	0
0	0	0	0	0	0	0	0	0	0
0	0	0	1	1	1	1	0	0	0
0	1	0	0	0	1	1	1	1	1
0	0	0	0	1	1	1	1	1	1
0	0	0	0	0	1	1	1	1	1
0	0	0	0	0	1	1	1	1	1

(b) Ising with $\theta = 3$

Figure 2.15: Realizations from the Ising model on a 10 × 10 grid with $\theta = 0$ and $\theta = 3$. When θ is larger, we observe fewer, larger clusters.

Much of the difficulty regarding parameter estimation stems from the normalizing constant $\Omega(\boldsymbol{\alpha}, \boldsymbol{\theta})$. Both maximum likelihood estimation and Bayesian methods require the computationally expensive task of evaluating $\Omega(\boldsymbol{\alpha}, \boldsymbol{\theta})$. Conditional on $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$, the normalizing constant requires the summation over all 2^n possible lattice layouts:

$$\Omega(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \sum_{\boldsymbol{z}} exp\Big\{ \sum_{i=1}^{n} \Big[\alpha_i(z(\boldsymbol{s}_i)) + \sum_{i \sim j} \theta_{i,j} w_{i,j} \delta\big(z(\boldsymbol{s}_i) = z(\boldsymbol{s}_j)\big) \Big] \Big\}. \tag{2.4}$$

Several methods exist to either estimate or circumvent the issue with the normalizing constant. We briefly discuss three approaches.

Pseudo-likelihood

Pseudo-likelihood functions are employed as an approximation of the actual likelihood function when it is too difficult to directly work with the likelihood. Therefore, one method to avoid computing the Ising normalizing constant (Equation 2.4) is to approximate the actual Ising likelihood function (Equation 2.3). Besag (1975) suggests the following:

$$L_n(\boldsymbol{\psi}) \approx \prod_{i=1}^{n} p_i(\boldsymbol{\psi}),$$

where $L_n(\boldsymbol{\psi})$ is the true likelihood with unknown parameters ,$\boldsymbol{\psi}$, and $p_i(\psi)$ is the sampling distribution of observation i, conditional on $\boldsymbol{\psi}$. Specifically for the Ising model, $z(\boldsymbol{s}_i) \sim Bernoulli(p_i)$ where p_i is the conditional probability that $z(\boldsymbol{s}_i) = 1$ given the neighbors of site i, given by

$$p_i = \left(1 + exp\left\{\alpha_i(z(\boldsymbol{s}_i)) + \sum_{i \sim j} \theta_{i,j} w_{i,j} \delta\big(z(\boldsymbol{s}_j) = 1\big)\right\}\right)^{-1}. \tag{2.5}$$

The structure of Equation 2.5 resembles that of the Logistic regression and there is no longer an intractable normalizing constant. Estimation can therefore proceed using standard Bayesian posterior analysis or maximum likelihood estimation. However, as θ becomes larger, this approximation becomes less reliable (Geyer & Thompson 1992).

Importance sampling

The most direct method of calculating the Ising normalizing constant (Equation 2.4) is to approximate it using importance sampling (**?**). Letting $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \boldsymbol{\theta})'$, we generate $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_B \sim f(\boldsymbol{x})$ and then calculate our estimate

$$\hat{\Omega}(\boldsymbol{\Theta}) \approx \frac{1}{B} \sum_{i=1}^{B} \frac{q_{\boldsymbol{\Theta}}(\boldsymbol{x}_i)}{f(\boldsymbol{x}_i)},$$

where $q_{\boldsymbol{\Theta}}(\boldsymbol{x}_i)$ is the unnormalized target distribution and the support of the distribution $q^*(\cdot)$ contains the support of $q_{\boldsymbol{\Theta}}(\cdot)$ (Gelman & Meng 1998). This approach requires that $g(\cdot)$ itself be normalized and performs best when $g(\cdot)$ is "close" to $q_{\boldsymbol{\Theta}}(\cdot)$. $g(\cdot)$ must also be entirely known, both the kernel and the normalizing constant. Choosing such a function for the Ising model can be difficult. A natural choice may be to simulate $\boldsymbol{x}_i \sim Ising(\tilde{\boldsymbol{\theta}})$, where $\tilde{\theta}$ is the maximum pseudo-likelihood estimate. However, while simulation from the Ising model is possible, we would not be able to evaluate the probability mass of the realization. This problem is avoided in situations when working with the ratio of normalizing constants, as opposed to the normalizing constant by itself.

While operating within a Bayesian framework with a prior distribution $\pi(\boldsymbol{\Theta})$, samples can be obtained from the posterior distribution using the Metropolis-Hastings algorithm (Hastings 1970). The acceptance ratio for a proposed transition from $\boldsymbol{\Theta}$ to $\boldsymbol{\Theta'}$ under transition kernel $g(\boldsymbol{\Theta'}|\boldsymbol{\Theta})$ is

$$H(\boldsymbol{\Theta'}) = \frac{\Omega(\boldsymbol{\Theta}) q_{\boldsymbol{\Theta'}}(\boldsymbol{z}) \pi(\boldsymbol{\Theta'}) g(\boldsymbol{\Theta}|\boldsymbol{\Theta'})}{\Omega(\boldsymbol{\Theta'}) q_{\boldsymbol{\Theta}}(\boldsymbol{z}) \pi(\boldsymbol{\Theta}) g(\boldsymbol{\Theta'}|\boldsymbol{\Theta})},$$

which includes the intractable normalizing constants under both sets of parameters $\boldsymbol{\Theta'}$ and $\boldsymbol{\Theta}$. The ratio $\Omega(\boldsymbol{\Theta})/\Omega(\boldsymbol{\Theta'})$, however, can be approximated by using importance sampling. Suppose we simulate $\boldsymbol{x}_i \sim f(\cdot)$ for $i = 1, \ldots, B$. Using importance sampling for both $\boldsymbol{\Theta'}$ and $\boldsymbol{\Theta}$, the ratio $\Omega(\boldsymbol{\Theta})/\Omega(\boldsymbol{\Theta'})$ becomes

$$\frac{\Omega(\boldsymbol{\Theta})}{\Omega(\boldsymbol{\Theta'})} = \frac{\frac{1}{B}\sum_{i=1}^{B} q_{\boldsymbol{\Theta}}(\boldsymbol{x}_i)/g(\boldsymbol{x}_i)}{\frac{1}{B}\sum_{i=1}^{B} q_{\boldsymbol{\Theta'}}(\boldsymbol{x}_i)/g(\boldsymbol{x}_i)}.$$

Again, choosing $f(\cdot)$ to be $Ising(\tilde{\boldsymbol{\Theta}})$, where $\tilde{\boldsymbol{\Theta}}$ is the maximum pseudo-likelihood estimate, both the numerator and denominator separately cannot be evaluated. However, the intractable normalizing constant $\Omega(\tilde{\boldsymbol{\Theta}})$ under $f(\cdot)$ cancels when computing the ratio. The ratio reduces to

$$\frac{\Omega(\boldsymbol{\Theta})}{\Omega(\boldsymbol{\Theta'})} = \frac{\frac{1}{B}\sum_{i=1}^{B} q_{\boldsymbol{\Theta}}(\boldsymbol{x}_i)/q_{\tilde{\boldsymbol{\Theta}}}(\boldsymbol{x}_i)}{\frac{1}{B}\sum_{i=1}^{B} q_{\boldsymbol{\Theta'}}(\boldsymbol{x}_i)/q_{\tilde{\boldsymbol{\Theta}}}(\boldsymbol{x}_i)},$$

where $q_{\tilde{\boldsymbol{\Theta}}}(\cdot)$ is the kernel of the Ising distribution under the maximum pseudo-likelihood estimate $\tilde{\boldsymbol{\Theta}}$. Performing this approximation within each iteration of Markov Chain Monte Carlo can be a computational burden, as simulating from an Ising distribution can be time-consuming. Møller proposed another method to avoid this simulation by augmenting the sampling of $\boldsymbol{\Theta}$ with an auxiliary variable.

Data augmentation

Møller et al. (2006) proposed a method for sampling from the posterior distribution of the Ising model. As with maximum likelihood estimation, sampling from the posterior distribution $P(\boldsymbol{\alpha}, \boldsymbol{\theta}|\boldsymbol{z})$ requires computing the normalizing constant. To sample from the posterior distribution of $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \boldsymbol{\theta})'$, the authors augment the actual Metropolis-Hastings (Hastings 1970) acceptance ratio by introducing an auxiliary variable x with conditional distribution $f(x|\boldsymbol{\Theta}, \boldsymbol{z})$. Then the Metropolis-Hastings acceptance ratio for the proposal $(\boldsymbol{\Theta}, x) \rightarrow (\boldsymbol{\Theta}', x')$ is

$$H(\boldsymbol{\Theta}', x') = \frac{f(x'|\boldsymbol{\Theta}', \boldsymbol{z})\pi(\boldsymbol{\Theta}')q_{\boldsymbol{\Theta}'}(\boldsymbol{z})q_{\boldsymbol{\Theta}}(x)g(\boldsymbol{\Theta}|\boldsymbol{\Theta}', x')}{f(x|\boldsymbol{\Theta}, \boldsymbol{z})\pi(\boldsymbol{\Theta})q_{\boldsymbol{\Theta}}(\boldsymbol{z})q_{\boldsymbol{\Theta}'}(x')g(\boldsymbol{\Theta}'|\boldsymbol{\Theta}, x)},$$

where $\pi(\cdot)$ is the prior distribution for $\boldsymbol{\Theta}$, $g(\cdot)$ is the proposal distribution, and $q(\cdot)$ is the target distribution without the normalizing constant. Once converged to the stationary distribution $P(\boldsymbol{\Theta}, x|\boldsymbol{z})$, the desired distribution $P(\boldsymbol{\Theta}|\boldsymbol{z}) = \int_x P(\boldsymbol{\Theta}, x|\boldsymbol{z})dx$ is trivial. As established by Møller., the technique appears similar to approximating the ratio of the normalizing constants $\Omega(\boldsymbol{\Theta})/\Omega(\boldsymbol{\Theta}')$ with two simple-sample importance sampling estimates within each iteration of MCMC.

For the Ising distribution, the authors recommend symmetric proposal distributions $g(\boldsymbol{\Theta}|\boldsymbol{\Theta'}) = g(\boldsymbol{\Theta}'|\boldsymbol{\Theta})$, flat prior distribution $\pi(\boldsymbol{\Theta}) = \delta(\boldsymbol{\Theta} \in [\boldsymbol{\Theta}_{min}, \boldsymbol{\Theta}_{max}])$, and auxiliary distribution $f(x|\boldsymbol{\Theta}') = Ising(\tilde{\boldsymbol{\Theta}})$, where $\tilde{\boldsymbol{\Theta}}$ is the maximum pseudo-likelihood estimate.

With these choices, the Metropolis-Hastings acceptance ratio reduces to

$$H(\boldsymbol{\Theta}', x') = \frac{q_{\tilde{\boldsymbol{\Theta}}}(\boldsymbol{x}')q_{\boldsymbol{\Theta}'}(\boldsymbol{y})q_{\boldsymbol{\Theta}}(\boldsymbol{x})}{q_{\tilde{\boldsymbol{\Theta}}}(\boldsymbol{x})q_{\boldsymbol{\Theta}}(\boldsymbol{y})q_{\boldsymbol{\Theta}'}(\boldsymbol{x}')}\delta(\boldsymbol{\Theta} \in [\boldsymbol{\Theta}_{min}, \boldsymbol{\Theta}_{max}]).$$

This ratio only involves evaluating the kernel of the Ising distribution for the auxiliary variable and observed data under the current, proposed, and pseudo-likelihood parameters. The normalizing constants in the numerator and denominator cancel. While similar to the importance sampling method, there is an importance difference. The importance sampler approximates the normalizing constant via Monte Carlo, where the approximation improves as more samples are collected. Møller's auxiliary variable method does not approximate the normalizing constant, but rather draws jointly from the posterior distribution of $\boldsymbol{\Theta}$ and $\boldsymbol{x}$.

2.4.2 Clipped Gaussian process

Another interpretation of a BRF is to assume that it is actually Gaussian random field that has been clipped at a fixed value ξ (De Oliveira 2000). Without the clipping, the analysis would simply be the standard Gaussian process utilized in many spatial analyses. The model is specified as follows:

$$z(\boldsymbol{s}_l) = \begin{cases} 1, & y(\boldsymbol{s}_l) \geq \xi, \\ 0, & y(\boldsymbol{s}_l) < \xi, \end{cases}$$

$$\boldsymbol{y} \sim Gaussian(\boldsymbol{X\beta}, \boldsymbol{\Sigma}),$$

where $\Sigma_{l,l'} = \sigma^2\rho(z(\boldsymbol{s}_l), z(\boldsymbol{s}_{l'})|\boldsymbol{\theta})$ describes the covariance of the unobservable latent process at locations $\boldsymbol{s}_l$ and $\boldsymbol{s}_{l'}$. The likelihood of the unknown parameters is then written as

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2, \xi|\boldsymbol{z}) = \int_{A(z(\boldsymbol{s_1}))} \cdots \int_{A(z(\boldsymbol{s_n}))} (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2}$$
$$exp\Big\{ -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta})\Big\}dy(\boldsymbol{s}_1)\cdots dy(\boldsymbol{s}_n),$$

where

$$A(z(\boldsymbol{s_l})) = \begin{cases} (-\infty, \xi], & z(\boldsymbol{s}_l) = 0, \\ [\xi, \infty), & z(\boldsymbol{s}_l) = 1. \end{cases}$$

Under this full model, the parameters are not identifiable, so it is convention to fix $\sigma^2 = 1$ and $\xi = 0$. Then, by specifying prior distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, the posterior distribution can be easily sampled using MCMC with the data augmentation strategy of Albert & Chib (1993).

Chapter 3

Methodology

In this chapter, we present the model statement, prior distributions, and constraints that collectively comprise the Hierarchical Gaussian Processes for Spatially Dependent Model Selection (HGP) methodology. We assign each spatial location into a cluster (as defined by the convex hull of the member coordinates). Within each cluster, we fit a stationary Gaussian process regression to the associated observations. We use an Ising distribution on the variable inclusion indicators across clusters to provide spatial correlation in the model selection. Similarly, for the nonzero linear effects, we use a CAR model for spatial smoothing. With these components, HGP provides nonstationary spatial modeling with spatially varying model selection and estimation.

3.1 Preliminary concepts and notation

3.1.1 Gaussian process

For overall consistency, we introduce the notation that will be utilized throughout the HGP methodology. We let $\boldsymbol{y} = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))'$ denote the $n \times 1$ response vector from a Gaussian process observed at locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n \in \mathbb{R}^2$. Further, we let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ be the $n \times p$

matrix of covariates with rows that correspond to locations $\boldsymbol{s}_1, \dots, \boldsymbol{s}_n$. Then, we write the distribution of $\boldsymbol{y}$ as

$$\boldsymbol{y} \sim Normal(\mu\mathbf{1} + \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}),$$

where μ is a scalar intercept applicable to each location, $\boldsymbol{\beta}$ is the $p \times 1$ vector of linear effects, and $\boldsymbol{\Sigma}$ is the covariance matrix. We will let $\boldsymbol{\Sigma}$ take on a reparameterized form of a Gaussian process covariance

$$\boldsymbol{\Sigma} = \tau^2\big(\boldsymbol{\rho}(\boldsymbol{D}|\phi) + g\boldsymbol{I}\big),$$

where $\boldsymbol{D}$ is the $n \times n$ symmetric matrix of pairwise Euclidean distances $d_{l,l'}$ and $\rho(d_{l,l'}) = exp(-d_{l,l'}^2/\phi)$, the Gaussian correlation function. Next, we generalize the single Gaussian process to a mixture of Gaussian processes.

3.1.2 Mixture of Gaussian processes

We assume that rather than observing one realization of a Gaussian process, we observe R realizations. The realizations are the results of separate processes in each of $r = 1, \dots, R$ clusters. We define $\boldsymbol{c} = (c_1, \dots, c_n)'$ to be indicators for cluster membership. Therefore, we say that $\boldsymbol{y}^{(r)} = \{y(\boldsymbol{s}_\ell) : c_\ell = r\}$. Each local Gaussian process has its own distinct mean and covariance structure. That is,

$$\begin{aligned}
\boldsymbol{y}^{(1)} &\sim Normal(\mu^{(1)}\mathbf{1} + \boldsymbol{X}^{(1)}\boldsymbol{\beta}^{(1)}, \boldsymbol{\Sigma}^{(1)}),\\
\boldsymbol{y}^{(2)} &\sim Normal(\mu^{(2)}\mathbf{1} + \boldsymbol{X}^{(2)}\boldsymbol{\beta}^{(2)}, \boldsymbol{\Sigma}^{(2)}),\\
&\vdots\\
\boldsymbol{y}^{(R)} &\sim Normal(\mu^{(R)}\mathbf{1} + \boldsymbol{X}^{(R)}\boldsymbol{\beta}^{(R)}, \boldsymbol{\Sigma}^{(R)}),
\end{aligned}$$

where $\boldsymbol{y}^{(r)}$ and $\boldsymbol{X}^{(r)}$ are of size $n_r \times 1$ and $n_r \times q_r$, respectively, and $\sum_{i=1}^{R} n_i = n$. $\Sigma^{(r)}$ is the $n_r \times n_r$ covariance matrix for $\boldsymbol{y}^{(r)}$ and involves scale, lengthscale, and nugget parameters

$\tau^{(r)}$, $\phi^{(r)}$, and $g^{(r)}$, respectively. The formulation implies that each vector $\boldsymbol{y}^{(r)}$ has its own set of covariates and covariance structure, and therefore, $\boldsymbol{y}^{(r)}$ and $\boldsymbol{y}^{(r')}$ are conditionally independent (given $\boldsymbol{c}$) for $r \neq r'$. Therefore, there exists clusters $1, \ldots R$ such that, within each cluster, there is an independent Gaussian process model.

Each group of observations $\boldsymbol{y}^{(r)}$ jointly follow a Gaussian process. However, we also want Gaussian process to be spatially distinct. Therefore, in Section 3.1.3 we discuss how we spatially partition our set of all observations. To begin, we first briefly review some geometric topics.

3.1.3 Convex hulls

Before directly utilizing convex hulls in our model statement, we briefly review. Let $\{p_1, p_2, \ldots, p_n\}$ be d-dimensional points in $\mathbb{R}^d$. Then, the convex hull of $\{p_1, \ldots, p_n\}$ in the Euclidean plane, denoted as $\text{Conv}(\{p_1, \ldots, p_n\})$ is the smallest convex set in $\mathbb{R}^d$ that contains $\{p_1, \ldots, p_n\}$. For our spatial setting, where coordinates exist in two dimensions, the convex hull can be defined more intuitively. The convex hull is the smallest polygon that contains all points $\{p_1, \ldots, p_n\}$, as well as all line segments connecting any two points p_i and p_j for $i = 1, \ldots, n$ and $j = 1, \ldots, n$. For example, consider the set of points $\{p_1, \ldots, p_6\}$ in Figure 3.1. The convex hull for this set is $\{p_1, \ldots, p_5\}$. If we connect $p_1, \ldots, p_5$ to form a polygon (black lines), then all points $p_1, \ldots, p_6$ are contained within the polygon, as well as all lines segments connecting each pair (blue lines). To contrast, now consider the polygon formed by sequentially connecting all points $p_1, \ldots, p_6$ in Figure 3.2. The line segments connecting points $p_2 - p_4$, $p_3 - p_4$ and $p_3 - p_5$ (red lines) are not completely contained within the polygon. Therefore, the polygon is not the convex hull for the set $\{p_1, \ldots, p_6\}$.

Further, we define the following notation: for two polygons P_1, P_2, we denote the overlap by $P_1 \cap P_2$. The overlap is the two-dimensional area shared by the two polygons. When all polygons are completely separated, we say that the overlap is $\emptyset$. Figures 3.3 and 3.4 show examples of non-overlapping and overlapping polygons, respectively.

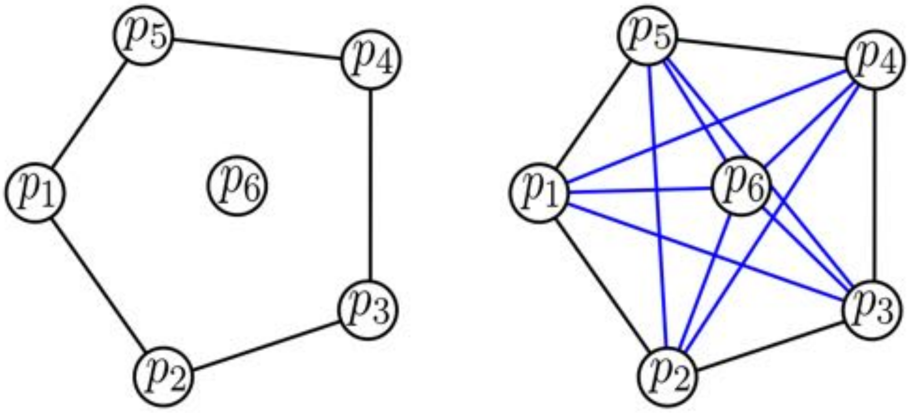


Figure 3.1: An example of the convex hull (black lines) of the set $\{p_1, p_2, \ldots, p_6\}$. Since it is a convex hull, the line segments connecting each pair of points (blue) are contained within the polygon.

To ensure that each Gaussian process model remains spatially distinct, we will constrain cluster assignments using non-overlapping convex hulls. We let $\boldsymbol{s}^{(r)} = \{\boldsymbol{s}_\ell : c_\ell = r\}$ denote the two-dimensional coordinates for observations that are assigned to cluster r. We constrain all cluster assignments such that $\text{Conv}(\boldsymbol{s}^{(r)}) \cap \text{Conv}(\boldsymbol{s}^{(r')}) = \emptyset$ for all $r \neq r'$. That is, from a two-dimensional latitude/longitude prospective, the convex hull for each cluster must be distinct from the others. Figure 3.5 shows examples of both acceptable and unacceptable clustering schemes.

3.1.4 Likelihood function

Due to the conditional independence of each cluster, the likelihood function for the unknown parameters can be factored into R separate pieces. Upon conditioning on the membership for one particular cluster, cluster r, the likelihood function is

$$L(\tau^{(r)}, g^{(r)}, \phi^{(r)}, \mu^{(r)}, \boldsymbol{\beta}^{(r)} | \boldsymbol{y}^{(r)}) =$$
$$(2\pi)^{-n_r/2} |\boldsymbol{\Sigma}^{(r)}|^{-1/2} exp\Big\{ -\frac{1}{2} \big(\boldsymbol{y}^{(r)} - \mu^{(r)}\mathbf{1} - \boldsymbol{X}^{(r)}\boldsymbol{\beta}^{(r)}\big)' (\boldsymbol{\Sigma}^{(r)})^{-1} \big(\boldsymbol{y}^{(r)} - \mu^{(r)}\mathbf{1} - \boldsymbol{X}^{(r)}\boldsymbol{\beta}^{(r)}\big) \Big\}.$$

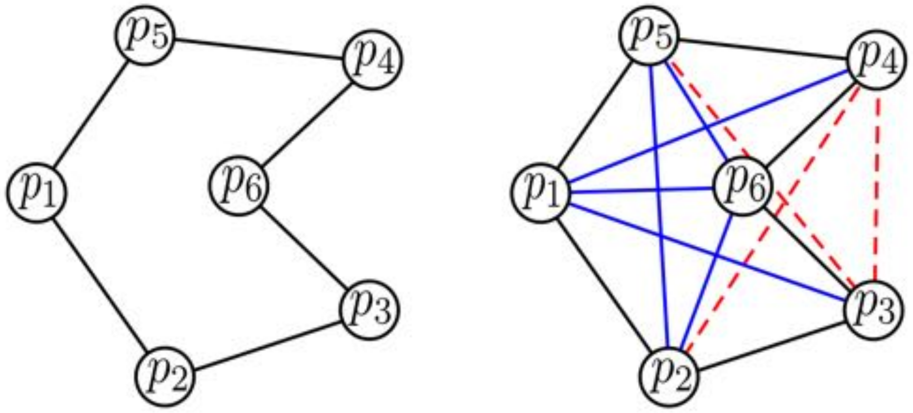


Figure 3.2: An example of a non-convex polygon (black lines) for the set $\{p_1, p_2, \ldots, p_6\}$. The line segments connecting points $p_2 - p_4$, $p_3 - p_4$, and $p_3 - p_5$ are not contained within the polygon.

It should be noted that the likelihood function is conditional on knowing which spatial sites are in cluster r. Combining the $r = 1, \ldots, R$ conditionally independent cluster likelihoods with the convex hull constraint on the spatial coordinates, we can define the joint likelihood for all R clusters,

$$L(\boldsymbol{\tau}, \boldsymbol{g}, \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{c}|\boldsymbol{y}^{(r)}) = \prod_{r=1}^{R} L(\tau^{(r)}, g^{(r)}, \phi^{(r)}, \mu^{(r)}, \boldsymbol{\beta}^{(r)}|\boldsymbol{y}^{(r)}) \Big[\prod_{r \neq r'} \delta\big(\text{Conv}(\boldsymbol{s}^{(r)}) \cap \text{Conv}(\boldsymbol{s}^{(r')}) = \emptyset\big) \Big],$$

with $\boldsymbol{\tau} = (\tau^{(1)}, \ldots, \tau^{(R)})'$, $\boldsymbol{g} = (g^{(1)}, \ldots, g^{(R)})'$, $\boldsymbol{\phi} = (\phi^{(1)}, \ldots, \phi^{(R)})'$, $\boldsymbol{\mu} = (\mu^{(1)}, \ldots, \mu^{(R)})'$, $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(R)})$, and $\boldsymbol{c} = (c_1, \ldots, c_n)'$.

3.2 Intra-cluster prior distributions

Since the major crux of HGP is to determine which covariates are most important in explaining the variability in the response, we specify a prior distribution for $\boldsymbol{\beta}$ that induces a posterior distribution with variable inclusion probabilities. To achieve this, we will use a modified version of Stochastic Search Variable Selection (SSVS) as developed by George &

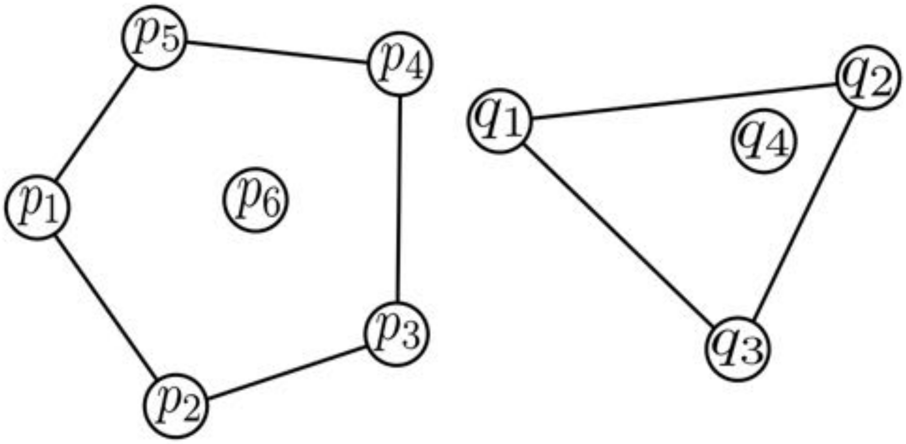


Figure 3.3: The polygons formed by $\text{Conv}(p_1, \ldots, p_6)$ and $\text{Conv}(q_1, \ldots, q_4)$ are distinct and have no overlap.

McCulloch (1993) and Geweke (1996). We let $\boldsymbol{\beta}^{(r)} = (\beta_1^{(r)}, \ldots, \beta_p^{(r)})'$ be the linear effects of covariates $j = 1, \ldots, p$ within cluster r and let $\boldsymbol{\gamma}^{(r)} = (\gamma_1^{(r)}, \ldots, \gamma_p^{(r)})'$ be binary indicators that determine if each of $\beta_j^{(r)}$ are included in the Gaussian process for cluster r. That is,

$$\begin{cases} \beta_j^{(r)} = 0, & \text{if } \gamma_j^{(r)} = 0, \\ \beta_j^{(r)} \neq 0, & \text{if } \gamma_j^{(r)} = 1. \end{cases}$$

Correlation in models between clusters is embedded in two separate but related ways.

1. $\beta_j^{(1)}$,. . . ,$\beta_j^{(R)}$ will be correlated with each other based on the proximity of clusters using a CAR model,

2. $\gamma_j^{(1)}$,. . . ,$\gamma_j^{(R)}$ will be associated with each other based on the proximity of clusters using the Ising distribution.

We thoroughly discuss the prior distributions for $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(R)}$ and $\boldsymbol{\gamma}^{(1)}, \ldots, \boldsymbol{\gamma}^{(R)}$ in Sections 3.2.1 and 3.2.2.

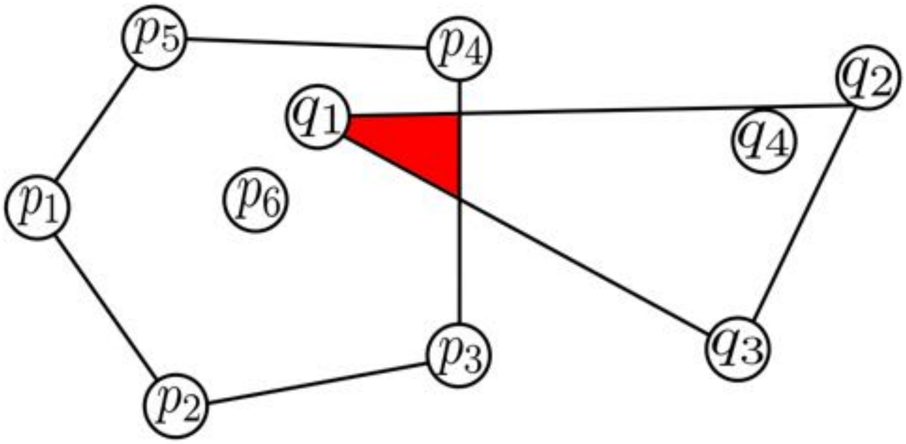


Figure 3.4: The polygons formed by $\mathrm{Conv}(p_1, \ldots, p_6)$ and $\mathrm{Conv}(q_1, \ldots, q_4)$ overlap each other. The region of overlap is colored red.

3.2.1 Conditional prior distribution for linear effects

The HGP methodology leverages the proximity of spatial clusters to smooth estimates for the nonzero $\beta_j^{(r)}$ for across space. One natural choice would be a Gaussian process for the linear effects, but each $\beta_j^{(r)}$ does not represent a point process. Rather, $\beta_j^{(r)}$ is representative of an entire cluster. Therefore, we approximate the Gaussian process by using a CAR model. We specify the joint prior distribution for $\beta_j^{(1)}, \ldots, \beta_j^{(R)}$ conditional upon $\gamma_j^{(1)}, \ldots, \gamma_j^{(R)}$ as

$$P(\beta_j^{(1)}, \ldots, \beta_j^{(R)} | \gamma_j^{(1)}, \ldots, \gamma_j^{(R)}) = \begin{cases} \delta(\beta_j^{(r)} = 0), & \text{for } \{r : \gamma_j^{(r)} = 0\}, \\ Normal(\mathbf{0}, \boldsymbol{K}_j), & \text{for } \{r : \gamma_j^{(r)} = 1\}, \end{cases}$$

where $\delta(\beta_j^{(r)} = 0)$ is the Dirac delta function at 0. Given the variable inclusion indicators $\gamma_j^{(1)}, \ldots, \gamma_j^{(R)}$, it is known for which clusters $\beta_j^{(r)}$ should be nonzero. Conditional upon $\{r : \gamma_j^{(r)} = 1\}$, the corresponding parameters $\beta_j^{(r)}$ follow a joint multivariate Normal distribution with a mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{K}_j$. $\boldsymbol{K}_j$ is a distance-based covariance matrix, much like $\boldsymbol{\Sigma}^{(r)}$ in Section 3.1.1. However the dimensionality of $\boldsymbol{K}_j$ will be determined by $\gamma_j^{(1)}, \ldots, \gamma_j^{(R)}$. That is, $\boldsymbol{K}_j$ is a $(\sum_{r=1}^R \gamma_j^{(r)}) \times (\sum_{r=1}^R \gamma_j^{(r)})$, symmetric matrix. $\boldsymbol{K}_j$ is specified

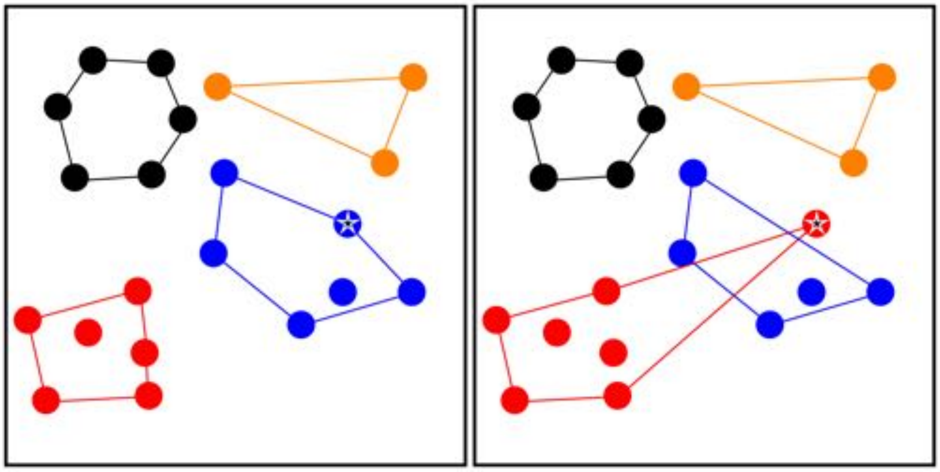


Figure 3.5: Example of acceptable (left) and unacceptable (right) cluster assignments. By reassigning the starred point on the left from the blue cluster to the red cluster, the newly computed convex hulls now overlap.

using the CAR model:

$$\boldsymbol{K}_j = \sigma_\beta^2 \big(\boldsymbol{I} - \phi_\beta \boldsymbol{W}_j\big)^{-1},$$

where $\boldsymbol{W}_j$ a weight matrix based on distances between the clusters for which $\gamma_j^{(r)} = 1$. Many options for distances between clusters exists, including single-linkage, complete-linkage, or centroid-to-centroid distance (Friedman et al. 2001). A discussion about the specification of $\boldsymbol{W}$ is found in Section 3.2.3. The CAR model prior distribution for the nonzero $\beta_j^{(r)}$ enforces a smoothing effect through space. Clusters that are close to each other are expected to have similar coefficients for each covariate. A spatial visualization of the prior distribution can be found in Figure 3.6. The parameters σ_β^2 and ϕ_β will be estimated and are constant for all covariates $j = 1, \ldots, p$. The prior distribution for the intercept of each cluster, $\mu^{(1)}, \ldots, \mu^{(R)}$ is specified in a similar way. Keeping with the notation for the regression coefficients, we use the following joint prior distribution for the intercepts:

$$(\mu^{(1)}, \ldots, \mu^{(R)})' \sim Normal(\boldsymbol{0}, \boldsymbol{K}_0)$$

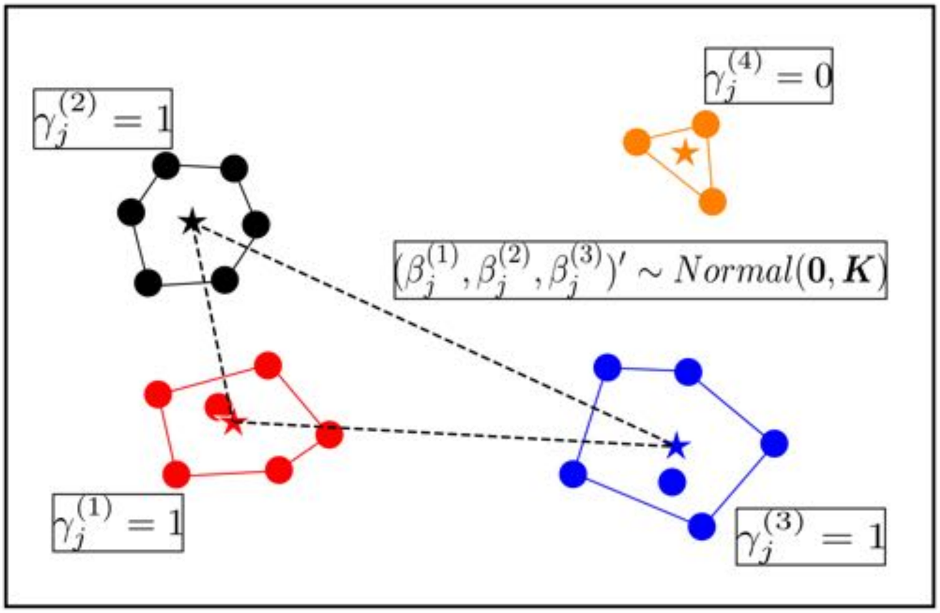

$\gamma_j^{(4)} = 0$
$\gamma_j^{(2)} = 1$
$(\beta_j^{(1)}, \beta_j^{(2)}, \beta_j^{(3)})' \sim Normal(\mathbf{0}, \boldsymbol{K})$
$\gamma_j^{(1)} = 1$
$\gamma_j^{(3)} = 1$

Figure 3.6: Example of joint prior distribution for $(\beta_j^{(1)}, \beta_j^{(2)}, \beta_j^{(3)} | \gamma_j^{(1)} = 1, \gamma_j^{(2)} = 1, \gamma_j^{(3)} = 1)$. Stars represent centroids and dashed-lines are centroid-to-centroid distances. Since $\gamma_j^{(4)} = 0$, $\beta_j^{(4)}$ is not included in the multivariate Normal distribution.

where $\boldsymbol{K}_0 = \sigma_\beta^2 \big(\boldsymbol{I} - \phi_\beta \boldsymbol{W}_0\big)^{-1}$ is a distance matrix with all pairwise distances between clusters. In essence, the prior distribution for the intercepts in each cluster is equivalent to that of the regression coefficients, except that no intercept can leave the model.

3.2.2 Prior distribution for variable inclusion indicators

Section 3.2.1 provided the first layer linking together the Gaussian process models in separate clusters. This implies that the coefficient estimates in nearby clusters should be similar. However, as variable selection is of primary interest, we induce sparsity in the models as well. To achieve spatially correlated sparsity, we implement the Ising distribution as the joint prior distribution for $\gamma_j^{(1)}, \ldots, \gamma_j^{(R)}$,

$$P(\gamma_j^{(1)}, \ldots, \gamma_j^{(R)})' = \frac{1}{\Omega(\theta)} exp\Big\{\theta \sum_{r \sim r'} w_{r,r'} \delta\big(\gamma_j^{(r)} = \gamma_j^{(r')}\big)\Big\} \text{ for } j = 1, \ldots, p.$$

θ is an overall strength parameter that determines the amount of spatial dependency between the variable inclusion in neighboring clusters and $w_{r,r'}$ is the weight between clusters r and r'. As with the cluster-to-cluster distance calculation from Section 3.2.1, there are many options for choosing $w_{r,r'}$. Choosing $w_{r,r'}$ is discussed in Section 3.2.3. Note, in this version of the Ising distribution, we set the external field to 0, implying that if there is no information from neighboring clusters, a priori the probability of variable inclusion is .5. Given the weights and the neighborhood structure, if covariate j is included in cluster r, this prior distribution places a higher probability on covariate j being included in cluster r'. Neighboring clusters with high weights $w_{r,r'}$ will have higher dependency as well. When $\theta = 0$, the model selection within each cluster will be completely independent of the other clusters. Consequently, the variable selection within each model would simplify to the prior distribution $P(\gamma_j^{(r)} = 1) = 0.5$.

3.2.3 Weight specification

The conditional prior distribution for $\beta_j^{(1)}, \ldots, \beta_j^{(R)}$ (Section 3.2.1) and the prior distribution for $\gamma_j^{(1)}, \ldots, \gamma_j^{(R)}$ (Section 3.2.2) require the specification of weights between each cluster. The CAR model on $\beta_j^{(1)}, \ldots, \beta_j^{(R)}$ is dependent on a weight matrix $\boldsymbol{W}$, while the Ising model for $\gamma_j^{(1)}, \ldots, \gamma_j^{(R)}$ needs weights $w_{r,r'}$ for $r \sim r'$. For simplicity and consistency, we will set both sets of weights to be identical, representing general intra-cluster weights.

Choosing weights for both CAR models and Ising models is dependent on the application. When clusters are represented as nodes on a lattice, one can use a neighborhood adjacency matrix for the weights. That is, $w_{r,r'} = 1$ if clusters r and r' are neighbors and $w_{r,r'} = 0$ otherwise. However, since there are no concrete boundaries between the spatial convex hulls of each cluster, a neighborhood structure does not exist. For example, consider the convex hulls in Figure 3.7 where the space is partitioned such that one convex hull is in each section. On the left, the black cluster shares a border with all three other clusters. However, if we use another partition on the right, which still has one cluster per section, the black cluster how only shares a border with the orange cluster. Unlike trees and tessellations, using the convex hull to define cluster assignments leaves no obvious way to draw boundaries. Instead of a neighborhood-based weight matrix, we will choose weights that decay with spatial distance between the clusters. We fix $w_{r,r'} = 1/d_{r,r'}$ for $r \neq r'$ and $w_{r,r} = 0$, where $d_{r,r'}$ is the Euclidean distance between clusters r and r'. As mentioned in Section 3.2.1, there are many options to measure distances between clusters, such as centroid-to-centroid, complete-linkage, and single linkage. Velasco-Cruz (2012) analyzes the effect of weights in the context regions defined by tessellations, where the choice can impact the analysis. Since the centroids change throughout the posterior sampling, we utilize the centroid-to-centroid distance for computational simplicity. For both the CAR model and the Ising model,

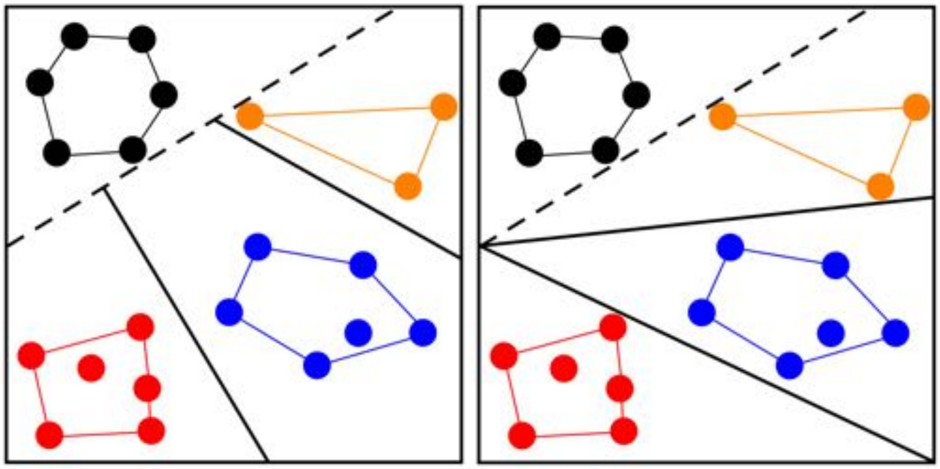


Figure 3.7: Two different partitioning schemes on the convex hull clusters. On the left, the black cluster has three shared borders. On the right, the black cluster has only one shared border.

$$\boldsymbol{W} = \begin{bmatrix} 0 & 1/d_{1,2} & \dots & 1/d_{1,R} \\ 1/d_{2,1} & 0 & \dots & 1/d_{2,R} \\ \vdots & \vdots & \ddots & \vdots \\ 1/d_{R,1} & 1/d_{R,2} & \dots & 0 \end{bmatrix}$$

is fixed and known, conditional upon the cluster assignments for all locations.

3.2.4 Prior distribution for CAR parameters

In Section 3.2.1, we introduced the CAR model that is used as the prior distribution for $\mu^{(1)}, \dots, \mu^{(R)}$ and $\beta_j^{(1)}, \dots, \beta_j^{(R)}$ for $j = 1, \dots, p$. The prior distribution on these intercepts and linear effects introduced two new parameters that need to be inferred from the data, σ_β^2 and ϕ_β. The prior distribution these parameters needs to be carefully chosen. For instance, given an $n \times n$ CAR model weight matrix $\boldsymbol{W}$ and its sorted eigenvalues $\lambda_1, \dots, \lambda_n$, we know ϕ_β must exist in the interval $(\lambda_1^{-1}, \lambda_n^{-1})$. Ren & Sun (2013) and De Oliveira (2012) derived and compared reference and Jeffreys prior distributions for the CAR model. We use the independent Jeffreys σ_β^2 and ϕ_β:

$$P(\phi_\beta, \sigma_\beta^2) \propto \frac{1}{\sigma_\beta^2}\left(\sum_{r=1}^{R}\left(\frac{\lambda_r}{1-\phi_\beta\lambda_r}\right)^2 - \frac{1}{R}\left(\sum_{r=1}^{R}\left(\frac{\lambda_r}{1-\phi_\beta\lambda_r}\right)\right)^2\right),$$

where $\lambda_1, \dots, \lambda_R$ are the sorted eigenvalues of $\boldsymbol{W}_0$, the weight matrix across all clusters. The independent Jeffreys prior distribution is improper, but it places a large amount of mass at the boundaries λ_1^{-1} and λ_R^{-1}, which represents large amounts of spatial dependence. An example of the this log-prior distribution when $\sigma_\beta^2 = 1$ can be seen in Figure 3.8. The density at the endpoints is infinite.

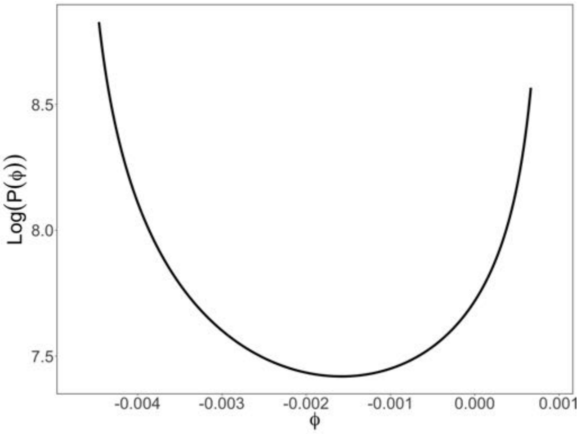


Figure 3.8: Example of log-prior distribution for ϕ for $n = 100$ evenly spaced points on the unit-square grid with $\boldsymbol{W}_{ij} = 1/d_{ij}$, where d_{ij} represents the Euclidean distance. Much like the $Beta(1/2, 1/2)$, there is infinite density at the endpoints.

3.2.5 Similarity to Stochastic Search Variable Selection

CAR model for linear effects

The prior distribution for $(\beta_j^{(1)}, \ldots, \beta_j^{(R)})$ of Section 3.2.1 is actually a generalization of the prior distribution specified by Geweke (1996) that we call Conditional Autoregressive Stochastic Search Variable Selection (CAR-SSVS). Recall the prior distribution for Stochastic Search Variable Selection (SSVS) (Section 2.1), which we refer to as Independent-SSVS:

$$P(\beta_j) = (1 - \gamma_j)\delta(\beta_j = 0) + \gamma_j Normal(0, \sigma_\beta^2).$$

Under our specified model with $r = 1, \ldots, R$ clusters, each with a separate model, the SSVS prior distribution becomes

$$P(\beta_j^{(r)}) = (1 - \gamma_j^{(r)})\delta(\beta_j^{(r)} = 0) + \gamma_j^{(r)} Normal(0, \sigma_\beta^2),$$

or, equivalently

$$P(\beta_j^{(1)}, \ldots, \beta_j^{(R)} | \gamma_j^{(1)}, \ldots, \gamma_j^{(R)}) = \begin{cases} \delta(\beta_j^{(r)} = 0), & \text{for } \{r : \gamma_j^{(r)} = 0\}, \\ Normal(\mathbf{0}, \sigma_\beta^2 \boldsymbol{I}), & \text{for } \{r : \gamma_j^{(r)} = 1\}. \end{cases}$$

The original SSVS prior distribution exactly matches our specified prior distribution if we had selected the covariance matrix $\boldsymbol{K}_j$ to be a diagonal matrix with constant variance σ_β^2. This implies that, if $\gamma_j^{(r)} = 1$, the conditional distribution of $\beta_j^{(r)}$ does not depend on $\beta_j^{(r')}$ for $r \neq r'$. The comparison becomes clearer when we analyze the conditional prior distribution for $\beta_j^{(r)}$ under our CAR-SSVS.

First, we consider the conditional prior distribution

$$P(\beta_j^{(r)} | \beta_j^{(1)}, \ldots \beta_j^{(r-1)}, \beta_j^{(r+1)}, \ldots, \beta_j^{(R)}, \gamma_j^{(1)}, \ldots, \gamma_j^{(R)})$$

under CAR-SSVS specification. Assume, without loss of generality, that $\gamma_j^{(r')} = 1$ for $r' = 1, \dots, r-1$ and $\gamma_j^{(r')} = 0$ for $r' = r+1, \dots, R$. Then $\beta_j^{(r)} = 0$ if $\gamma_j^{(r)} = 0$ or $(\beta_j^{(1)}, \dots, \beta_j^{(r)})' \sim Normal(\mathbf{0}, \boldsymbol{K}_j)$ if $\gamma_j^{(r)} = 1$. Using the properties of jointly distributed Normal random variables, we write the conditional prior distribution for $\beta_j^{(r)}$ as

$$P(\beta_j^{(r)}|\beta_j^{(1)}, \dots \beta_j^{(r-1)}, \gamma_j^{(1)}, \dots, \gamma_j^{(r)}) = (1-\gamma_j^{(r)})\delta(\beta_j^{(r)} = 0) + \gamma_j^{(r)} Normal(m, V),$$

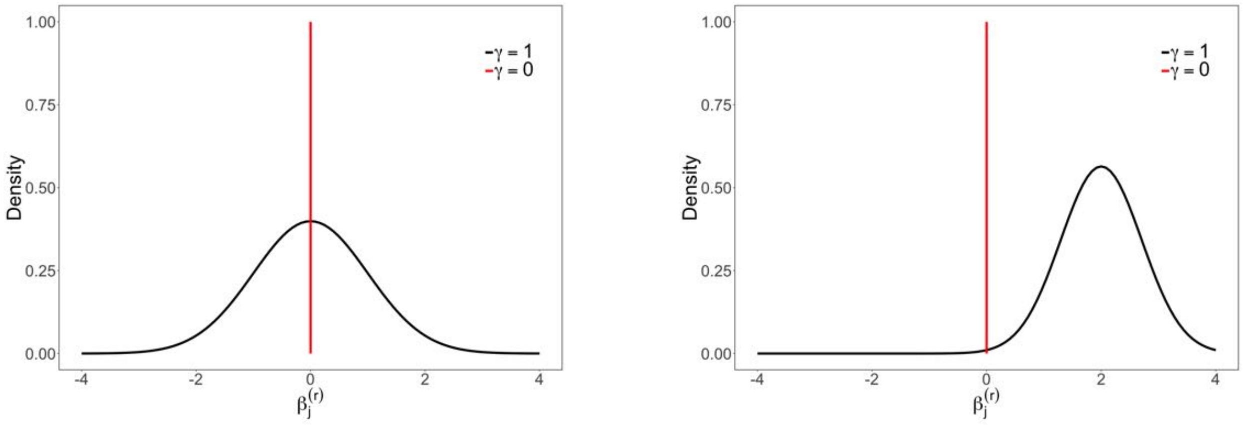
where

$$\begin{aligned} V &= \boldsymbol{K}_{r,r} - \boldsymbol{K}_{r,1:(r-1)}(\boldsymbol{K}_{1:(r-1),1:(r-1)})^{-1}\boldsymbol{K}_{1:(r-1),1}, \\ m &= \boldsymbol{K}_{r,1:(r-1)}(\boldsymbol{K}_{1:(r-1),1:(r-1)})^{-1}(\beta_j^{(1)}, \dots, \beta_j^{(r-1)})'. \end{aligned}$$

In contrast to Independent-SSVS, the conditional prior distribution of $\beta_j^{(r)}$ under CAR-SSVS combines a point mass at 0 with a nonzero-centered Normal distribution. By sharing information across clusters, we remove a significant portion of prior mass from 0 and shift it towards the values $\beta_j^{(1)}, \dots, \beta_j^{(r-1)}$. This restructuring implies that a priori, the linear effect of cluster r should be near that of its neighbors, rather than centered at 0. The difference in prior distributions between Independent-SSVS and CAR-SSVS can be seen visually in Figure 3.9.

Ising distribution for variable inclusion indicators

While Section 3.2.5 discusses the similarity between CAR-SSVS and SSVS in terms of the joint distribution for $\beta_j^{(1)}, \dots, \beta_j^{(R)}$, there is another connection with Independent-SSVS in terms of the Ising distribution. Recall, the Ising model is used as a joint prior distribution over the variable inclusion indicators $\gamma_j^{(1)}, \dots, \gamma_j^{(R)}$. First, consider again the prior specifica-



(a) Independent-SSVS

(b) CAR-SSVS

Figure 3.9: The conditional prior distribution for $\beta_j^{(r)}$ under Independent-SSVS and CAR-SSVS specifications. The spatial structure of CAR-SSVS shifts the prior mean and shrinks the prior variance.

tion for $\beta_j^{(r)}$ and $\gamma_j^{(r)}$ under Independent-SSVS:

$$\begin{aligned} P(\beta_j^{(r)}|\gamma_j^{(r)}) &= (1-\gamma_j^{(r)})\delta(\beta_j^{(r)}=0)+\gamma_j^{(r)}Normal(0,\sigma_\beta^2), \\ P(\gamma_j^{(r)}) &= p_j^{(r)}. \end{aligned}$$

Often $p_j^{(r)}$ is selected to be 0.5, indicating that, a priori, the probability that $\beta_j^{(r)} \neq 0$ with probability .5. While we looked carefully at the $P(\beta_j^{(r)}|\gamma_j^{(r)}=1)$ in Section 3.2.5, we now examine the prior distribution on $\gamma_j^{(r)}$. Rather than $p_j^{(r)}=0.5$, we use the Ising model for $\gamma_j^{(1)},\ldots,\gamma_j^{(R)}$:

$$P(\gamma_j^{(1)},\ldots,\gamma_j^{(R)})' = \frac{1}{\Omega(\theta)}exp\Big\{\theta\sum_{r\sim r'}w_{r,r'}\delta\big(\gamma_j^{(r)}=\gamma_j^{(r')}\big)\Big\}.$$

While this joint distribution requires the computation of the intractable normalizing constant, the conditional distribution $P(\gamma_j^{(r)}|\boldsymbol{\gamma}_j^{-(r)}) = P(\gamma_j^{(r)}|\gamma_j^{(1)},\ldots,\gamma_j^{(r-1)},\gamma_j^{(r+1)},\ldots,\gamma_j^{(R)})$

does not. Since $\gamma_j^{(r)} \in \{0,1\}$, $P(\gamma_j^{(r)}|\boldsymbol{\gamma}_j^{-(r)})$ must be a Bernoulli distribution. That is,

$$P(\gamma_j^{(r)}|\boldsymbol{\gamma}_j^{-(r)}) \propto exp\Big\{\theta \sum_{r\sim r'} w_{r,r'}\delta\big(\gamma_j^{(r)} = \gamma_j^{(r')}\big)\Big\}.$$

By appropriately normalizing $P(\gamma_j^{(r)}|\boldsymbol{\gamma}_j^{-(r)})$, we can find $p_j^{(r)} = P(\gamma_j^{(r)} = 1|\boldsymbol{\gamma}_j^{-(r)})$:

$$\begin{aligned} P(\gamma_j^{(r)} = 1|\boldsymbol{\gamma}_j^{-(r)}) &= \frac{P(\gamma_j^{(r)} = 1|\boldsymbol{\gamma}_j^{-(r)})}{P(\gamma_j^{(r)} = 1|\boldsymbol{\gamma}_j^{-(r)}) + P(\gamma_j^{(r)} = 0|\boldsymbol{\gamma}_j^{-(r)})} \\ &= \left(1 + exp\Big\{\theta \sum_{r\sim r'} w_{r,r'}\Big[\delta\big(\gamma_j^{(r')} = 0\big) - \delta\big(\gamma_j^{(r')} = 1\big)\Big]\Big\}\right)^{-1}. \end{aligned}$$

Rather than fixing $p_j^{(r)}$ a priori, we use information about $\gamma_j^{-(r)}$ to influence the prior probability for $\gamma_j^{(r)}$. Note, if $\theta = 0$, $p_j^{(r)}$ reduces to 0.5, and $\gamma_j^{(r)}$ is independent of $\boldsymbol{\gamma}_j^{-(r)}$. However, if $\theta > 0$, the indicators for neighboring clusters are related to each other.

Let us consider an example with $R = 4$. We analyze the variable indicator for $\gamma_j^{(1)}$ conditional on $\gamma_j^{(2)}$, $\gamma_j^{(3)}$, $\gamma_j^{(4)}$. We further assume that $w_{1,2} = w_{1,3} = w_{1,4} = \frac{1}{3}$. We show the effect of θ on $P(\gamma_j^{(r)} = 1|\gamma_j^{-(r)})$ under the following scenarios:

1. $\gamma_j^{(2)} = 0$,$\gamma_j^{(3)} = 0$,$\gamma_j^{(4)} = 0$,
2. $\gamma_j^{(2)} = 1$,$\gamma_j^{(3)} = 0$,$\gamma_j^{(4)} = 0$,
3. $\gamma_j^{(2)} = 1$,$\gamma_j^{(3)} = 1$,$\gamma_j^{(4)} = 0$,
4. $\gamma_j^{(2)} = 1$,$\gamma_j^{(3)} = 1$,$\gamma_j^{(4)} = 1$.

First, consider Scenarios 3 and 4. In Scenario 4, all neighboring clusters have variable inclusion indicators of 1. Consequently, for $\theta > 0$, the probability of variable inclusion in cluster 1 increases sharply as a function of θ. However, for Scenario 3, only 2 out of the 3 other clusters have variable inclusion indicators of 1. Overall, this still increases the probability of variable inclusion, but not as drastically as Scenario 4. Very similar results hold for Scenarios

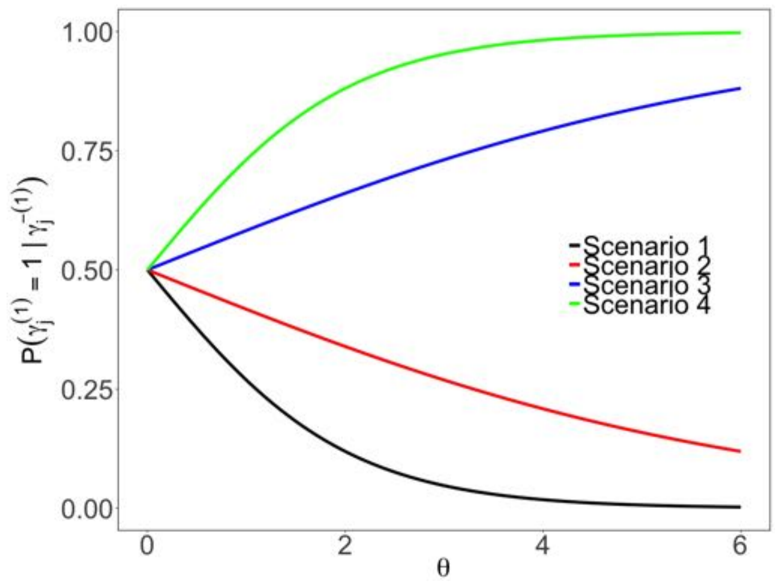


Figure 3.10: Paths denoting the change in $P(\gamma_j^{(1)} = 1|\boldsymbol{\gamma}_j^{-(1)})$ with respect to θ under the four scenarios. When θ is larger, the effect on the probability is more dramatic.

1 and 2, except the probabilities decreases from .5 rather than increase, since the majority of the other clusters have variable inclusion indicators of 0. Notice, when $\theta = 0$, all scenarios have a inclusion probability of 0.5 since there is no neighborhood structure and the external field is 0.

3.2.6 Summary of HGP methodology

We now follow the detailed descriptions of Sections 3.1.2, 3.1.3, 3.2.1, and 3.2.2 with a concise summary of the methodology. Our model statement and prior distributions are as follows:

$$\begin{aligned}
(\boldsymbol{y}^{(r)}|\mu^{(r)}, \boldsymbol{\beta}^{(r)}, \boldsymbol{\Sigma}^{(r)}, \boldsymbol{c}) &\sim Normal(\mu^{(r)}\mathbf{1} + \boldsymbol{X}^{(r)}, \boldsymbol{\Sigma}^{(r)}) \\
&\qquad \text{subject to } \mathrm{Conv}(\boldsymbol{s}^{(r)}) \cap \mathrm{Conv}(\boldsymbol{s}^{(r')}) = \emptyset, \\
c_i &\sim Multinomial(1/R, \ldots, 1/R), \\
(\mu^{(1)}, \ldots, \mu^{(R)}|\sigma^2_\beta, \phi_\beta)' &\sim Normal(\mathbf{0}, \boldsymbol{K}_0), \\
P(\beta_j^{(1)}, \ldots, \beta_j^{(R)}|\gamma_j^{(1)}, \ldots, \gamma_j^{(R)}, \sigma^2_\beta, \phi_\beta) &= \begin{cases} \delta(\beta_j^{(r)} = 0), & \text{for } \{r : \gamma_j^{(r)} = 0\}, \\ Normal(\mathbf{0}, \boldsymbol{K}_j), & \text{for } \{r : \gamma_j^{(r)} = 1\}, \end{cases} \\
(\gamma_j^{(1)}, \ldots, \gamma_j^{(R)}|\theta)' &\sim Ising(\theta), \\
P(\phi_\beta, \sigma^2_\beta) &\propto \frac{1}{\sigma^2_\beta}\left(\sum_{r=1}^{R} \Big(\frac{\lambda_r}{1-\phi_\beta\lambda_r}\Big)^2 - \frac{1}{R}\left(\sum_{r=1}^{R} \Big(\frac{\lambda_r}{1-\phi_\beta\lambda_r}\Big)\right)^2 \right),
\end{aligned}$$

where $\boldsymbol{K}_0 = \sigma^2_\beta(\boldsymbol{I} - \phi_\beta \boldsymbol{W})^{-1}$, λ_r is the r^{th} largest eigenvalue of $\boldsymbol{W}$, and $\boldsymbol{W}$ the weight matrix with elements $w_{r,r'} = 1/d_{r,r'}$ for $r \neq r'$ and $w_{r,r} = 0$. $d_{r,r'}$ is the Euclidean centroid-to-centroid distance between clusters r and r'. The prior distribution for the Gaussian process parameters $\tau^{(r)}$, $\phi^{(r)}$, and $g^{(r)}$ is application specific. We discuss the prior distribution for θ in Section 5.1.

Chapter 4

Considerations for HGP implementation

In Chapter 3, we present the likelihood function and prior distributions for HGP. Now, we address the implementation and posterior sampling of our hierarchical model using MCMC. The posterior distribution of the cluster assignment labels and Ising parameter require special care. We describe the computational considerations for the cluster assignments and the approximation that we utilize to easily sample the Ising parameter.

4.1 Full conditional distributions

Recall the hierarchical model listed in Section 3.2.6. To perform inference on the unknown parameters, we use MCMC to sample from the joint posterior distribution. For purposes of discussion, we group the unknown parameters as follows: 1) mean structure ($\boldsymbol{\mu}$, $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$), 2) CAR parameters ($\phi_\beta, \sigma^2_\beta$), 3) covariance structure ($\boldsymbol{\phi}$,$\boldsymbol{\tau}$,$\boldsymbol{g}$) 4) cluster labels ($\boldsymbol{c}$), and 5) Ising parameter (θ). The mean structure and CAR parameters can be easily sampled from their full conditional distributions (Appendix A) using either a Gibbs step or a Metropolis-within-

Gibbs step. We also include the generic full conditional distributions for the covariance structure parameters (Appendix A), however the choice of prior distribution for ($\boldsymbol{\phi}$,$\boldsymbol{\tau}$,$\boldsymbol{g}$) is specific to the application. The cluster labels and Ising parameter require more care during implementation. Therefore, we discuss the posterior sampling of $\boldsymbol{c}$ and θ in Sections 4.2 and 4.4.

4.2 Clusters assignments with convex hull constraint

The model selection and estimation methodology of Sections 3.2.1 and 3.2.2 is predicated on knowing cluster assignment labels for observations $1, \ldots, n$. These labels are unknown model parameters and need to be inferred from the data. The assignment of each observation to one of the R available clusters is computationally demanding. Therefore, we discuss the procedure and considerations for the implementation of this portion of the methodology.

4.2.1 Mixture of Gaussian process models

We begin by discussing the process by which cluster assignments are made when no constraints exist. The likelihood function for the parameters, conditional on cluster assignment labels, is written as

$$L(\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\phi}, \boldsymbol{g} | \boldsymbol{y}, \boldsymbol{c}) = \prod_{r=1}^{R} Normal(\boldsymbol{y}^{(r)} | \mu^{(r)} + \boldsymbol{X}^{(r)}, \Sigma^{(r)}),$$

where $\Sigma^{(r)}$ is based on the Gaussian correlation function with parameters $\tau^{(r)}$, $\phi^{(r)}$, and $g^{(r)}$ (further details in Section 3.1.2). Since $\boldsymbol{c}$ is unknown, we sample from the full conditional distribution using MCMC using the prior distribution $P(c_i = r) = 1/R$.

Consider the proposal of observation i to cluster r. Then the coordinates $\boldsymbol{s}_i$ would be added to the Gaussian correlation function within the covariance $\boldsymbol{\Sigma}^{(r)}$. Treating observation

i as the first element in covariance matrix, we can then write $\mathbf{\Sigma}^{(r)}$ as

$$\mathbf{\Sigma}^{(r)} = \begin{bmatrix} \Sigma_{1,1}^{(r)} & \Sigma_{1,2}^{(r)} \\ \Sigma_{2,1}^{(r)} & \Sigma_{2,2}^{(r)} \end{bmatrix}$$

where $\Sigma_{1,1}^{(r)} = (\tau^{(r)})^2(1+g^{(r)})$, $\Sigma_{1,2}^{(r)} = \boldsymbol{\rho}(\boldsymbol{s}_i, \{\boldsymbol{s}_j : c_j = r\})$, and
$\Sigma_{2,2}^{(r)} = (\tau^{(r)})^2(\boldsymbol{\rho}(\{\boldsymbol{s}_j : c_j = r\}, \{\boldsymbol{s}_j : c_j = r\}) + g^{(r)}\boldsymbol{I})$. Similarly, we can augment the covariate matrix and response vector for cluster r as

$$\boldsymbol{X}^{(r)} = \begin{bmatrix} \boldsymbol{x}_1' \\ \boldsymbol{X}_2 \end{bmatrix}$$

and

$$\boldsymbol{y}^{(r)} = \begin{bmatrix} y_1 \\ \boldsymbol{y}_2 \end{bmatrix}$$

where $\boldsymbol{x}_1$ and y_1 are the vector of covariates and scalar response for observation i, respectively, and $\boldsymbol{X}_2$ and $\boldsymbol{y}_2$ are the matrix of covariates and response vector for the other observations in cluster r. Then, the full conditional distribution for the cluster assignment of observation i is given by

$$(c_i|\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{g}, \boldsymbol{\tau}, \boldsymbol{y}) \sim Multinomial(\pi_1, \ldots, \pi_R) \tag{4.1}$$

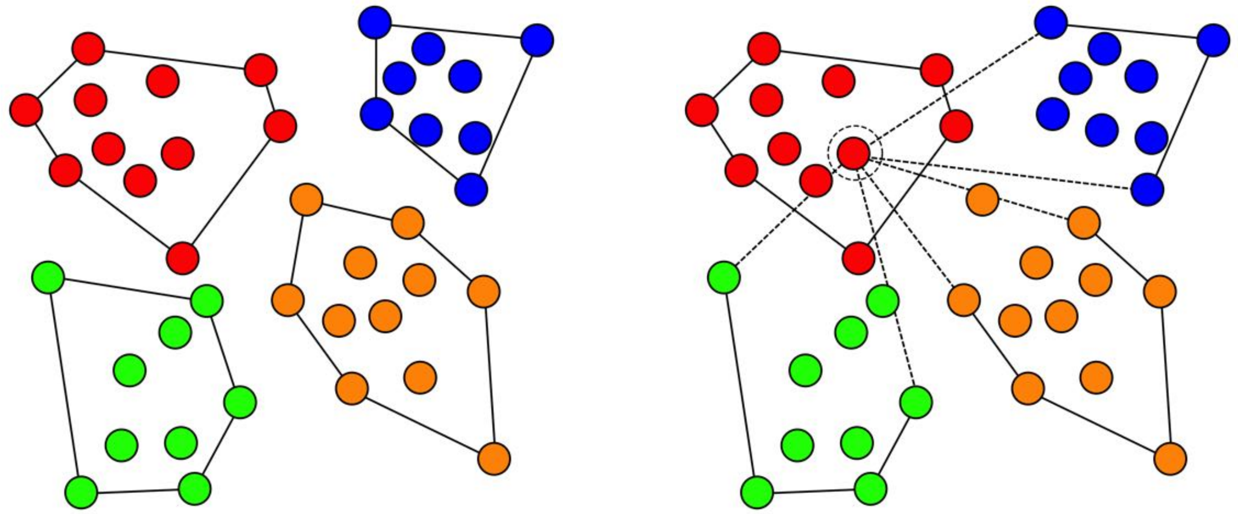
where π_r is proportional to the predictive density of y_i given $\{y_j : c_j = r\}$. That is,

$$\pi_r \propto v^{-1/2} exp\Big\{ -\frac{1}{2v^{1/2}}\big(y_i - e\big)^2 \Big\} \tag{4.2}$$

where $v = \Sigma_{1,1}^{(r)} - \Sigma_{1,2}^{(r)}\big(\Sigma_{2,2}^{(r)}\big)^{-1}\Sigma_{2,1}^{(r)}$ and $e = \mu^{(r)} + \boldsymbol{x}_1'\boldsymbol{\beta}^{(r)} + \Sigma_{1,2}^{(r)}\big(\Sigma_{2,2}^{(r)}\big)^{-1}(\boldsymbol{y}_2 - \mu^{(r)} - \boldsymbol{X}_2\boldsymbol{\beta}^{(r)})$. Therefore, for every point $i = 1, \ldots, n$, we must calculate π_r, resulting in nR calculations every MCMC iteration.

4.2.2 Constrained cluster assignments

We now add the non-overlapping convex hull constraint (Section 3.1.3) to the full conditional distribution for c_i (Equation 4.1). Recall, $\boldsymbol{s}^{(r)}$ is the collection of spatial coordinates for the observations that are assigned to cluster r; that is, $\boldsymbol{s}^{(r)} = \{\boldsymbol{s}_\ell : c_\ell = r\}$. A set of cluster assignments $\boldsymbol{c}$ is valid if and only if $\text{Conv}(\boldsymbol{s}^{(r)}) \cap \text{Conv}(\boldsymbol{s}^{(r')}) = \emptyset$ for all $r \neq r'$. Suppose that at a given MCMC iteration, the clusters are assigned in a valid manner (Figure 4.1a). Then if a point does not currently exist on the convex hull, assigning it to a new cluster will lead to overlapping convex hulls (Figure 4.1b).

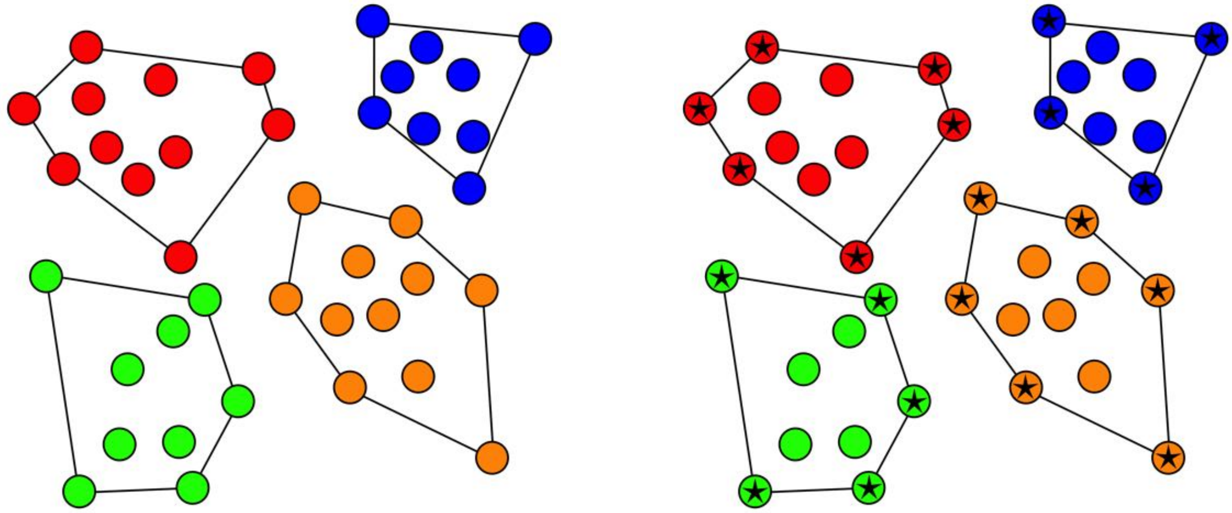


(a) Valid clusters

(b) Overlapping convex hulls

Figure 4.1: If a point does not exist on the convex hull, then assigning it to any other cluster will lead to overlapping convex hulls. The dashed lines represent the new convex hulls if the circled point were reassigned.

Since we only consider cluster assignments such that $\text{Conv}(\boldsymbol{s}^{(r)}) \cap \text{Conv}(\boldsymbol{s}^{(r')}) = \emptyset$, on any given MCMC iteration we only propose that a point moves clusters if it exists on the convex hull of its current cluster. In Figure 4.2a, there are 40 observations in four existing clusters. However, we only need to compute $P(c_i|\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{g}, \boldsymbol{\tau}, \boldsymbol{y})$ for the points with stars (Figure 4.2b), greatly reducing the number of computations per iteration.

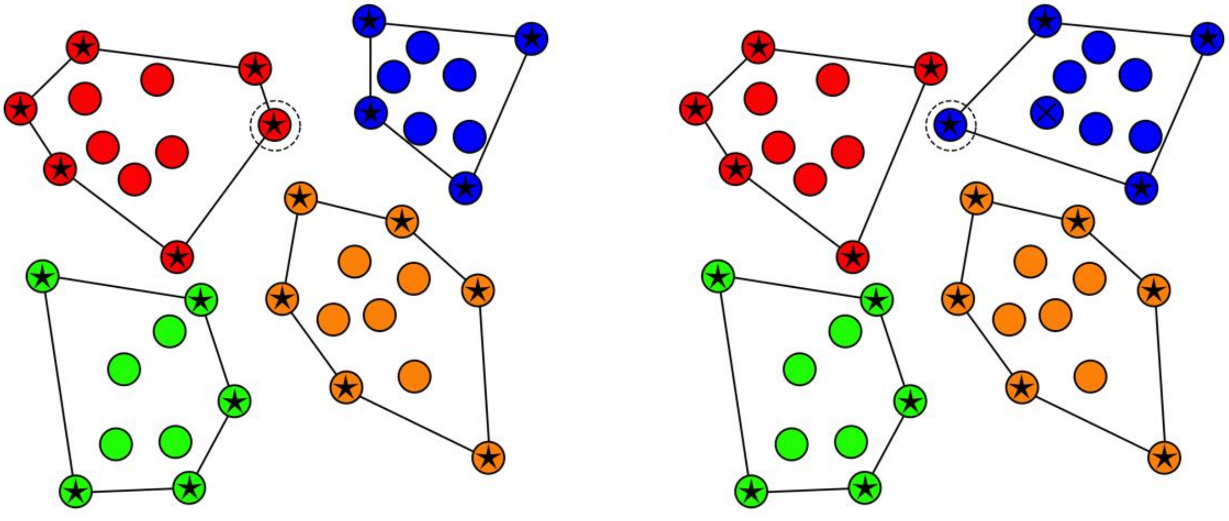


(a) Valid clusters

(b) Points that are eligible for reassignment

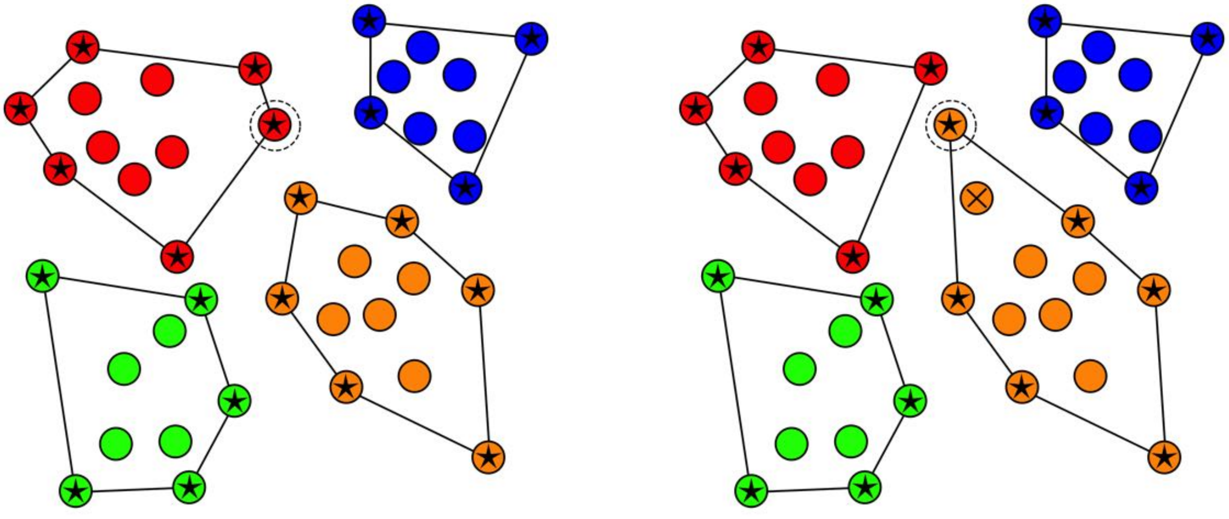
Figure 4.2: Cluster assignments with stars denoting the observations that exist on the convex hull.

Let i be a point on the convex hull of its cluster, r. Then, observation i is eligible for movement from cluster r to another existing cluster. The draw from $P(c_i|\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{g}, \boldsymbol{\tau}, \boldsymbol{y})$ will involve computing $\pi_1, \ldots, \pi_R$. Although it is possible for observation i to move, it may not be able to move all R clusters. Consider the initial cluster configuration in Figure 4.3a where observation i is circled. Suppose we propose to move the point from the red cluster to the blue cluster (Figure 4.3b). The resulting convex hulls would not overlap, therefore the move would be valid. We obtain the same result when we propose to move the point into the orange cluster (Figures 4.4a and 4.4b). However, if we propose a move into the green cluster (Figures 4.5a and 4.5b), the proposed red and green convex hulls will overlap. For a point that is on its convex hull, we first check to see which other clusters is it able to join (without introducing overlapping convex hulls). Therefore, we only calculate π_k for the proposed cluster assignments that are valid.



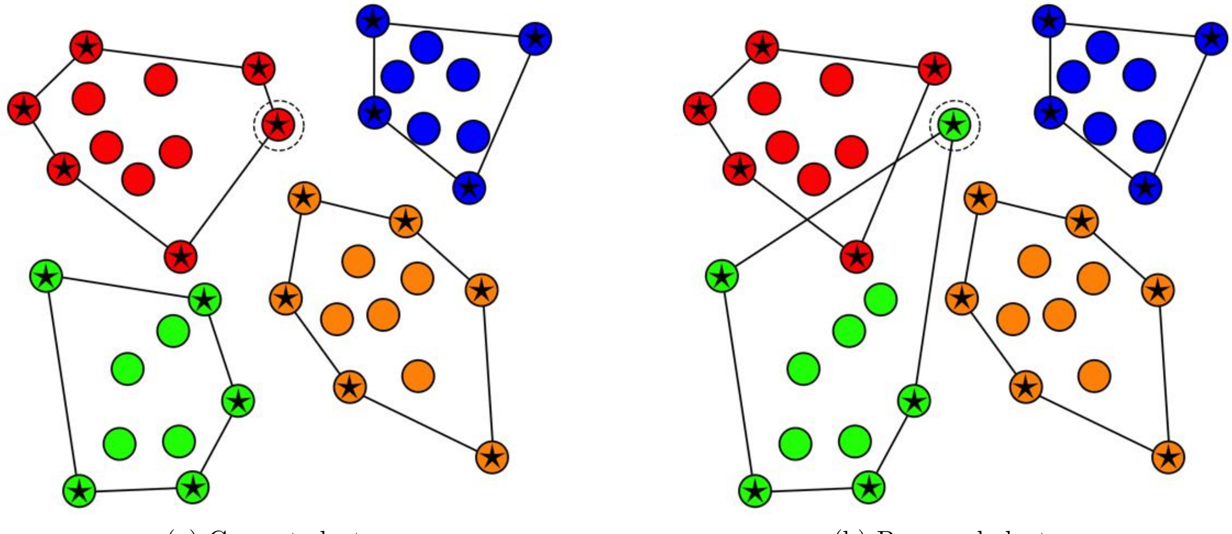
(a) Initial configuration (b) Proposed clusters

Figure 4.3: Proposing to move point (circle) from red cluster to blue cluster. Since the resulting convex hulls do not overlap, the move is legal.



(a) Current clusters (b) Proposed clusters

Figure 4.4: Proposing to move point (circle) from red cluster to orange cluster. Since the resulting convex hulls do not overlap, the move is legal.



(a) Current clusters

(b) Proposed clusters

Figure 4.5: Proposing to move point (circle) from red cluster to green cluster. Since the resulting convex hulls overlap, the move is unacceptable.

4.2.3 Summary of posterior sampling of cluster labels

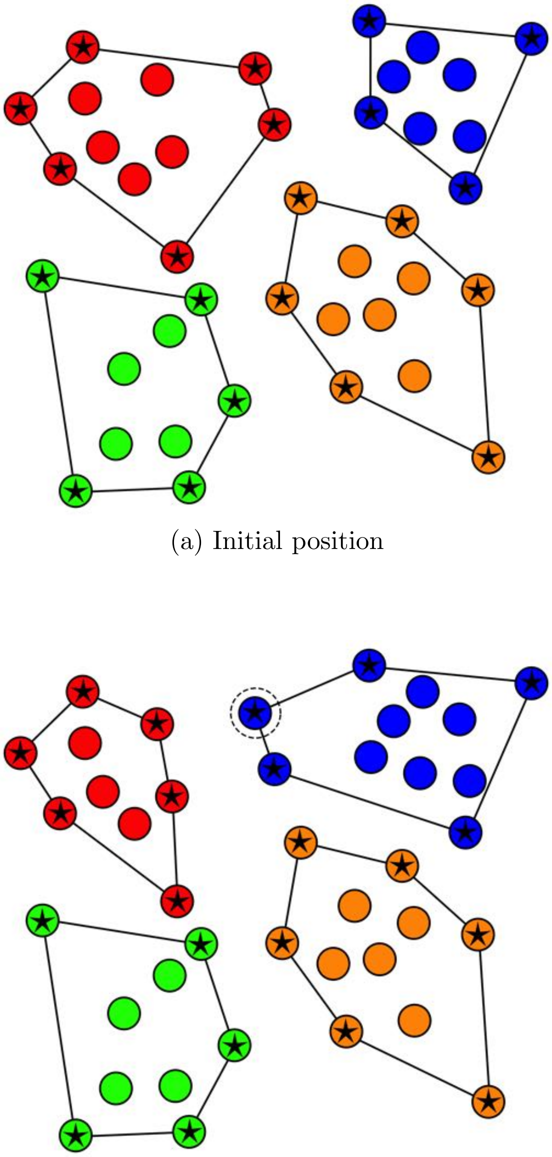
Combining Sections 4.2.1 and 4.2.2, we can concisely describe the posterior sampling of the cluster labels $\boldsymbol{c}$ in a simple algorithm. For observation i, we perform the following:

Algorithm 1: Steps for drawing c_i from its full conditional distribution

```
if $\boldsymbol{s}_i$ is on convex hull of cluster $c_i$ then
    for $r \in \{1, \ldots, R\}$ do
        if $c_i \to r$ is valid then
            Compute $\pi_r$ (Equation 4.2)
        else
            $\pi_r = 0$
        end
    end
    Normalize $\pi_1, \ldots, \pi_R$
    Draw $c_i \sim Multinomial(\pi_1, \ldots, \pi_R)$
else
    Do not update $c_i$
end
```

4.3 MCMC mixing of cluster assignments

The posterior sampling of the cluster labels $\boldsymbol{c}$ is straightforward and only requires proposals for observations that exist on the convex hull. The fallback, however, is the rate at which we are able to explore the posterior distribution. While it is easy for a single observation to move into a new cluster, it can be difficult for a cluster to grow or shrink significantly. Consider the example starting position in Figure 4.6a. We first move a red point into the blue cluster (circled) (Figure 4.6b). Next, we move a second red point into the blue cluster (Figure 4.6c). Continuing the process (Figures 4.6d-4.6f), the blue cluster grows, spanning the entire top part of the space. But this type of cluster growth requires the consecutive acceptances for proposals that move the red points into the blue cluster. Therefore, the movement of the convex hulls and cluster assignments throughout the posterior sampling resembles “crawling”. To improve the exploration of the posterior distribution, we use a special cluster proposal, which we detail in Section 4.3.1.



(a) Initial position

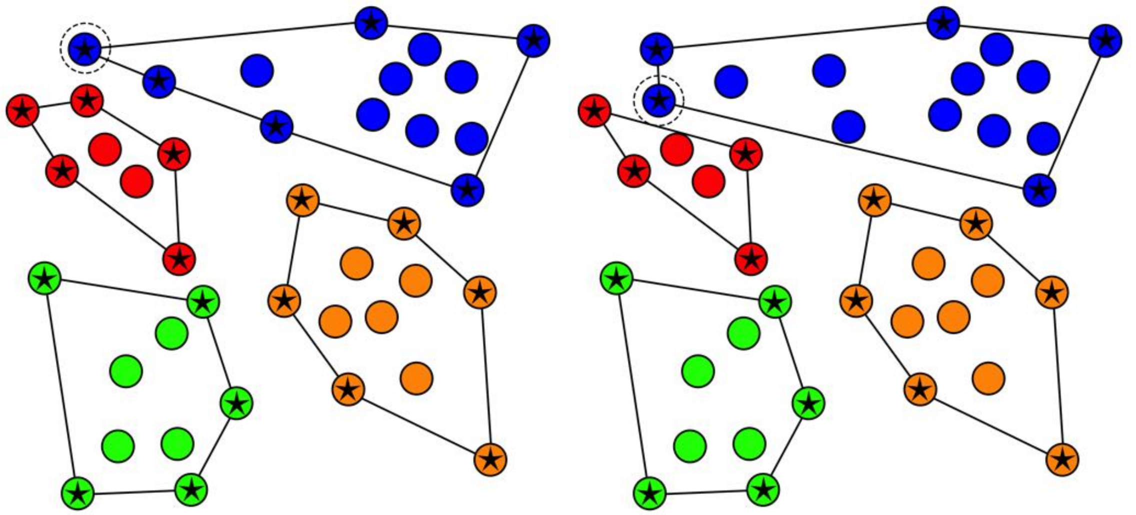
(b) Move 1

(c) Move 2

(d) Move 3

4.3.1 MCMC proposal for improved mixing

As can be seen visually in Figures 4.6a-4.6f, although it is simple to move one point to a new cluster, it can be demanding to drastically shift the overall shape of the clusters.



(e) Move 4

(f) Move 5

To remedy this issue, we utilize a cluster assignment proposal distribution that reassigns many points simultaneously. We propose R new cluster centroids $\mathcal{C}_1^*, \ldots, \mathcal{C}_R^*$ and assign each observation to the cluster with the closest (Euclidean distance) centroid, obtaining proposed cluster labels $c_1^*, \ldots, c_n^*$. We then propose new parameters $\mu^{(r)*}$,$\boldsymbol{\beta}^{(r)*}$, $g^{(r)*}$, $\phi^{(r)*}$, and $\tau^{(r)*}$ based on the maximum likelihood estimate of each new cluster (conditional upon $\boldsymbol{\gamma}^{(r)}$).

First, we propose new cluster centroids $(\mathcal{C}_1^*, \ldots, \mathcal{C}_R^*) \sim Uniform(a_1, b_1) \times Uniform(a_2, b_2)$, where a_1, a_2, b_1, b_2 are the minimum longitude, minimum latitude, maximum longitude, and maximum latitude, respectively. To assign observation i to its new cluster, we solve $\underset{j}{\text{ArgMin}}||s_i - \mathcal{C}_j||$, which places the observation in the cluster with proposed centroid closest to s_i. Given $c_1^*, \ldots, c_n^*$, we conditionally propose the other parameters from the distributions

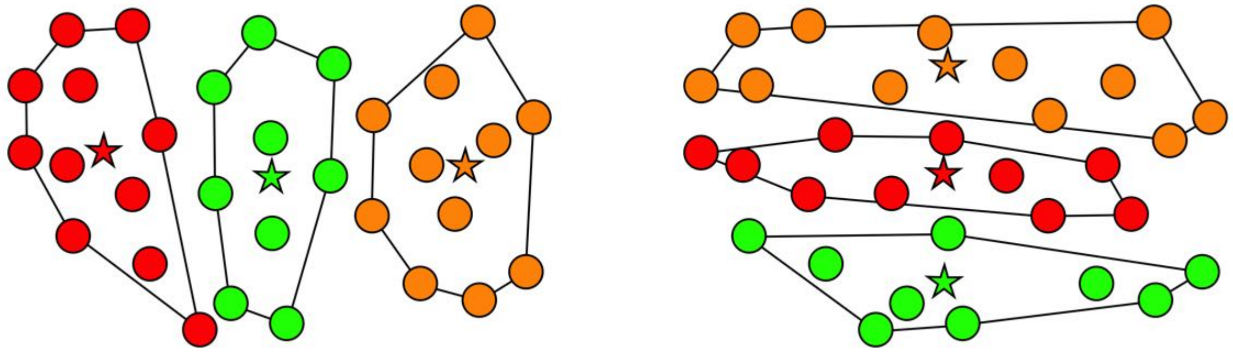
$$\begin{aligned} \boldsymbol{\beta}^{(r)*}|\boldsymbol{\gamma}^{(r)} &\sim Normal(\hat{\boldsymbol{\beta}}^{(r)*}, v_1), \\ \mu^{(r)*} &\sim Normal(\hat{\mu}^{(r)*}, v_2), \\ g^{(r)*} &\sim Normal(\hat{g}^{(r)*}, v_3), \\ \phi^{(r)*} &\sim Normal(\hat{\phi}^{(r)*}, v_4), \\ \tau^{(r)*} &\sim Normal(\hat{\tau}^{(r)*}, v_5), \end{aligned}$$

where $v_1, \ldots, v_5$ are chosen to be very small numbers (e.g. 10^{-8}) and $\hat{\boldsymbol{\beta}}^{(r)*}$, $\mu^{(r)*}$, $\hat{g}^{(r)*}$, $\hat{\phi}^{(r)*}$, and $\hat{\tau}^{(r)*}$ are the maximum likelihood estimates using $\boldsymbol{X}^{(r)*}$ and $\boldsymbol{y}^{(r)*}$, the data now proposed for assignment to cluster r. It must be noted that $\hat{\boldsymbol{\beta}}^{(r)*}$ has the same dimensionality as $\boldsymbol{\beta}^{(r)}$. This constraint prevents the proposed move from being transdimensional. We center proposal distribution at the maximum likelihood estimates to increase the probability that the move is accepted. Using $\boldsymbol{\Sigma}^{(r)*}$ to denote the new covariance matrix under $g^{(r)*}$, $\phi^{(r)*}$, and $\tau^{(r)*}$ and $\boldsymbol{W}^*$ to denote the updated weight matrix, the Metropolis ratio is

$$\alpha = \frac{\prod_{r=1}^{R} L(\boldsymbol{y}^{(r)*}|\mu^{(r)*}, \boldsymbol{\beta}^{(r)*}, \boldsymbol{\Sigma}^{(r)*}) P(\boldsymbol{\mu}^*|\boldsymbol{W}^*, \phi_\beta, \sigma_\beta^2) \prod_{j=1}^{p} P(\hat{\boldsymbol{\beta}}_j^*|\boldsymbol{W}^*, \phi_\beta, \sigma_\beta^2) \prod_{r=1}^{R}}{\prod_{r=1}^{R} L(\boldsymbol{y}^{(r)}|\mu^{(r)}, \boldsymbol{\beta}^{(r)}, \boldsymbol{\Sigma}^{(r)}) P(\boldsymbol{\mu}|\boldsymbol{W}, \phi_\beta, \sigma_\beta^2) \prod_{j=1}^{p} P(\hat{\boldsymbol{\beta}}_j|\boldsymbol{W}, \phi_\beta, \sigma_\beta^2) \prod_{r=1}^{R}} \times \frac{P(g^{(r)*}, \phi^{(r)*}, \tau^{(r)*})}{P(g^{(r)}, \phi^{(r)}, \tau^{(r)})},$$

and we accept the proposal with probability $Min(1, \alpha)$.

The largest advantage to our proposal is the ability to get out of a local mode in the posterior distribution without the need to sequentially move one point at a time. For instance, consider an MCMC chain with current iteration shown in Figure 4.7a. However, suppose that the true cluster assignments are horizontal, rather than vertical. As seen as an example in Figures 4.6a-4.6f, this substantial movement would require many small steps to attain. Instead, we can propose new centroids (Figure 4.7b) and the corresponding cluster assignments to immediately propose the horizontal cluster alignment. We recommend using this proposal periodically in conjunction with the one-at-a-time sampling to traverse the posterior distribution of the cluster assignments.

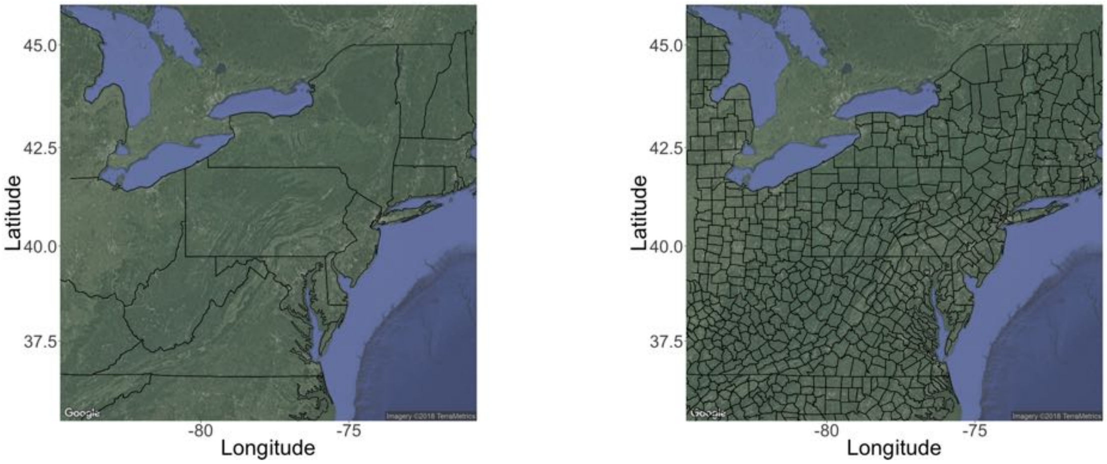


(a) Initial layout

(b) Proposed layout

4.3.2 Commentary on clustering with convex hulls

While we use the CAR model (Section 3.2.5) within the HGP methodology, it is often utilized in situations with areal data, which aggregates data over some spatial unit (e.g. county-level). However, the inference associated with the CAR model is dependent upon the resolution of the areal units. For example, consider Figures 4.7a and 4.7b. Each plot contains states from the mid-Atlantic region (bottom) to New England (top). Figure 4.7a draws boundaries based on the state lines and Figure 4.7b displays county lines. Depending on whether the data were aggregated by state or county, the neighborhood structure and inference would change.



(a) State-level

(b) County-level

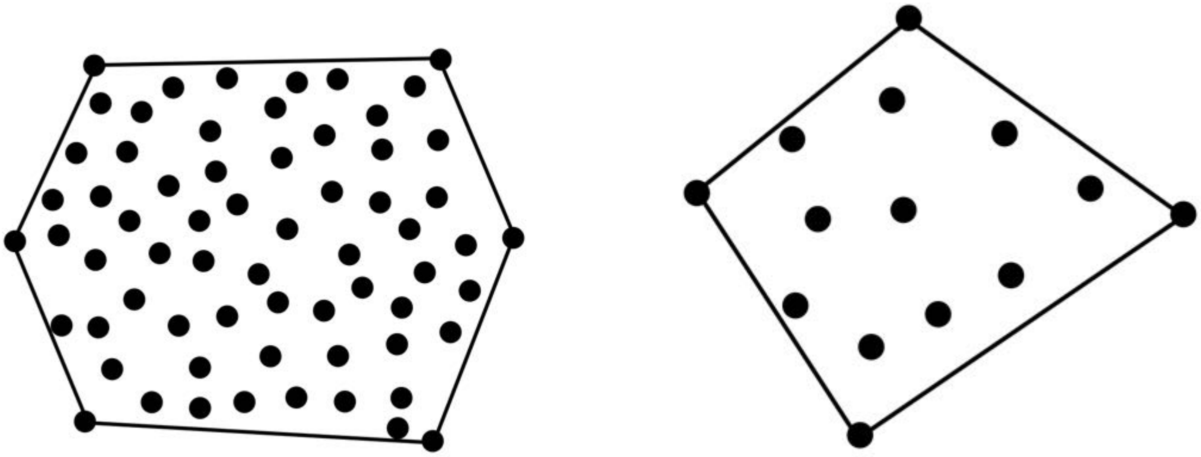
Figure 4.7: A map of the northeastern United States with state-level and county-level resolution.

As part of HGP, we use the spatial coordinates of each observation to form spatially contiguous clusters as defined by convex hulls. The non-zero linear effects across clusters follow the CAR model (Section 3.2.5). The cluster assignments are jointly inferred with the other unknown parameters and, therefore, we do not have to specify a level of resolution for the CAR model. The convex hulls form bins for the observations and the Gaussian process mixture allows observations to shift from bin to bin according to the data. In essence, we are able to learn the correct level of resolution for our CAR model.

Another consequence of the convex hull clustering is the effect of outliers within a cluster. Since a point that is not on the convex hull cannot shift cluster assignments (Section 4.2.2), the conditional likelihood of that point does not play an immediate role in the cluster assignment. If an observation is not predicted well given the model in its current cluster, the convex hull constraint will prohibit movement unless the move will not cause overlapping clusters. While other techniques for Gaussian process mixtures exist that would allow interior points to move (e.g. Rasmussen & Ghahramani (2002)), HGP enforces a hard contraint to ensure non-overlapping clusters.

4.3.3 Scalability

Within the HGP methodology, there are two sets of parameters for which inference is computationally intensive: 1) the Gaussian process covariance parameters and 2) the cluster membership indicators. Sampling from the full conditional distribution for $\phi^{(r)}$ and $g^{(r)}$ (Appendix A) requires the determinant and inverse of $\mathbf{\Sigma}^{(r)}$, which is an $O(n_r^3)$ operation. These need to be sampled for each for R clusters, however this is parallelizable and R is generally a small number. Similarly, the predictive distribution for $y_i^{(r)}$ (Equation 4.2), has computation complexity $O(n_r^2)$ and needs to be repeated for each observation being considered for reassignment. However, as the number of observations in a cluster increases, a smaller proportion is eligible for move at any given time (Figures 4.8a and 4.8b).



(a) Dense cluster

(b) Sparse cluster

Figure 4.8: Examples of dense and sparse clusters. A smaller proportion of points are on the convex hull of the dense cluster.

Consequently, from a computational standpoint, the matrix inversions and determinant calculations are the biggest computational limitations. Assuming, on average, n/R points per cluster, the limitations depend on both the total number of observations and the number of clusters. The number of observations n can be higher when R is selected to be larger. However, R must be iterated over to choose an optimal number of clusters. Therefore, we recommend using the HGP methodology for moderately sized datasets (e.g. $n < 5,000$) depending on the computing power available.

4.4 Estimating the Ising model parameter

In Section 2.4.1 we presented a broad overview of the Ising model and its use in modeling correlated binary data. In Section 3.2.2, we discussed the prior distribution specification of the variable inclusion indicators $\gamma_j^{(1)}, \dots, \gamma_j^{(R)}$ for $j = 1, \dots, p$. Recall the prior distribution

$$P(\gamma_j^{(1)}, \dots, \gamma_j^{(R)})' = \frac{1}{\Omega(\theta)} exp\bigg\{\theta \sum_{r \sim r'} w_{r,r'} \delta\big(\gamma_j^{(r)} = \gamma_j^{(r')}\big)\bigg\} \text{ for } j = 1, \dots, p,$$

where $\Omega(\theta)$ is the normalizing constant that involves summing over all 2^R grid layouts. For small R, this is feasible, but in a general framework this is too computationally demanding. While we will use the true Ising model for the estimation of the variable inclusion indicators (Section 3.2.5), to estimate θ we will use the pseudo-likelihood function. The pseudo-likelihood for the Ising model is given by

$$L(\theta|\gamma_j^{(1)}, \dots \gamma_j^{(r)}) \propto \prod_{r=1}^{R} p_r^{\gamma_j^{(r)}} (1 - p_r)^{1-\gamma_j^{(r)}},$$

where $p_r = \big(1 + exp\{\theta \sum_{r \sim r'} w_{r,r'}(\delta(\gamma_j^{(r')} = 0) - \delta(\gamma_j^{(r')} = 1))\}\big)^{-1}$. The parameter θ is no longer embedded in a large summation over grids and can be estimated more easily. While we no longer need to calculate $\Omega(\theta)$, we must be careful with the propriety of the posterior distribution throughout the MCMC sampling.

Traditional applications of the Ising model involve a collection of fixed and known binary responses, including both ones and zeroes. The variable inclusion indicators, however, are unknown and updated with each sample from the posterior distribution. As a result, it is not implausible for one posterior sample to have no variability (i.e. $\gamma_j^{(r)} = 1$ for all $j = 1, \dots, p$,

$r = 1, \ldots, R$). In this case the posterior distribution for θ is improper if we assume a flat prior distribution, $P(\theta) \propto 1$ (Proof in Appendix B). To utilize the computational savings of the pseudo-likelihood while ensuring a proper posterior distribution for θ, we choose a proper prior distribution for θ in the Gamma family.

Since we only consider positive values of θ, implying positive spatial correlation, we choose the Gamma distribution as a prior. We consider the Gamma distribution with parameterization

$$P(x) \propto x^{a-1} e^{-bx}$$

for $x > 0$. To specify the hyperparameters a and b in our Gamma prior, we first fix a and then tune b based on a criterion. We analyze the ability to accurately estimate θ under both $a = 1$ and $a = 1.5$ in a simulation study in Section 5.1.1. Here, we restrict our discussion to the procedure for choosing b conditional upon a.

Without loss of generality, we consider $\gamma_j^{(1)}$, the indicator with the largest associated weights, $\sum_{r' \sim 1} w_{1,r'}$. The conditional probability $P(\gamma_j^{(1)} = 1 | \gamma_j^{(2)}, \ldots, \gamma_j^{(R)})$ is given by

$$p_1 = \left(1 + exp\{\theta \sum_{r' \sim 1} w_{1,r'} (\delta(\gamma_j^{(r')} = 0) - \delta(\gamma_j^{(r')} = 1))\}\right)^{-1}.$$

When $\gamma_j^{(2)} = 1, \ldots, \gamma_j^{(R)} = 1$, the conditional probability for $\gamma_j^{(1)}$ simplifies to

$$\tilde{p}_1 = \left(1 + exp\{-\theta \sum_{r' \sim 1} w_{1,r'}\}\right)^{-1},$$

which is the largest value p_1 can attain. Clearly, the value of θ is dependent on the scale of the weights $w_{1,r'}$. To determine the Gamma hyperparameter b, we consider values of θ to prevent $\tilde{p}_1$ from becoming exactly one. The procedure is as follows:

First, we solve for θ_{max} by bounding $\tilde{p}_r$ by p_{max}, as detailed by:

$$\theta_{max} = -\Big(\sum_{r' \sim 1} w_{1,r'}\Big)^{-1} Log\Big(p_{max}^{-1} - 1\Big).$$

Then, given θ_{max}, we solve for the optimal value b that places $(1 - p_{tail})\%$ of the Gamma distribution mass below θ_{max}:

$$b = \underset{b}{\text{ArgMin}} \ \Big| \int_b^\infty Gamma(x|a = 1.5, b)dx - p_{tail}\Big|.$$

While the choices for p_{max} and p_{tail} are arbitrary, we choose $p_{max} = 0.99$ and $p_{tail} = 0.05$ to place substantial mass on values for θ that does not allow conditional probabilities rise above 0.99. These values perform well in our simulations in Section 5.1.1.

Combining our chosen prior distribution with the p Ising models for $(\gamma_j^{(1)}, \dots, \gamma_j^{(R)})'$ for $j = 1, \dots, p$, we can write the full conditional distribution for θ as

$$P(\theta|\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_p) \propto \prod_{j=1}^{p} \Big\{ \prod_{r=1}^{R} p_r^{\gamma_j^{(r)}} (1 - p_r)^{1-\gamma_j^{(r)}} \Big\} \theta^{a-1} e^{-b\theta},$$

where $p_r = \big(1 + exp\{\theta \sum_{r \sim r'} w_{r,r'}(\delta(\gamma_j^{(r')} = 0) - \delta(\gamma_j^{(r')} = 1))\}\big)^{-1}$. While the full conditional distribution is not a recognizable form that can utilize a Gibbs step, it does not require any difficult computations. A Metropolis-Hasting step is easy to implement with any proposal distribution with support $\mathbb{R}^+$.

Chapter 5

Simulations

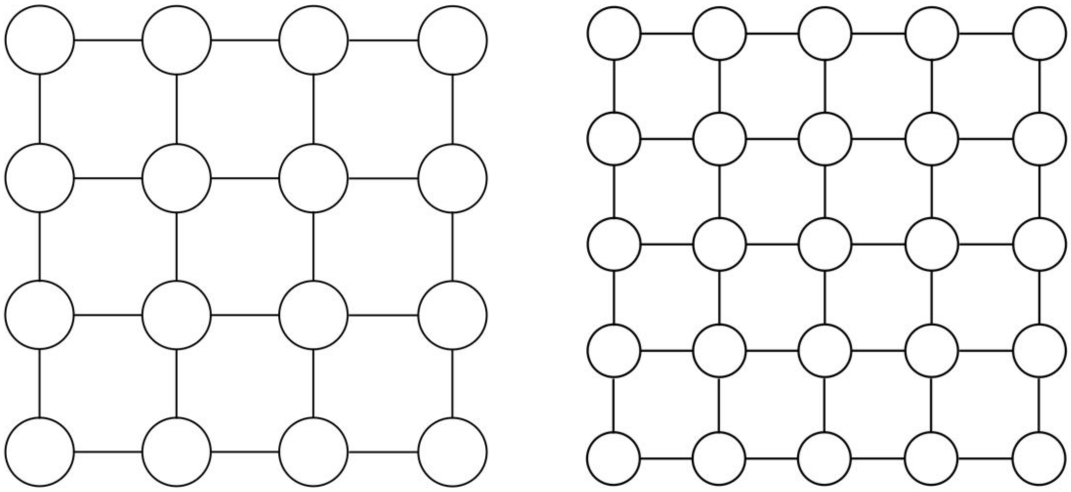
The HGP methodology has several layered components that combine to provide spatially dependent model selection. Before applying our methodology to real data, we perform controlled simulations to evaluate the benefits and performance of the individual pieces. In this section, we present comprehensive simulation studies to compare and contrast different methods for estimating the Ising parameter θ (Section 5.1), specifying a prior distribution for the linear effects (Section 5.2), and specifying a prior distribution for the prior probability for variable inclusion (Section 5.3).

5.1 Estimation of the Ising parameter

We now compare various methods for estimating the spatial strength parameter θ associated with the Ising distribution. We simulate 100 realizations from the Ising distribution under varying grid sizes and true values of θ. For each simulated dataset, we estimate θ using the posterior median with different model assumptions. Before presenting results, we first discuss the simulation process and the competing methods for estimating θ.

5.1.1 Simulating binary random fields

In practice, we expect the number of spatial clusters to be relatively small ($\approx 3 - 15$) in most settings. Consequently, we focus our simulations on smaller binary random fields. We simulate square grids with sizes $n = 4$, $n = 9$, $n = 16$, and $n = 25$ (Figure 5.1). For each grid, the weight matrix $\boldsymbol{W}$ is specified by: $w_{i,i'} = 1$ if vertices i and i' are connected with an edge and $w_{i,i'} = 0$ otherwise. Given this choice for $\boldsymbol{W}$, the sum of the weights for node i is $2 \leq \sum_{i'} w_{i,i'} \leq 4$.


(a) 4×4 grid ($n = 16$)

(b) 5×5 grid ($n = 25$)

Figure 5.1: Visualizations of the size $n = 16$ and $n = 25$ grids used in the simulation.

To choose true values of θ for the simulation, we analyze the conditional distribution $\tilde{p}_i = P(y_i = 1|y_{i'} = 1, i \sim i') = \left(1 + e^{-\theta \sum_{i'} w_{i,i'}}\right)^{-1}$. For a node with the highest total spatial weight, $p_i = \left(1 + e^{-4\theta}\right)^{-1}$ Therefore, use the values $\theta \in \{0, .275, .55, 1.15\}$, which corresponds to $\tilde{p}_i \in \{.50, .75, .90, .99\}$. For each combination of grid size and θ, we generate 100 realizations.

5.1.2 Methods of estimation

For each realization, we sample from the posterior distribution $P(\theta|y_1,\ldots,y_n)$ using four methods.

1. Pseudo-likelihood with a Gamma distribution prior on θ:

$$P(\theta|y_1,\ldots y_n) \propto \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}\theta^{a-1}e^{-b\theta}$$

where $p_i = \left(1+exp\{\theta\sum_{i\sim i'} w_{i,i'}(\delta(y_{i'}=0)-\delta(y_{i'}=1))\}\right)^{-1}$. We set the hyperparameter $a=1.5$ to obtain a dispersed Gamma distribution with a nonzero mode. The rate parameter b is chosen based on the methodology of Section 4.4.

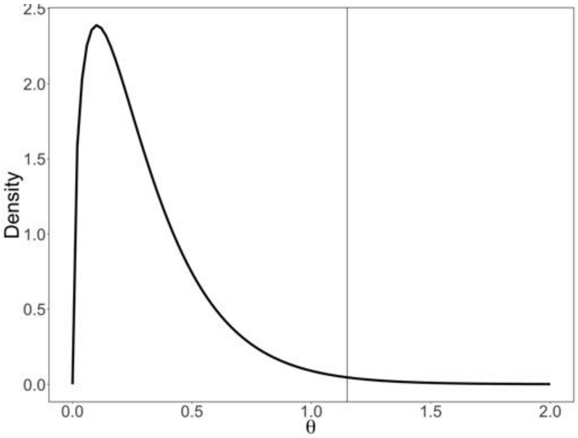


Figure 5.2: Gamma distribution with shape $a=1.5$ and rate $b=4.93$, corresponding to $p_{max}=.99$ and $p_{tail}=.05$ for weights based on edge-sharing on a grid. The vertical line is drawn at θ_{max}.

2. Pseudo-likelihood with an Exponential prior distribution on θ:

$$P(\theta|y_1,\ldots y_n) \propto \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}e^{-b\theta}$$

where $p_i = \left(1+exp\{\theta\sum_{i\sim i'} w_{i,i'}(\delta(y_{i'}=0)-\delta(y_{i'}=1))\}\right)^{-1}$. The hyperparameter b is computed using the same logic as the Gamma distribution, except that $a=1$ rather than $a = 1.5$. This choice places significantly more mass near 0 than the Gamma distribution (Figure 5.3).

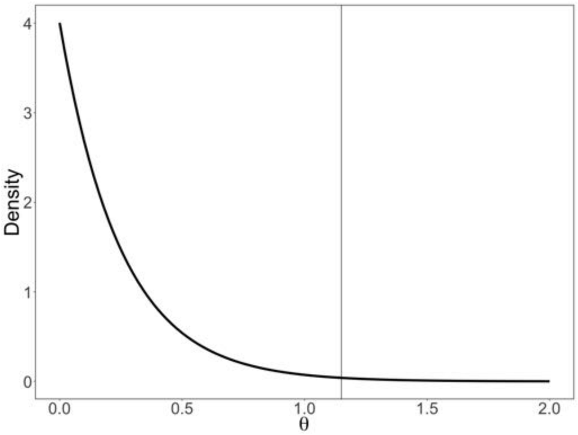


Figure 5.3: Exponential distribution with rate $b = 4$, corresponding to $p_{max} = .99$ and $p_{tail} = .05$ for weights based on edge-sharing on a grid. The vertical line is drawn at θ_{max}.

3. Pseudo-likelihood with a flat prior distribution on θ:

$$P(\theta|y_1, \dots y_n) \propto \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

where $p_i = \big(1 + exp\{\theta \sum_{i\sim i'} w_{i,i'}(\delta(y_{i'} = 0) - \delta(y_{i'} = 1))\}\big)^{-1}$. This prior distribution places no additional mass on small values of θ.

4. Ising likelihood with a flat prior distribution on θ:

$$P(\theta|y_1, \dots, y_n) \propto \frac{1}{\Omega(\theta)} exp\Big\{\theta \sum_{i=1}^{n} \sum_{i\sim i'} w_{i,i'}\delta(y_{i'} = y_i)\Big\}$$

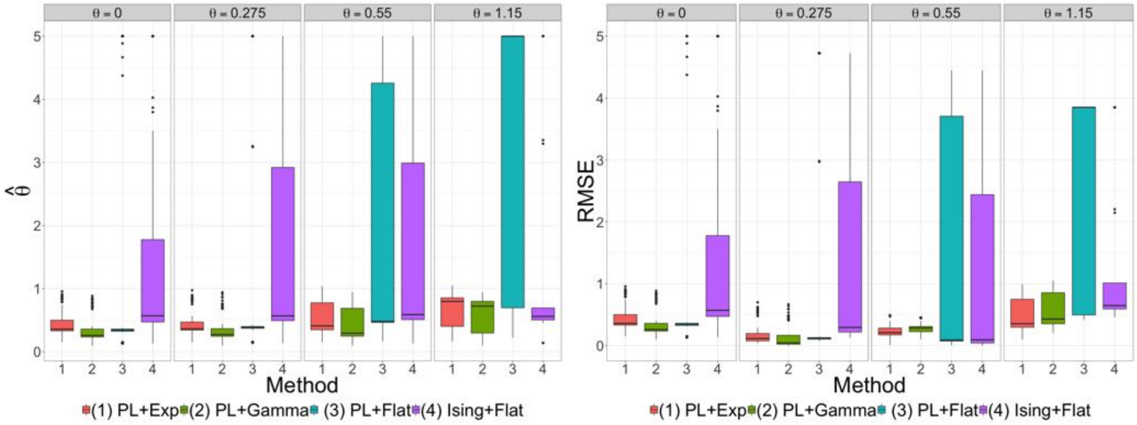
where $\Omega(\theta)$ is an intractable normalizing constant that depends on θ. While $\Omega(\theta)$ cannot be easily computed, the ratio of normalizing constants $\Omega(\theta')/\Omega(\theta)$ that appears in the Metropolis-Hastings acceptance ratio can be approximated using importance sampling. We approximate this ratio using 500 samples from an Ising distribution using the true value θ.

5.1.3 Simulation results

2 × 2 grid :

For a 2 × 2 grid, each simulation only provides a sample size of $n = 4$. As expected, each method struggles the most with estimating θ with this setup (Figures 5.4a and 5.4b). However, the pseudo-likelihood methods with proper prior distributions perform more consistently than its counterparts. While both the the pseudo-likelihood with flat prior appears to give reasonable results when $\theta = 0$, the boxplot are somewhat misleading. For 16 of the 100 simulations, the resulting grids led to improper posterior distributions under the pseudo-likelihood with flat prior. These results are removed from the boxplots. This occurs when a simulated grid has no variance in $\boldsymbol{y}$ (proof can be found in Section B). Similarly, the true Ising model is also improper when a grid has no variance in the response.

For true values $\theta \in \{0, .275, .55, 1.15\}$, the simulations resulted in improper posterior distributions for the pseudo-likelihood and Ising model with flat priors 16%, 16%, 30%, and 76% of simulations, respectively (Table 5.1).



(a) θ vs. $\hat{\theta}$

(b) θ vs. root mean-squared error

Figure 5.4: Simulation results on a 2 × 2 grid where the true model is Ising with spatial strength parameter θ (x-axis). θ was estimated with one of four models to find the posterior median (left) and root mean-squared error (right).

Both the pseudo-likelihood with Gamma and Exponential priors performed similarly. The

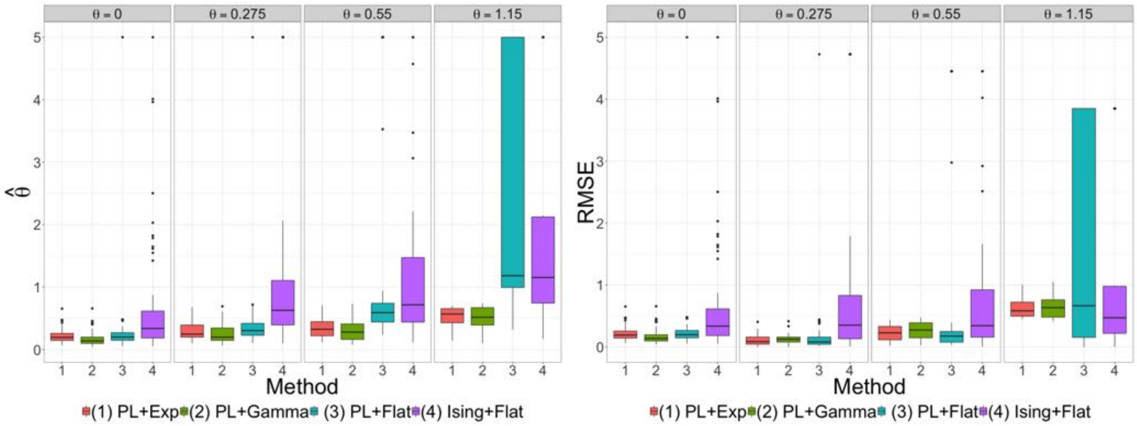
Exponential prior distribution did outperform the Gamma distribution for smaller values of θ, which is expected since the Exponential distribution places more mass near 0. As θ increased towards its maximum value, both methods had similar root mean-squared error values, on average.

Table 5.1: Counts (per 100 simulations) of grids with no variability in the response.

	Grid Size			
$\boldsymbol{\theta}$	4	9	16	25
0	16	1	0	0
.275	15	1	0	0
.55	27	10	1	0
1.15	67	47	30	21

3 × 3 grid :

The 3×3 grid provided more than twice the sample size as the 2×2 grid, leading to much improved estimation for most methods. With the larger grid size, it was much more uncommon to observe grids with no variability (Table 5.1). This only occurred 1%, 1%, 7%, and 45% for $\theta \in \{0, .275, .55, 1.15\}$, respectively. Consequently, the pseudo-likelihood with flat prior did a much better job at estimating θ than the 2×2 grid case. The Gamma and Exponential priors obtained very similar estimates for θ (Tables 5.5a and 5.5b). As expected, with more data available, the Ising model had the lowest root mean-squared error (on average), when the posterior is proper.



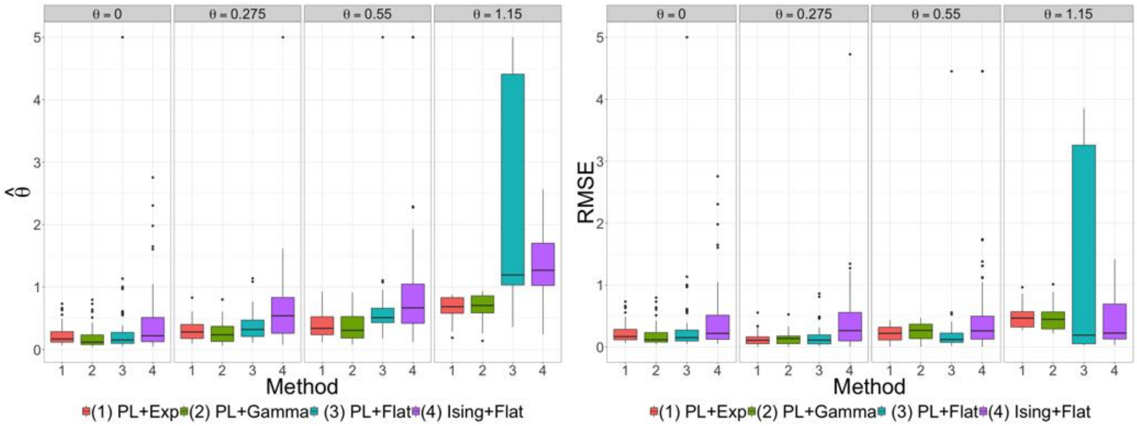
(a) θ vs. $\hat{\theta}$

(b) θ vs. root mean-squared error

Figure 5.5: Simulation results on a 3×3 grid where the true model is Ising with spatial strength parameter θ (x-axis). θ was estimated with one of four models to find the posterior median (left) and root mean-squared error (right).

4 × 4 grid :

Once the grid size reaches $n = 16$ total observations, it becomes uncommon to observe grids with no variability unless θ is large. For grids with variation in the response, the Ising distribution with flat prior generally outperforms the other competitors. However, for $\theta \in \{0, .275, .55\}$, there is not much of a difference between the Ising distribution with flat prior and the pseudo-likelihood with Gamma and Exponential priors (Table 5.6b).



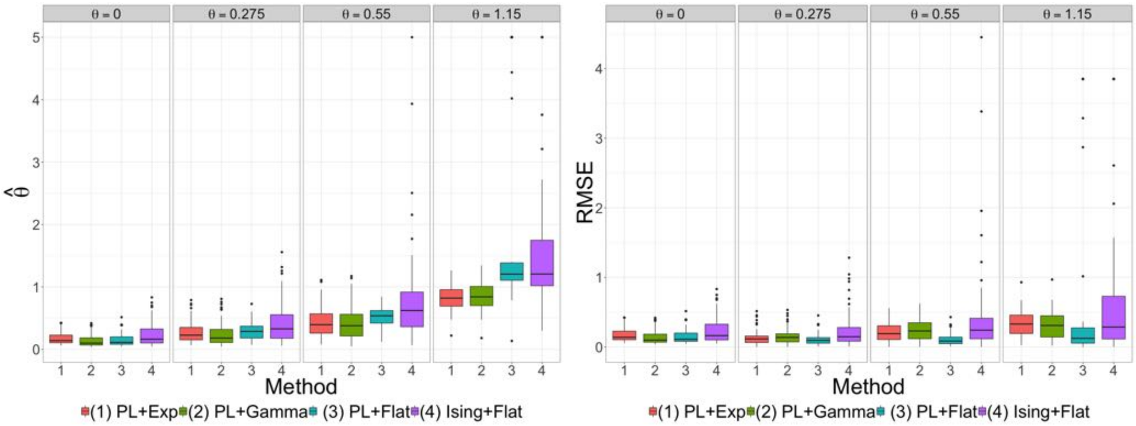
(a) θ vs. $\hat{\theta}$

(b) θ vs. root mean-squared error

Figure 5.6: Simulation results on a 4 × 4 grid where the true model is Ising with spatial strength parameter θ (x-axis). θ was estimated with one of four models to find the posterior median (left) and root mean-squared error (right).

5 × 5 grid :

For grids with variability in the response, the Ising model with flat prior outperforms all other methods across θ. However, even with grids of size $n = 25$, when $\theta = 1.15$ we observe 21/100 degenerate grids for which the pseudo-likelihood with flat prior and Ising distribution with flat prior lead to improper posterior distributions. The root mean-squared error for both the pseudo-likelihood with Exponential and Gamma prior distribution is reasonably close to the Ising distribution with flat prior (Table 5.7b) while not leading to improper posterior distributions nor requiring importance sampling.



(a) θ vs. $\hat{\theta}$

(b) θ vs. root mean-squared error

Figure 5.7: Simulation results on a 5×5 grid where the true model is Ising with spatial strength parameter θ (x-axis). θ was estimated with one of four models to find the posterior median (left) and root mean-squared error (right).

Summary of results:

As expected, when a simulated grid has variation in the response, using the true Ising distribution with flat prior outperforms all other methods. However, the Ising distribution requires importance sampling to estimate the normalizing constant, which increases the computational burden. Also, for small grids, it is not uncommon to see realizations that are degenerate. The pseudo-likelihood with flat prior does not perform well in most scenarios and suffers from an improper posterior distribution in many simulations. The pseudo-likelihood with Exponential and Gamma priors both outperform their flat prior counterpart across all configurations. While Gamma and exponential priors have similar performance when θ is large, the additional mass that the Exponential distribution places near 0 leads to superior performance when θ is smaller.

5.2 Independent-SSVS versus CAR-SSVS

5.2.1 Simulating data

To isolate the difference in variable selection performance between Independent-SSVS and CAR-SSVS, we run simulations with various settings of total sample size n, number of clusters R, number of possible covariates per cluster p, proportion of sparsity p_j, spatial strength in the CAR model ϕ_β, and the underlying mean level of the nonzero coefficients $E\left[\beta\right]$. We discuss the simulation procedure below.

1. Given n, generate a square grid of size $\sqrt{n} \times \sqrt{n}$.

2. Generate R cluster centroids and n cluster assignments using K-means algorithm.

3. Given the cluster centroids, calculate the CAR weight matrix $\boldsymbol{W}$, where $w_{r,r'} = 1/d_{r,r'}$ for $r \neq r'$, $w_{r,r} = 0$, and $d_{r,r'}$ is the Euclidean distance between the centroids of clusters r and r'.

4. Using the maximum eigenvalue of $\boldsymbol{W}$, λ_{max}, let the CAR parameter $\phi_\beta \in \{0, .48\lambda_{max}^{-1}, .975\lambda_{max}^{-1}\}$ to simulate a variety of spatial strength in the nonzero coefficients.

5. Choose σ_β^2 such that the CAR covariance matrix $\boldsymbol{K}_0 = \sigma_\beta^2(\boldsymbol{I} - \phi_\beta \boldsymbol{W})^{-1}$ has diagonal elements 1.

6. Simulate indicators for each cluster and each covariate $\gamma_j^{(r)} \sim Bernoulli(p_j)$ for $j = 1, \ldots, p$ and $r = 1, \ldots, R$.

7. Simulate the vector of intercepts $\boldsymbol{\mu} \sim Normal(E\left[\beta\right]\mathbf{1}, \boldsymbol{K}_0)$.

8. For each covariate j such that $\gamma_j^{(r)} = 1$, simulate $\{\beta_j^{(r)}|\gamma_j^{(r)} = 1\} \sim Normal(E\left[\beta\right]\mathbf{1}, \boldsymbol{K}_j)$.

9. For each cluster, simulate a covariate matrix $\boldsymbol{X}^{(r)}$ such that each element $x_{ij} \sim Normal(0, 1)$.

10. For each cluster, simulate the response vector $\boldsymbol{y}^{(r)} \sim Normal(\mu^{(r)} + \boldsymbol{X}^{(r)}\boldsymbol{\beta}^{(r)}, \boldsymbol{I})$.

For each combination, we simulate 100 datasets. It is assumed that we know the true values for all parameters except for ϕ_β, σ^2_β, $\boldsymbol{\mu}$, each $\gamma_j^{(r)}$, and each $\beta_j^{(r)}$. The goal of our inference is to correctly estimate the variable inclusion indicators $\gamma_j^{(r)}$ for $r = 1, \dots, R$ and $j = 1, \dots, p$.

5.2.2 Methods of estimation and evaluation

For each simulated dataset, we use separate methods to infer the unknown parameters: Independent-SSVS and CAR-SSVS. For Independent-SSVS, we assume the prior distribution $(\beta_j^{(r)}|\gamma_j^{(r)} = 1) \sim Normal(0, 1)$. Therefore, for this model we do not need to estimate σ^2_β or ϕ^2_β. For CAR-SSVS, we assume the prior distribution $\{\beta_j^{(r)}|\gamma_j^{(r)} = 1\} \sim Normal(\mathbf{0}, \boldsymbol{K}_j)$, where $\boldsymbol{K}_j$ is determined by σ^2_β and ϕ^2_β.

To evaluate the performance of each technique, we use the average fraction correctly classified for fit (AFCCF) (Wilkinson 1999) on the variable inclusion indicators $\gamma_j^{(r)}$. The formula for AFCCF is

$$AFCCF = \frac{1}{pR}\Big[\sum_{r=1}^{R}\sum_{j=1}^{p}\gamma_j^{(r)}P(\gamma_j^{(r)} = 1|\boldsymbol{y}) + (1-\gamma_j^{(r)})P(\gamma_j^{(r)} = 0|\boldsymbol{y})\Big],$$

where $P(\gamma_j^{(r)} = 1|\boldsymbol{y})$ is the posterior marginal probability that $\gamma_j^{(r)} = 1$. When a method perfectly infers the unknown variable inclusion indicators, $AFCCF = 1$. At its minimum, $AFCCF = 0$. In Section 5.2.3, we compare AFCCF for both methods for the settings listed in Table 5.2.

Table 5.2: Configurations for each simulation. λ_{max} denotes the largest eigenvalue of $\boldsymbol{W}$, which is determined by R.

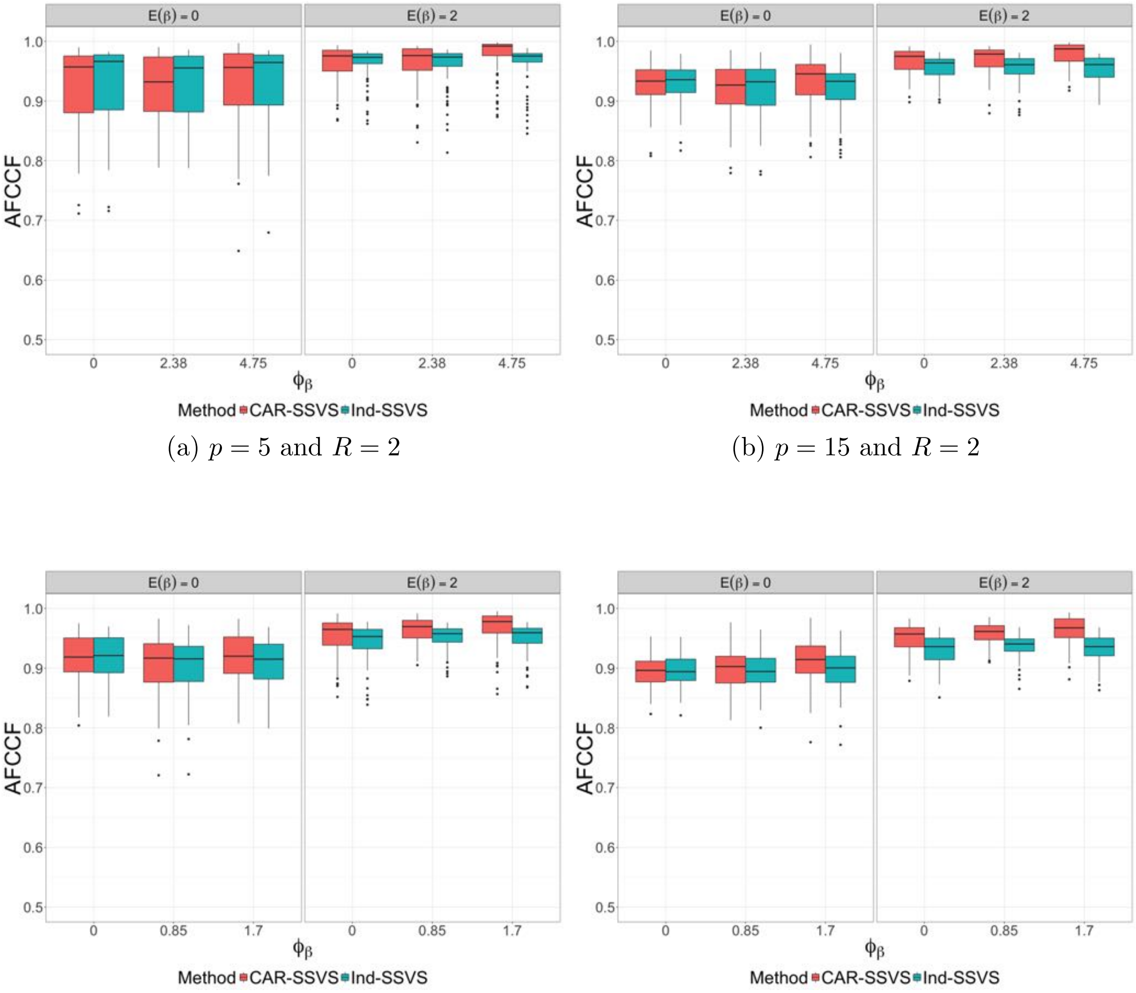
Parameter	**Values**
n	$\{100, 225\}$
p	$\{5, 15\}$
R	$\{2, 4, 8\}$
p_j	$\{.1, .5, .9\}$
ϕ_β	$\{0, .48\lambda_{max}^{-1}, .975\lambda_{max}^{-1}\}$
$E\left[\beta\right]$	$\{0, 2\}$

5.2.3 Simulation results

We organize the simulation results in the following manner. Each grouping of six plots corresponds to a combination of n and p_j. Given n and p_j, the left and right columns are for $p = 5$ and $p = 15$, respectively. The rows represent $R = 2$, $R = 4$, and $R = 8$. Since the most notable difference between methods occurs as a function of $E\left[\beta\right]$, each plot is segmented by $E\left[\beta\right] = 0$ versus $E\left[\beta\right] = 2$.

Sample size $n = 100$ with inclusion probability $p_j = 0.1$:

With a sample size of $n = 100$, each cluster has between 50 observations for $R = 2$ clusters and ≈ 12 observations for $R = 8$ clusters. However, since $p_j = 0.1$, only a small proportion of covariates are actually active. When there are only $p = 5$ covariates, both Independent-SSVS and CAR-SSVS perform similarly, especially when $E[\beta] = 0$ (Tables 5.8a, 5.8c, and 5.8e). However, when $p = 5$ and $E[\beta] = 2$, CAR-SSVS has higher AFCCF. We see similar results when $p = 15$ (Tables 5.8b, 5.8d, and 5.8f); both methods have similar AFCCF, except when $E[\beta] = 2$. Across all scenarios, CAR-SSVS tends to perform better as ϕ_β increases.

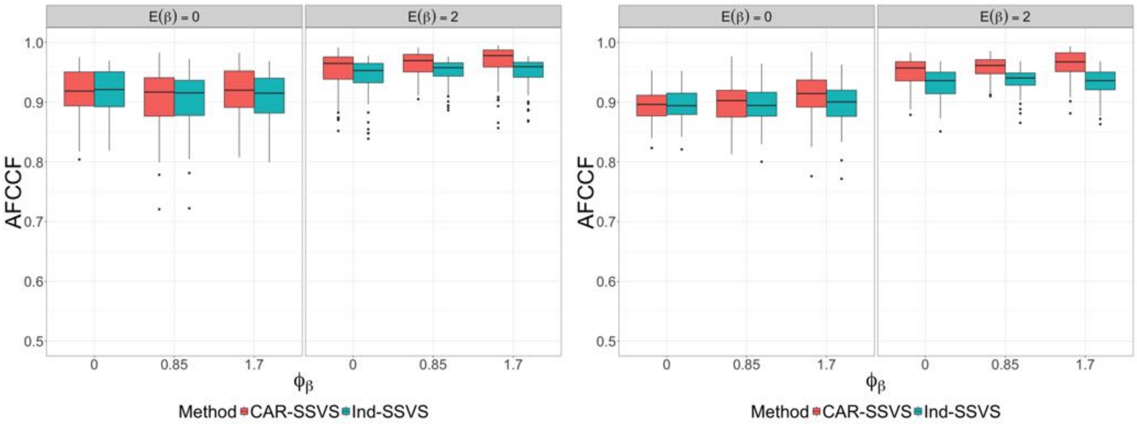


(a) $p = 5$ and $R = 2$

(b) $p = 15$ and $R = 2$

(c) $p = 5$ and $R = 4$

(d) $p = 15$ and $R = 4$

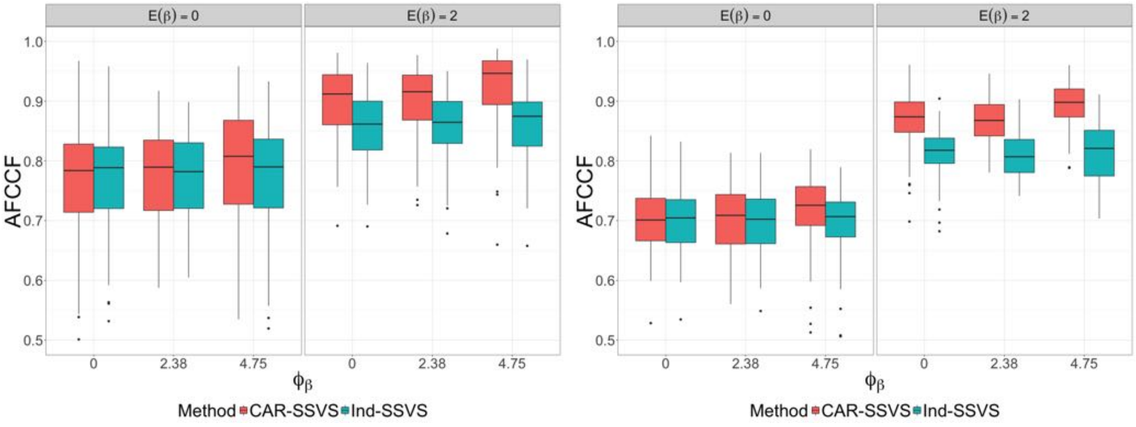


(e) $p = 5$ and $R = 8$ (f) $p = 15$ and $R = 8$

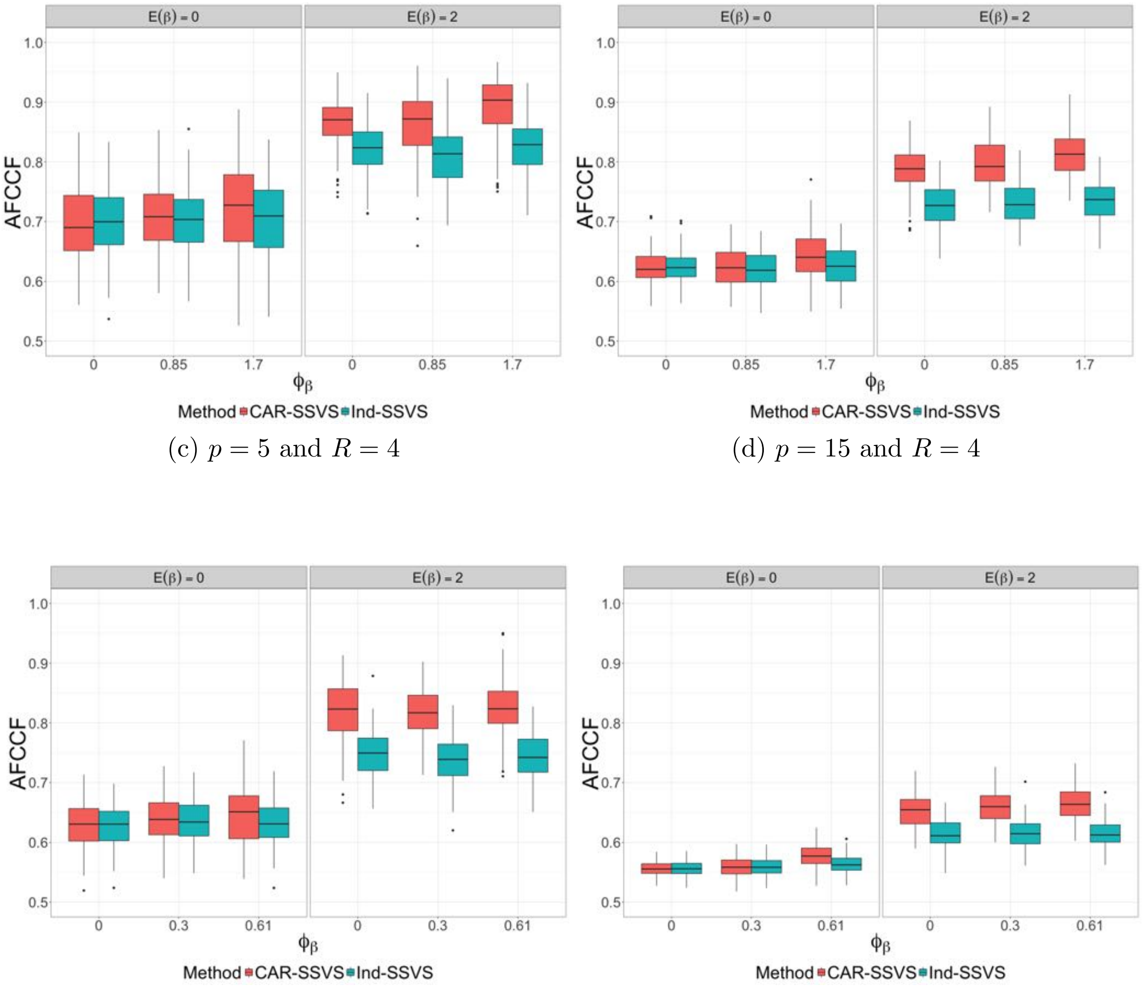
Figure 5.8: Boxplots for AFCCF for Independent-SSVS and CAR-SSVS with sample size $n = 100$ and inclusion probability $p_j = 0.1$.

Sample size $n = 100$ with inclusion probability $p_j = 0.5$:

As with the $n = 100$ $p_j = 0.1$ configuration, the most drastic difference between methods comes when $E[\beta] = 2$ rather than $E[\beta] = 0$. The separation between methods is stronger when $p = 15$ (Tables 5.9b, 5.9d, and 5.9f), since Independent-SSVS has the burden of estimating more parameters without the additional prior information that comes from inter-cluster sharing.



(a) $p = 5$ and $R = 2$ (b) $p = 15$ and $R = 2$



(c) $p = 5$ and $R = 4$

(d) $p = 15$ and $R = 4$

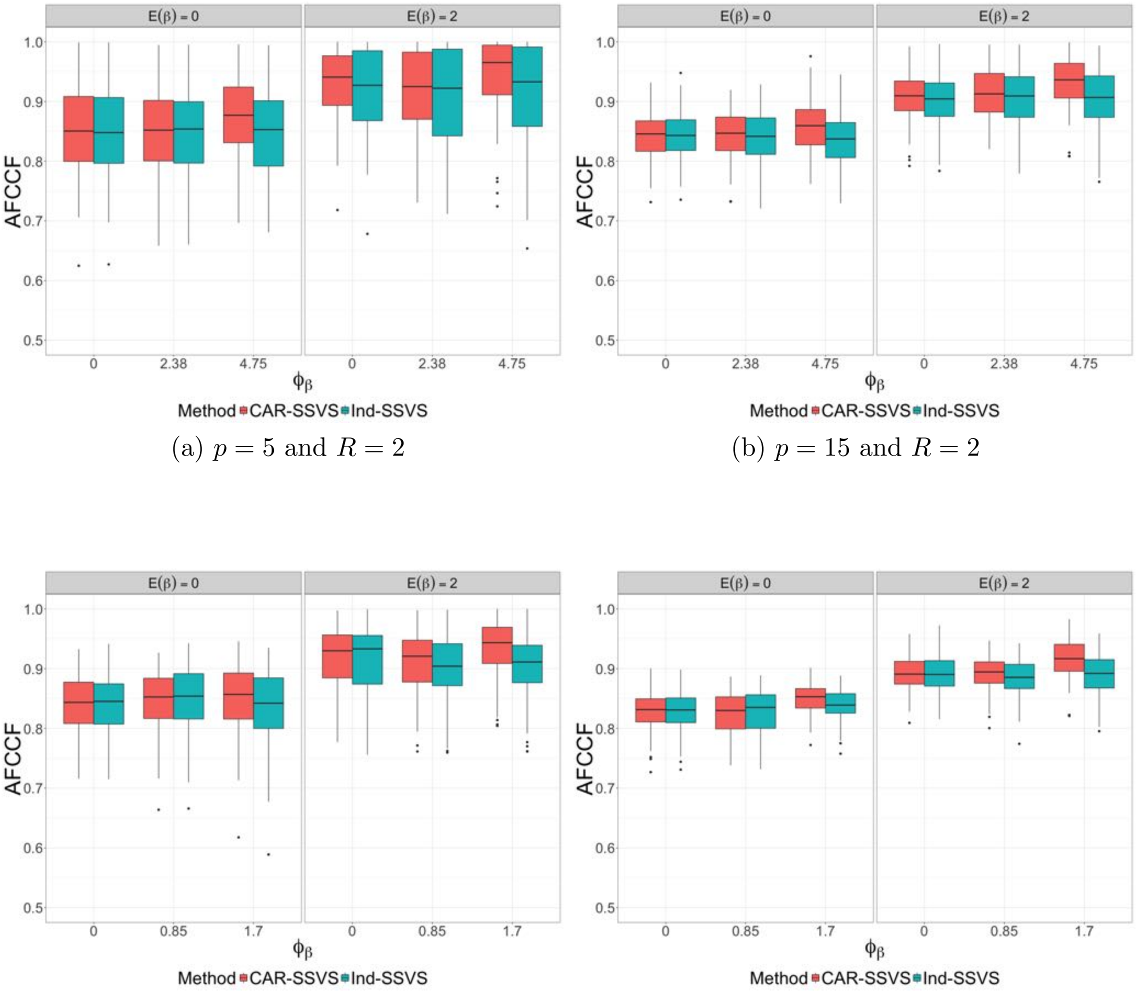
(e) $p = 5$ and $R = 8$

(f) $p = 15$ and $R = 8$

Figure 5.9: Boxplots for AFCCF for Independent-SSVS and CAR-SSVS with sample size $n = 100$ and inclusion probability $p_j = 0.5$.

Sample size $n = 100$ with inclusion probability $p_j = 0.9$:

Overall, CAR-SSVS does not provide much of a benefit when the probability of inclusion $p_j = 0.9$. The only separation occurs at high levels of ϕ_β. With $p_j = 0.9$, the model space is dense and there are not many opportunities to find unimportant covariates. CAR-SSVS does best when some clusters have large effects for a given covariate, while the covariate has no effect in others. It is this type of discontinuity that allows CAR-SSVS to identify unimportant covariates. However, this situation does not arise often when most covariates are active in all clusters.

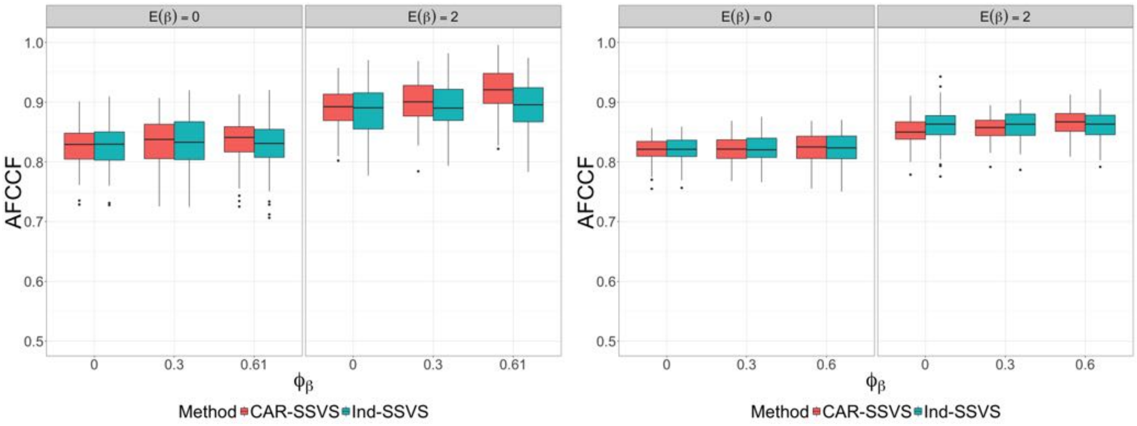


(a) $p = 5$ and $R = 2$

(b) $p = 15$ and $R = 2$

(c) $p = 5$ and $R = 4$

(d) $p = 15$ and $R = 4$



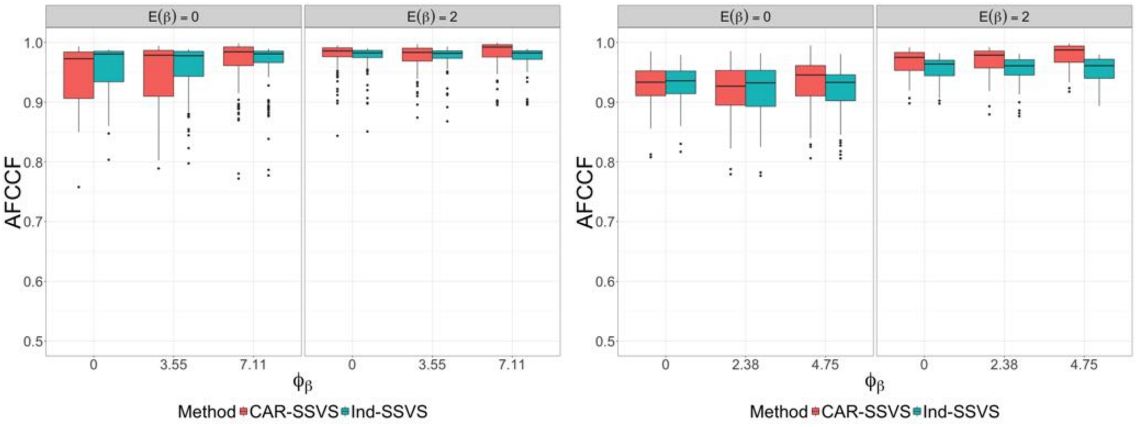
(e) $p = 5$ and $R = 8$

(f) $p = 15$ and $R = 8$

Figure 5.10: Boxplots for AFCCF for Independent-SSVS and CAR-SSVS with sample size $n = 100$ and inclusion probability $p_j = 0.9$.

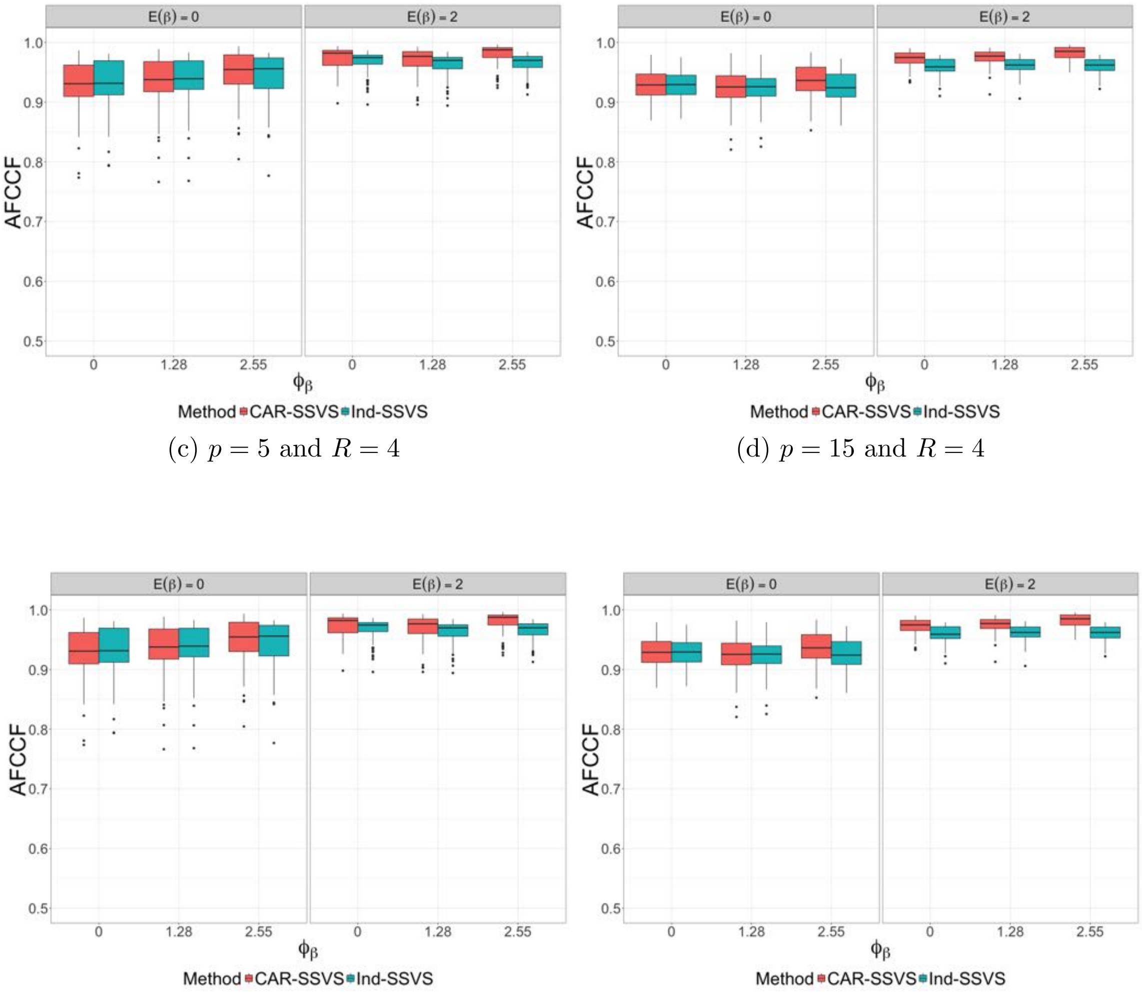
Sample size $n = 225$ with inclusion probability $p_j = 0.1$:

The combination of increasing overall sample size to $n = 225$ with the sparse signal space ($p_j = 0.1$) leads to a configuration with few covariates and enough data to estimate each well. As was the case when $n = 100$, CAR-SSVS performs better when $E[\beta] = 2$, but it is not a significant improvement.



(a) $p = 5$ and $R = 2$

(b) $p = 15$ and $R = 2$



(c) $p = 5$ and $R = 4$

(d) $p = 15$ and $R = 4$

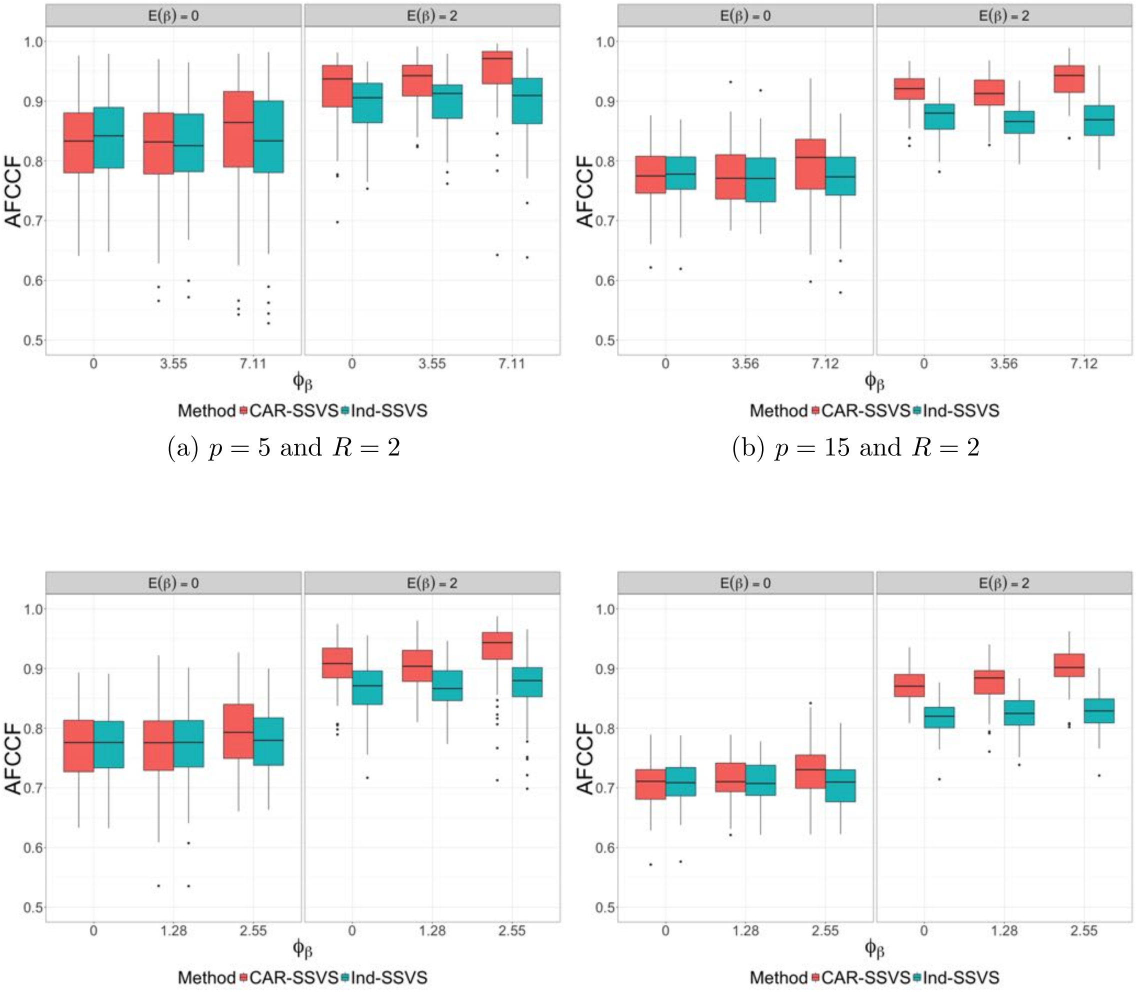
(e) $p = 5$ and $R = 8$

(f) $p = 15$ and $R = 8$

Figure 5.11: Boxplots for AFCCF for Independent-SSVS and CAR-SSVS with sample size $n = 225$ and inclusion probability $p_j = 0.1$.

Sample size $n = 225$ with inclusion probability $p_j = 0.5$:

As expected, an increased sample size of $n = 225$ led to overall higher AFCCF values across both methods. Again, when $E[\beta] = 2$, CAR-SSVS is able to leverage the shared information to better infer the variable inclusion indicators. The discrepancy between methods is exaggerated with the number of clusters is large (Tables 5.12e and 5.12f).

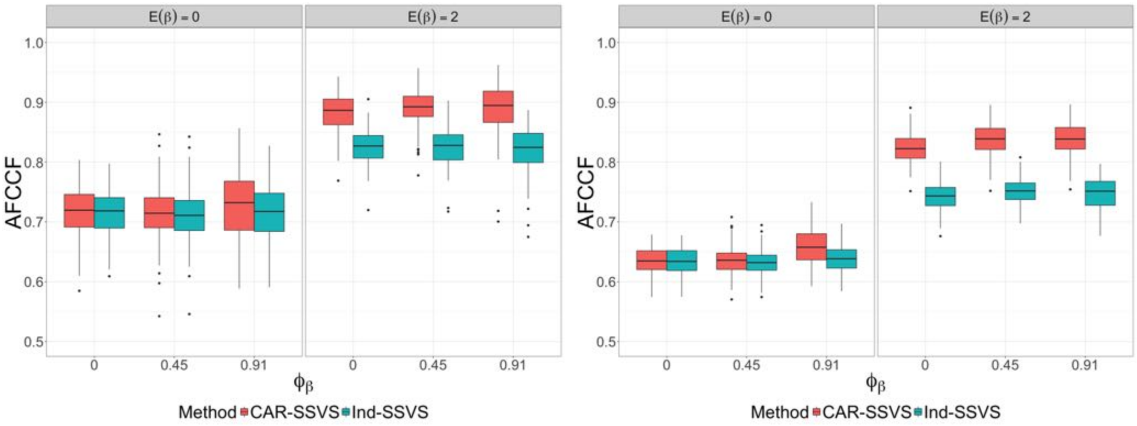


(a) $p = 5$ and $R = 2$

(b) $p = 15$ and $R = 2$

(c) $p = 5$ and $R = 4$

(d) $p = 15$ and $R = 4$



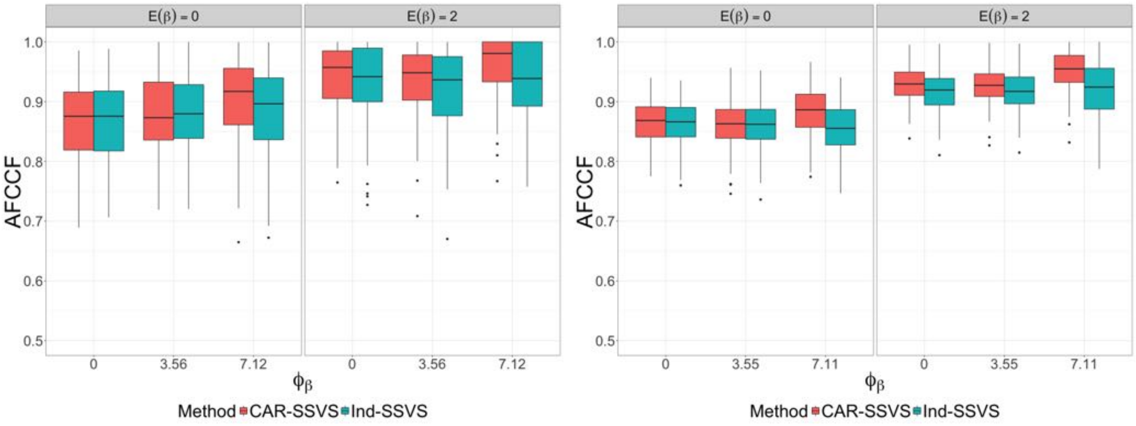
(e) $p = 5$ and $R = 8$

(f) $p = 15$ and $R = 8$

Figure 5.12: Boxplots for AFCCF for Independent-SSVS and CAR-SSVS with sample size $n = 225$ and inclusion probability $p_j = 0.5$.

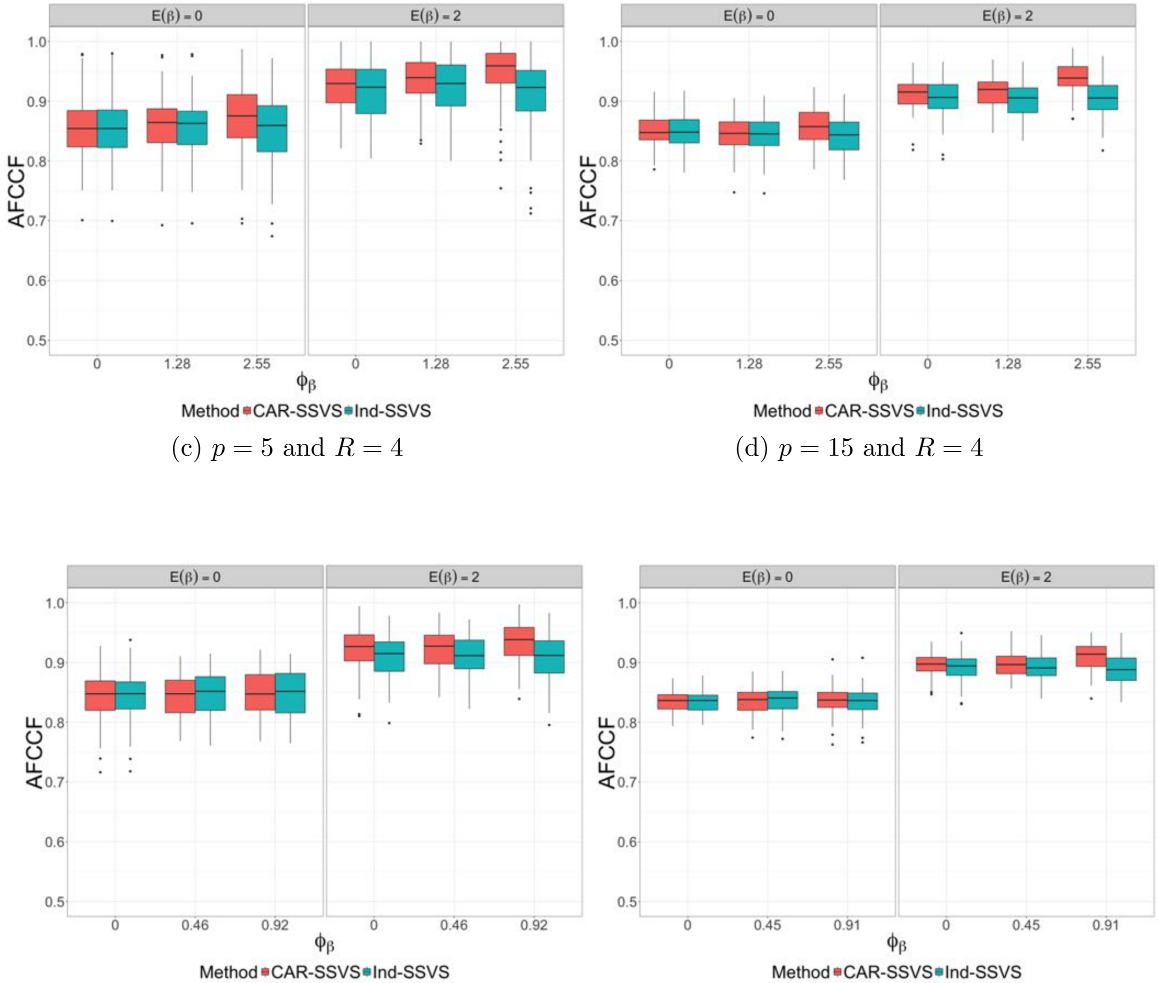
Sample size $n = 225$ with inclusion probability $p_j = 0.9$:

Both Independent-SSVS and CAR-SSVS obtain similar AFCCF values when $p_j = 0.9$. CAR-SSVS has a slight advantage when $E[\beta] = 2$ and ϕ_β is large. Otherwise, we see no real advantage in using the CAR model on the linear coefficients.



(a) $p = 5$ and $R = 2$

(b) $p = 15$ and $R = 2$



0

(c) $p = 5$ and $R = 4$

(d) $p = 15$ and $R = 4$

(e) $p = 5$ and $R = 8$

(f) $p = 15$ and $R = 8$

Figure 5.13: Boxplots for AFCCF for Independent-SSVS and CAR-SSVS with sample size $n = 225$ and inclusion probability $p_j = 0.9$.

Summary of results:

In many scenarios, Independent-SSVS and CAR-SSVS lead to very similar results. However, CAR-SSVS is advantageous in situations where there are large effects across clusters that can be shared. When the mean effect $E[\beta]$ is large, the implied correlation between clusters makes it easier to determine if an effect is 0 (details in Section 3.2.5). Further, the stronger the correlation is, measured by ϕ_β, the better CAR-SSVS performs. The information sharing is leveraged more with a larger number of clusters. When R is large, Independent-SSVS needs to estimate more parameters without additional observations, while CAR-SSVS uses intra-cluster estimation In the absence of these advantageous situations, CAR-SSVS does not perform significantly worse than Independent-SSVS.

5.3 Independent-SSVS versus SSVS with Ising

5.3.1 Simulating data

As with Section 5.2.3, we perform simulations to understand the effects of specifying an Ising distribution as a prior distribution for the variable inclusion indicators. We compare the Ising distribution with the common choice, $\gamma_j^{(r)} \sim Bernoulli(p = .5)$. Datasets are simulated under various settings of total sample size n, number of clusters R, number of possible covariates per cluster p, spatial strength in the Ising model θ, and the underlying mean level of the nonzero coefficients $E\left[\beta\right]$. We discuss the simulation procedure below.

1. Given n, generate a square grid of size $\sqrt{n} \times \sqrt{n}$.

2. Generate R cluster centroids and n cluster assignments using K-means algorithm.

3. Given the cluster centroids, calculate the Ising weight matrix $\boldsymbol{W}$, where $w_{r,r'} = 1/d_{r,r'}$ for $r \neq r'$, $w_{r,r} = 0$, and $d_{r,r'}$ is the Euclidean distance between the centroids of clusters r and r'.

4. We choose θ_{max} such that the maximum conditional probability of $\gamma_j^{(r)} = 1$ is .95. Then, we generate a uniform sequence of θ between 0 and θ_{max} to simulate a variety of spatial strengths in the Ising distribution.

5. Simulate indicators for each cluster and each covariate $(\gamma_j^{(1)}, \ldots \gamma_j^{(R)}) \sim Ising(\theta)$ for $j = 1, \ldots, p$.

6. Simulate the vector of intercepts $\boldsymbol{\mu} \sim Normal(E\left[\beta\right] \mathbf{1}, 1)$.

7. For each covariate j such that $\gamma_j^{(r)} = 1$, simulate $\{\beta_j^{(r)} | \gamma_j^{(r)} = 1\} \sim Normal(E\left[\beta\right], 1)$.

8. For each cluster, simulate a covariate matrix $\boldsymbol{X}^{(r)}$ such that each element $x_{ij} \sim Normal(0, 1)$.

9. For each cluster, simulate the response vector $\boldsymbol{y}^{(r)} \sim Normal(\mu^{(r)} + \boldsymbol{X}^{(r)}\boldsymbol{\beta}^{(r)}, \boldsymbol{I})$.

For each combination, we simulate 100 datasets. It is assumed that we know the true values for all parameters except for θ, each $\gamma_j^{(r)}$, and each $\beta_j^{(r)}$. The goal of our inference is to correctly estimate the variable inclusion indicators $\gamma_j^{(r)}$ for $r = 1, \ldots, R$ and $j = 1, \ldots, p$.

5.3.2 Methods of estimation and evaluation

For each simulated dataset, we use separate methods to infer the unknown parameters: SSVS with prior probability of variable inclusion $p_j^{(r)} = 0.5$ and SSVS with Ising distribution on the variable inclusion indicators. For both models, we assume the prior distribution $(\beta_j^{(r)} | \gamma_j^{(r)} = 1) \sim Normal(0, 1)$. The only difference between the two methods lies in the estimation of θ. Under the Ising model, we infer θ using the pseudo-likelihood with Exponential prior distribution (Section 4.4). For Independent-SSVS method, we fix $\theta = 0$, resulting in a priori independent $\gamma_j^{(r)}$ across all clusters and covariates.

To evaluate the performance of each technique, we use the average fraction correctly classified for fit (AFCCF) (Section 5.2.3) on the variable inclusion indicators $\gamma_j^{(r)}$. The formula for AFCCF is again, given by

$$AFCCF = \frac{1}{pR}\Big[\sum_{r=1}^{R}\sum_{j=1}^{p}\gamma_j^{(r)}P(\gamma_j^{(r)} = 1|\boldsymbol{y}) + (1 - \gamma_j^{(r)})P(\gamma_j^{(r)} = 0|\boldsymbol{y})\Big],$$

where $P(\gamma_j^{(r)} = 1|\boldsymbol{y})$ is the posterior marginal probability that $\gamma_j^{(r)} = 1$. 100 datasets were simulated for each combination of settings.

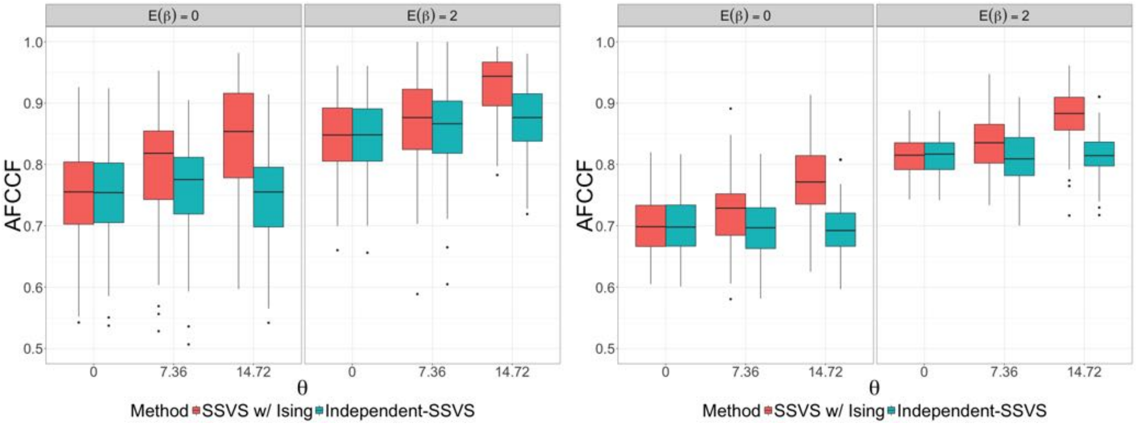
Table 5.3: Configurations for each simulation. θ_{max} is the value of θ, conditional on $\boldsymbol{W}$, that induces a maximum conditional probability of .95 among the variable inclusion indicators.

Parameter	Values
n	$\{100, 225\}$
p	$\{5, 15\}$
R	$\{2, 4, 8\}$
θ	$\{0, .5\theta_{max}, \theta_{max}\}$
$E[\beta]$	$\{0, 2\}$

5.3.3 Simulation results

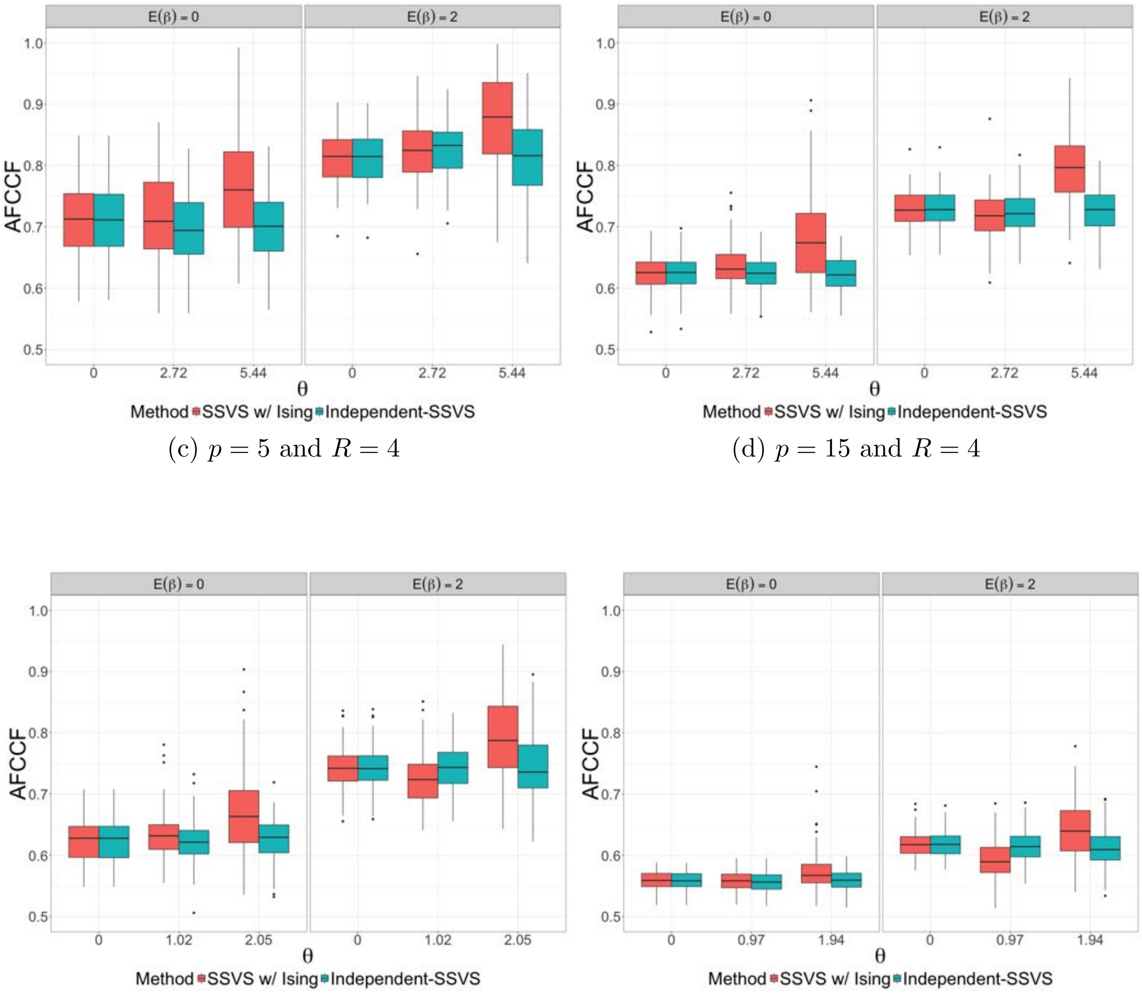
Unlike the CAR-SSVS versus Independent-SSVS comparison (Section 5.2.3), the driving force in discrepancies between methods is not the effect size $E[\beta_p]$. The greatest advantage from using the Ising distribution occurs as the true spatial dependency between clusters becomes larger. The Ising model has the flexibility to use information about neighboring clusters to bolster the prior belief in the current cluster. When $\theta = 0$, both methods result in almost identical results. Since θ is estimated, when the spatial dependency is not present, the posterior distribution of θ is concentrated near smaller values of θ, leading to similar results as Independent-SSVS.

Sample size of $\boldsymbol{n = 100}$:



(a) $p = 5$ and $R = 2$

(b) $p = 15$ and $R = 2$



(c) $p = 5$ and $R = 4$

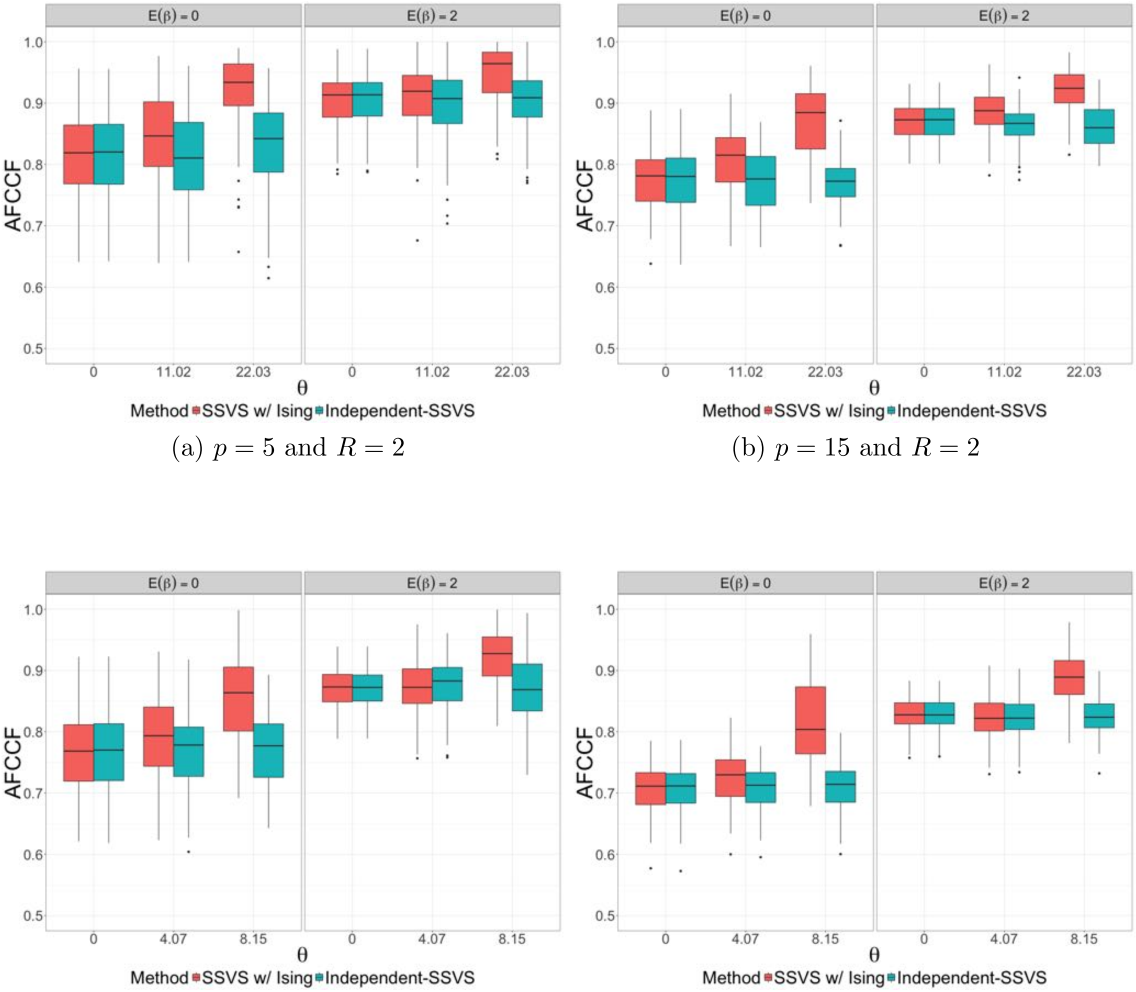
(d) $p = 15$ and $R = 4$

(e) $p = 5$ and $R = 8$

(f) $p = 15$ and $R = 8$

Figure 5.14: Boxplots for AFCCF for Independent-SSVS and SSVS with Ising with sample size $n = 100$.

Sample size of $n = 225$:

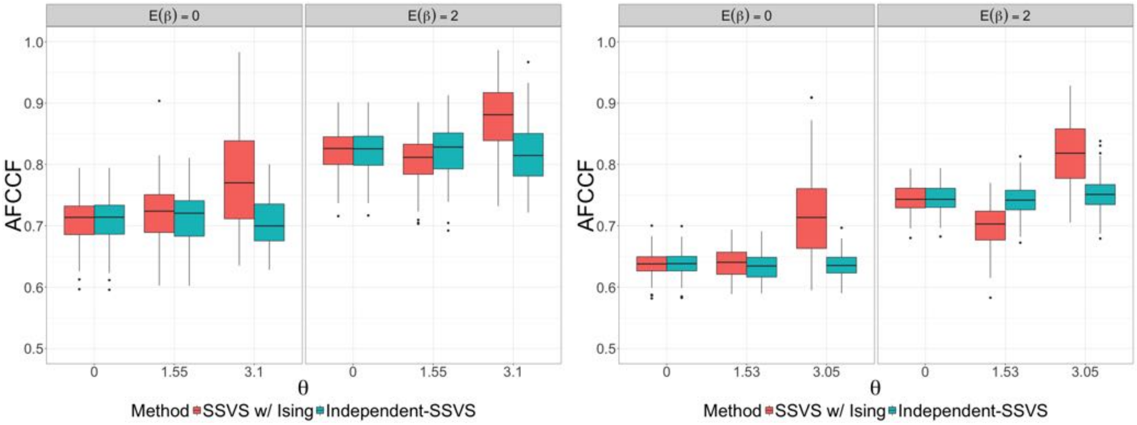


(a) $p = 5$ and $R = 2$

(b) $p = 15$ and $R = 2$

(c) $p = 5$ and $R = 4$

(d) $p = 15$ and $R = 4$



(e) $p = 5$ and $R = 8$

(f) $p = 15$ and $R = 8$

Figure 5.15: Boxplots for AFCCF for Independent-SSVS and SSVS with Ising with sample size $n = 225$.

Chapter 6

East Brook Trout Joint Venture Analysis

6.1 Overview of data

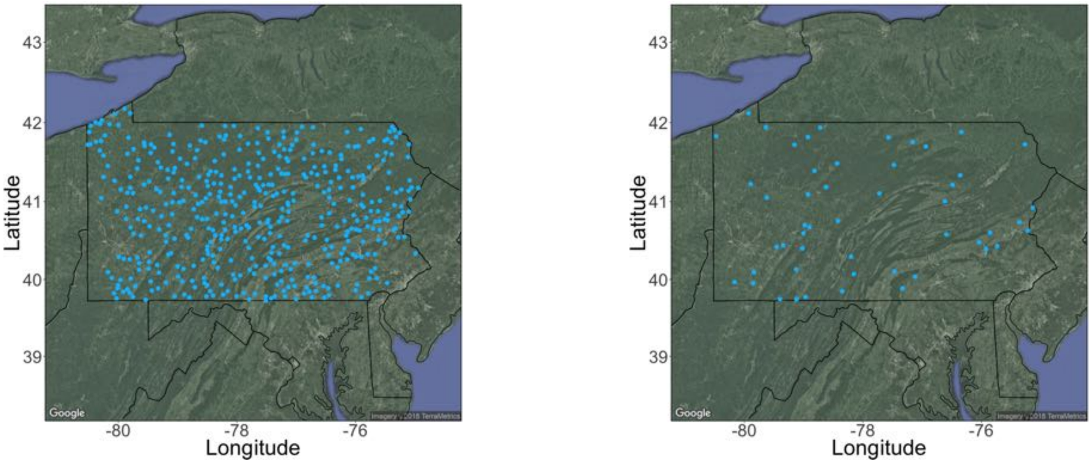
We now apply the HGP methodology to the East Brook Trout Joint Venture (EBTJV) dataset. The complete dataset contains $3,337$ sampled subwatershed locations throughout the eastern United States (Hudy et al. 2008). At each location, we have information on physical characteristics, such as landscape and anthropogenic variables, as well as the response of interest, the presence or absence of brook trout. Section 1 contains additional details regarding the data collection and previous attempts at statistical analysis. Ultimately, we compare the HGP methodology to the stationary Gaussian process model with model selection.

We choose to analyze the covariates chosen by Velasco-Cruz (2012). Velasco-Cruz considered the five covariates used in the analysis by Zhang et al. (2008), along with three additional linear covariates. With these eight linear covariates, we use longitude and latitude to control for spatial correlation when predicting the response, the presence or absence of Brook trout. The list of all covariates considered in our HGP methodology can be found in Table 6.1.

Table 6.1: All variables used in the the analysis. Longitude and latitude are only used for the Gaussian process covariance.

Variable	**Usage**	**Description**
Status	Response	The presence (Status= 1) or absence (Status= 0) of Brook trout
Longitude	Covariance	Longitude of subwatershed
Latitude	Covariance	Latitude of subwatershed
Road_Density	Linear	Road density of subwatershed
Indust_Trans	Linear	Proportion of commercial/industrial/transportation of subwatershed
Transitional	Linear	Proportion of transitional areas of sparse vegetation of subwatershed
Ag	Linear	Proportion of subwatershed used for agriculture
Mixed_Forest	Linear	Proportion of mixed forest in subwatershed
Elevation	Linear	Mean elevation of subwatershed
Total_Forest	Linear	Combined proportion of deciduous, evergreen, and mixed forest of subwatershed
Log_Chem	Linear	Logarithm of environmental information using NO3 and SO4

We focus our analysis of the EBTJV on a random sample of 500 subwatershed locations from Pennsylvania. Previous work by Zhang et al. (2008) and Velasco-Cruz (2012) have suggested that Pennsylvania benefits from using localized models with the EBTJV data. For each model we compare, we use 90% of the data for estimation and the remaining 10% for cross-validation. This is repeated ten times to ensure each observation is reserved for prediction exactly one time. Figure 6.1 shows the layout of both the training (left) and testing (right) sets for one instance of model fitting. A full map showing the distribution of presence/absence can be found in Figure 1.2 within Section 1.



(a) 450 subwatersheds for model training (b) 50 subwatersheds for cross-validation

Figure 6.1: Sample of 500 subwatersheds across Pennsylvania. 450 are used for fitting HGP, while the remaining 50 are used for cross-validation.

6.2 Accommodating the binary response

The HGP methodology (Section 3.1.1) addressed the mixture of Gaussian process models with a continuous response, $y_i^{(r)} \in (-\infty, \infty)$. The EBTJV dataset uses the presence or absence of Brook trout in subwatershed locations as the response of interest. Consequently, we must build an additional layer into our methodology to link the binary response with the Gaussian process mixtures. De Oliveira (2000) developed the clipped Gaussian process for analyzing spatially correlated binary data. This involves treating the continuous Gaussian process as an unobserved, latent variable. The latent variable is assumed to be truncated to either 0 or one. That is,

$$z_i^{(r)} = \begin{cases} 1, & y_i^{(r)} \geq 0, \\ 0, & y_i^{(r)} < 0, \end{cases}$$

$$\boldsymbol{y}^{(r)} \sim Normal(\mu^{(r)}\mathbf{1} + \boldsymbol{X}^{(r)}\boldsymbol{\beta}^{(r)}, \boldsymbol{\Sigma}^{(r)}),$$

where $z_i^{(r)}$ is the binary status (presence/absence) of Brook trout at location i, which is located in cluster r. The latent variables $\boldsymbol{y}^{(r)}$ are unknown and are integrated over in the posterior sampling. Further, another consequence of using the clipped Gaussian process is that not all parameters are identifiable. De Oliveira (2000) recommends fixing parameters in the Gaussian process covariance. We alter the Gaussian process covariance function

$$\boldsymbol{\Sigma}^{(r)} = (\tau^{(r)})^2\big(\boldsymbol{\rho}(\boldsymbol{D}|\phi^{(r)}) + g^{(r)}\boldsymbol{I}\big),$$

by setting $\tau^{(r)} = 1$ and $g^{(r)} = 0$, resulting in $\boldsymbol{\Sigma}^{(r)} = \boldsymbol{\rho}(\boldsymbol{D}|\phi^{(r)})$. Conditional upon the current values for $\boldsymbol{y}^{(r)}$ at any given MCMC iteration, the model is Gaussian. Details about sampling $\boldsymbol{y}^{(r)}$ can be found in Albert & Chib (1993).

Another advantage to the clipped Gaussian process is the procedure for prediction at new spatial locations. Since the latent observations follow a Gaussian process, we can predict using properties of multivariate Normal random variables. For ease of notation, assume only one cluster exists, thus the model for $\boldsymbol{y}^{(r)}$ becomes $\boldsymbol{y} \sim Normal(\mu\boldsymbol{1} + \boldsymbol{X\beta}, \Sigma)$, where $\Sigma = \boldsymbol{\rho(D|\phi)}$ is a matrix with correlation measurements between all locations $\boldsymbol{s}_\ell$ and $\boldsymbol{s}_{\ell'}$. Then, to make a prediction at a new location $\boldsymbol{s}_0$ with covariates $\boldsymbol{x}_0$, it is

$$\begin{aligned} P(z_0 = 1|z_1, \ldots, z_n) &= P(y_0 > 0|\boldsymbol{y}) \\ &\approx \frac{1}{B}\sum_{b=1}^{B} \delta(y_0^{(b)} > 0|\boldsymbol{y}), \end{aligned}$$

where $y_0^{(b)}$ for $b = 1, \ldots, B$ are MCMC draws after burn-in from the distribution $P(y_0|\mu, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{y})$. This distribution is given by

$$P(y_0|\mu, \boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{y}) = Normal(\mu_0, \sigma_0^2),$$

with $\sigma_0^2 = 1 - \boldsymbol{\rho}_0\boldsymbol{\rho}^{-1}\boldsymbol{\rho}_0'$, $\mu_0 = \mu + \boldsymbol{x}_0'\boldsymbol{\beta} + \boldsymbol{\rho}_0\boldsymbol{\rho}^{-1}(\boldsymbol{y} - \mu\boldsymbol{1} - \boldsymbol{X\beta})$, and $\boldsymbol{\rho}_0$ is a $1 \times n$ correlation matrix between $\boldsymbol{s}_0$ and $\boldsymbol{s}_\ell$ for $\ell = 1, \ldots, n$. To evaluate the performance of the HGP methodology under different model assumptions for number of clusters, we use the the predictive probability of Brook trout presence for our testing test.

To evaluate the overall predictive performance, we utilize the average fraction correctly classified for fit (AFCCF) statistic. Unlike scenarios where the data are simulated from a true model, we cannot calculate AFCCF on the variable inclusion indicators, since they are unknown. Rather, we use AFCCF with the posterior probability that each testing location has Brook trout present or absent. If we denote the Status (presence/absence) at each testing location as $z_0^{(1)}, \ldots, z_0^{(n_0)}$, AFCCF is easily computed as

$$AFCCF = \frac{1}{n_0}\bigg(\sum_{\ell=1}^{n_0} z_0^{(\ell)} P(z_0^{(\ell)} = 1|\boldsymbol{z}) + (1 - z_0^{(\ell)}) P(z_0^{(\ell)} = 0|\boldsymbol{z})\bigg).$$

6.3 Results for Pennsylvania EBTJV data

We perform the analysis for the Pennsylvania EBTJV data using the model stated in Section 6.2. All prior distributions for $\boldsymbol{\mu}$, $\boldsymbol{\beta}$, ϕ_β, and σ_β^2 follow from the HGP methodology in Chapter 3. We place an Exponential prior distribution on θ (Section 4.4). For this special binary case, we use the reference prior from Berger et al. (2001) for the lengthscale parameters $\phi^{(r)}$. In all scenarios, we obtain $100,000$ samples from the posterior distribution after burn-in is removed.

6.3.1 Choosing the number of clusters

First, we decide the appropriate number of spatial clusters for Pennsylvania. We use 10-fold cross validation on the same 500 locations for each $R \in \{1, 2, 3, 4\}$ to compare the AFCCF values (Figure 6.2). On average, $R = 3$ clusters results in an AFCCF value that is 7% higher than the stationary model ($R = 1$) and 3.5% higher than when $R = 2$ clusters. Even the worst predictive set for $R = 3$ outperforms the median AFCCF for the stationary model.

Increasing the number of clusters to $R = 4$ provides no improvement over $R = 3$, therefore we proceed with the parameter inference with the $R = 3$ cluster model.

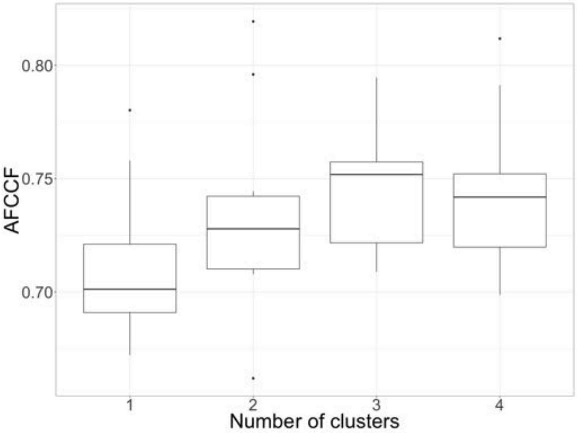
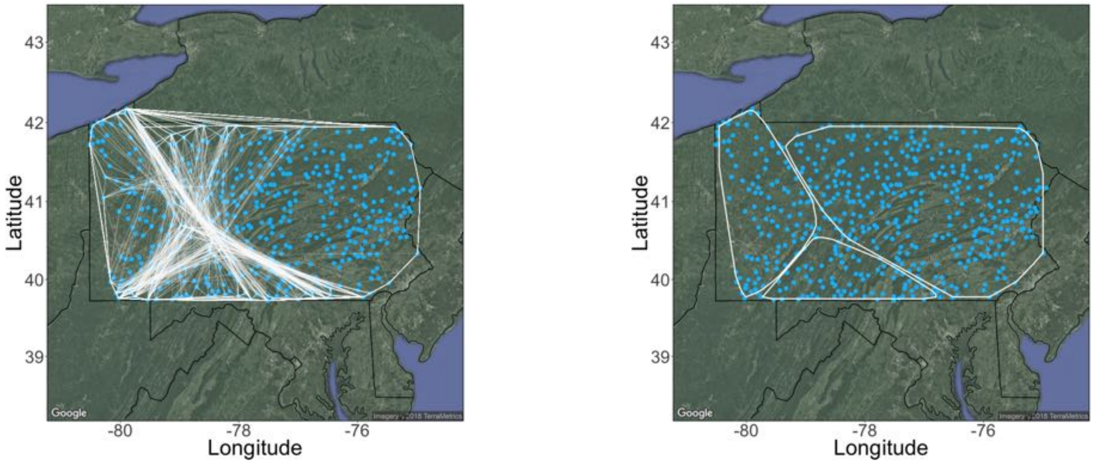


Figure 6.2: Boxplots containing the AFCCF values across 10 cross-validated sets for $R \in \{1, 2, 3, 4\}$. We select $R = 3$ clusters to perform inference.

As discussed in Section 4.2.2, the convex hull constrained cluster assignments need to be sampled throughout the MCMC. The data supports a clustering scheme that groups western Pennsylvania into one cluster, central and eastern Pennsylvania into an other cluster, and a small section of southern Pennsylvania into the third cluster. Figure 6.3a shows 250 realizations from the posterior distribution of cluster assignments (convex hulls). Areas with a large amount of overlapping white imply more certainty in the existence of the grouping. The thinner white lines are cluster assignments that are visited less frequently in the posterior distribution. The posterior "average" for the cluster assignments is the set of convex hulls resulting from putting each location in its most probable cluster (Figure 6.3b). The covariance of the attributes within each cluster can be found in Tables 6.2, 6.3, and 6.4.



(a) 250 draws from posterior distribution of cluster assignments

(b) Posterior “average” convex hull assignments

Figure 6.3: Draws from the posterior distribution for cluster assignments.

Table 6.2: Covariance of attributes associated with the most probable southern cluster.

	Road Density	Indust Trans	Transitional	Ag	Mixed Forest	Elevation	Total Forest	Log Chem
Road_Density	0.31	0.25	0.01	0.18	-0.06	-0.11	-0.24	-0.10
Indust_Trans	0.25	0.35	0.11	0.08	-0.03	-0.12	-0.14	0.03
Transitional	0.01	0.11	1.86	-0.46	0.12	-0.01	0.39	0.13
Ag	0.18	0.08	-0.46	1.30	-0.41	-0.83	-1.22	-0.59
Mixed_Forest	-0.06	-0.03	0.12	-0.41	0.37	0.31	0.39	0.12
Elevation	-0.11	-0.12	-0.01	-0.83	0.31	1.44	0.79	0.66
Total_Forest	-0.24	-0.14	0.39	-1.22	0.39	0.79	1.18	0.54
Log_Chem	-0.10	0.03	0.13	-0.59	0.12	0.66	0.54	1.08

Table 6.3: Covariance of attributes associated with the most probable western cluster.

	Road Density	Indust Trans	Transitional	Ag	Mixed Forest	Elevation	Total Forest	Log Chem
Road_Density	0.77	0.62	0.02	0.07	-0.22	-0.24	-0.31	-0.11
Indust_Trans	0.62	0.75	0.01	-0.02	-0.03	-0.17	-0.22	-0.14
Transitional	0.02	0.01	0.37	-0.16	0.21	0.03	0.13	-0.01
Ag	0.07	-0.02	-0.16	0.43	-0.28	-0.13	-0.39	-0.14
Mixed_Forest	-0.22	-0.03	0.21	-0.28	0.95	0.20	0.32	0.04
Elevation	-0.24	-0.17	0.03	-0.13	0.20	0.25	0.22	0.20
Total_Forest	-0.31	-0.22	0.13	-0.39	0.32	0.22	0.47	0.22
Log_Chem	-0.11	-0.14	-0.01	-0.14	0.04	0.20	0.22	0.46

Table 6.4: Covariance of attributes associated with the most probable eastern cluster.

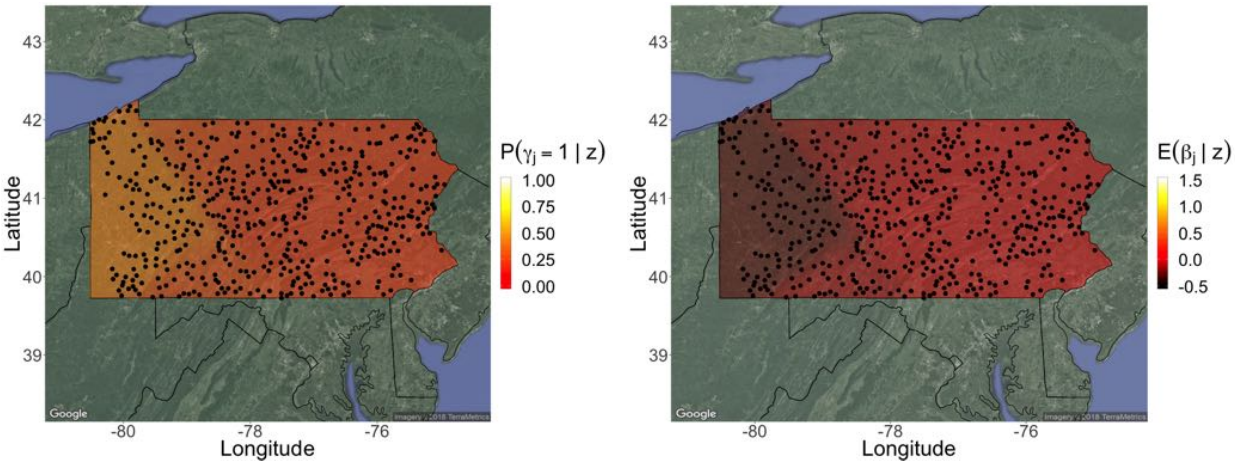
	Road Density	Indust Trans	Transitional	Ag	Mixed Forest	Elevation	Total Forest	Log Chem
Road_Density	1.22	0.90	-0.05	0.50	-0.50	-0.80	-0.78	-0.46
Indust_Trans	0.90	1.23	0.17	0.14	-0.33	-0.45	-0.49	-0.24
Transitional	-0.05	0.17	0.99	-0.29	0.19	0.18	0.20	0.12
Ag	0.50	0.14	-0.29	1.08	-0.58	-0.79	-1.02	-0.29
Mixed_Forest	-0.50	-0.33	0.19	-0.58	1.13	0.63	0.64	0.31
Elevation	-0.80	-0.45	0.18	-0.79	0.63	1.14	0.89	0.42
Total_Forest	-0.78	-0.49	0.20	-1.02	0.64	0.89	1.11	0.37
Log_Chem	-0.46	-0.24	0.12	-0.29	0.31	0.42	0.37	1.16

6.3.2 Posterior probabilities and mean effects

Using the posterior “average” convex hulls, we can analyze the posterior probabilities for variable inclusion and linear effects for each cluster. However, since cluster assignments change through the posterior sampling (Figure 6.3a), we begin by analyzing the smoothed linear effects. For any given MCMC iteration, we know the current state of the cluster assignments $\boldsymbol{c}$, as well as $\gamma_j^{(r)}$ and $\beta_j^{(r)}$ for $r = 1, \ldots 3$ and $j = 1, \ldots, 8$. By matching each spatial location to the corresponding ($\gamma_j^{(r)}$,$\beta_j^{(r)}$) pair, we create a heat map of both the variable inclusion probability and linear effect through space. This procedure marginalizes the cluster assignments. In Figures 6.4-6.9, we show the posterior mean for $\gamma_j^{(r)}$ and $\beta_j^{(r)}$ as a function of the spatial location. The smoothing effect is achieved by interpolating the effects at each observed location.

Covariates with low posterior probability of inclusion:

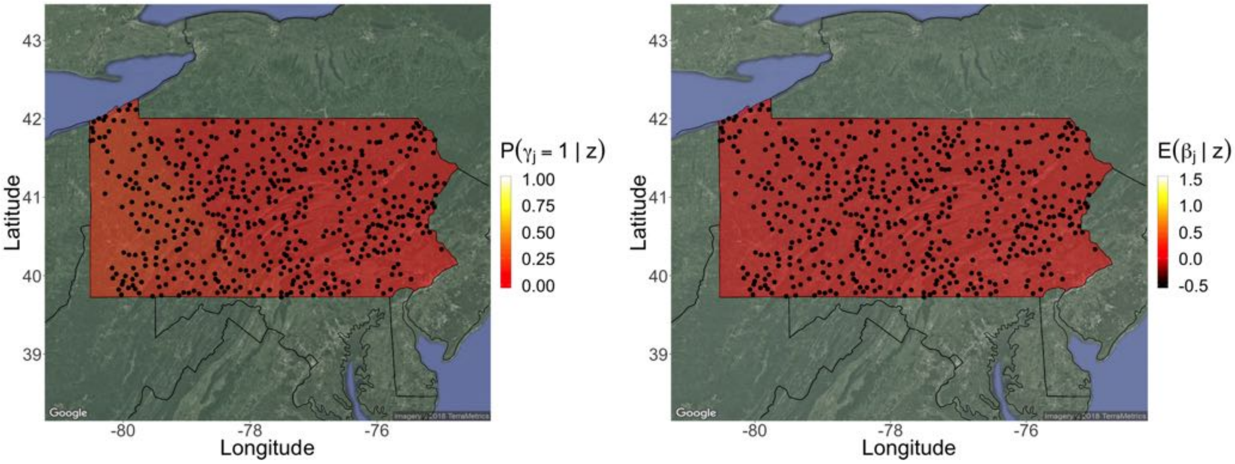
The covariates Road_Density, Indust_Trans, Transitional, Ag, Mixed_Forest, and Log_Chem collectively have the lowest posterior probabilities for variable inclusion. These probabilities range from 0.09 to 0.369 (Figures 6.4a-6.9a). Log_Chem is the least important, with posterior probability of inclusion ranging from 0.047 to 0.152 throughout Pennsylvania. The other covariates have posterior probabilities of inclusion near 0.30. Consequently, we observe fairly consistent, near-zero posterior mean effects for Indust_Trans (Figure 6.5b), Transitional (Figure 6.6b), Mixed_Forest (Figure 6.8b), and Log_Chem (Figure 6.9b. The linear effects do vary, however, for Road_Density and Ag. In western Pennsylvania, we observe a negative posterior mean effect that does not exist in other parts of the state (Figure 6.4b). Therefore, we would predict a lower probability of brook trout presence in western Pennsylvania as road density increases. We see similar behavior for Ag in the western part of the state (Figure 6.7b), however the effect is not quite as strong.



(a) Posterior mean for variable inclusion

(b) Posterior mean linear effect

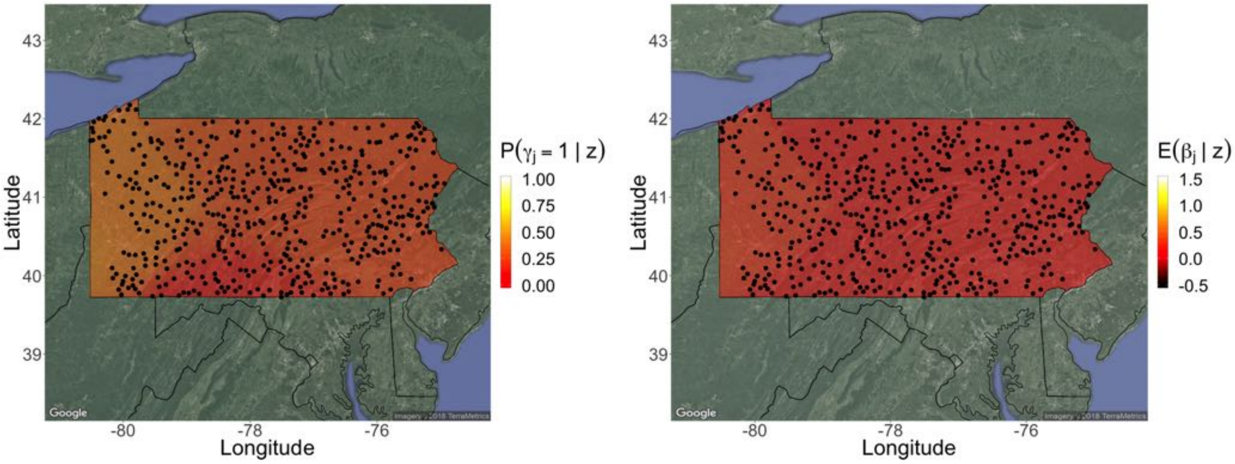
Figure 6.4: Posterior mean across space of γ_j and β_j for Road_Density.



(a) Posterior mean for variable inclusion

(b) Posterior mean linear effect

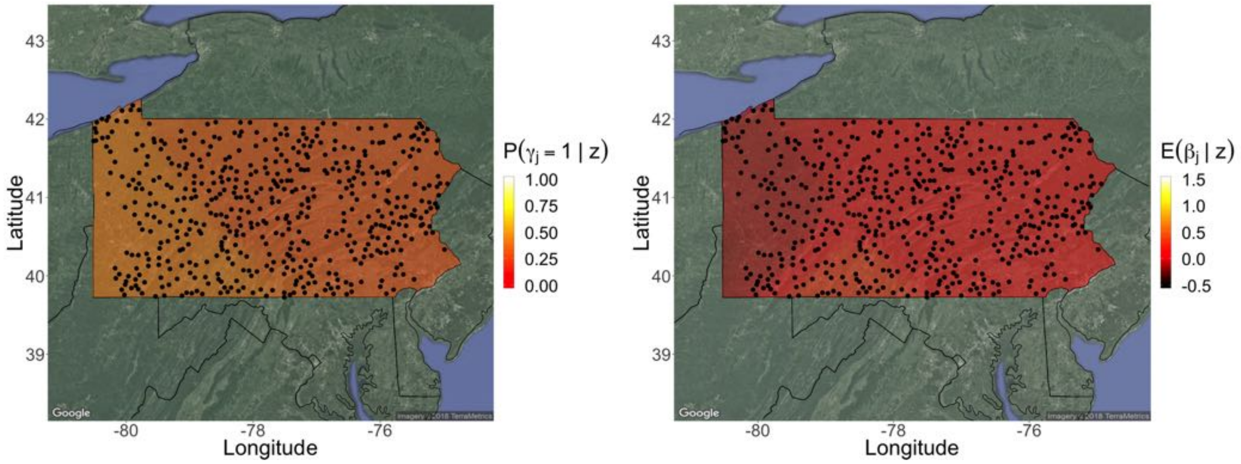
Figure 6.5: Posterior mean across space of γ_j and β_j for Indust_Trans.



(a) Posterior mean for variable inclusion

(b) Posterior mean linear effect

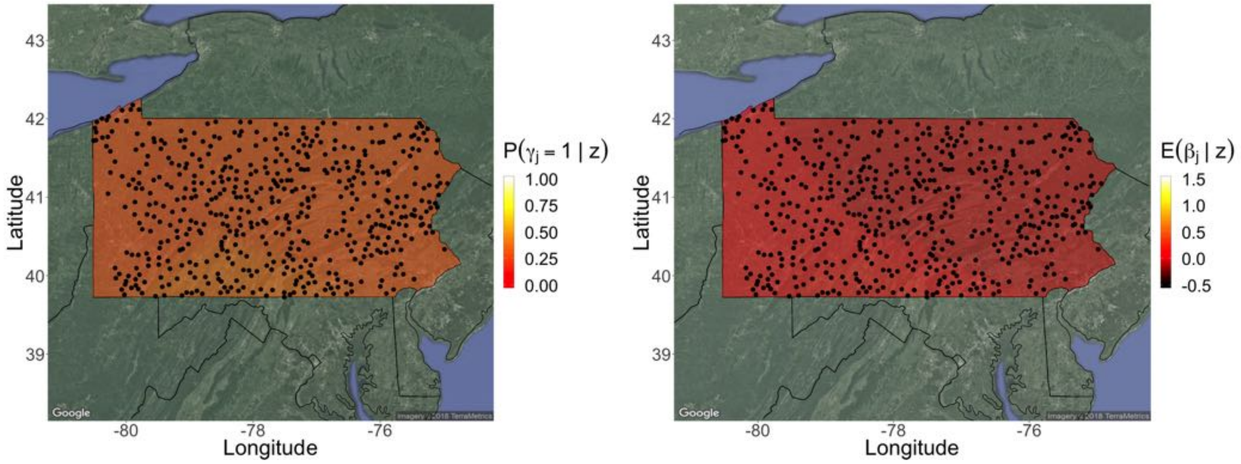
Figure 6.6: Posterior mean across space of γ_j and β_j for Transitional.



(a) Posterior mean for variable inclusion

(b) Posterior mean linear effect

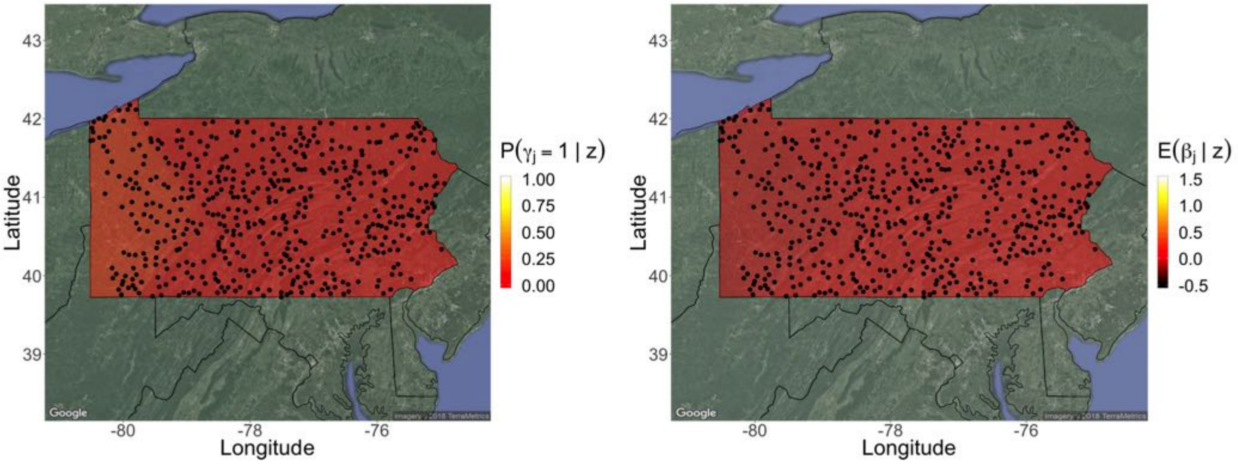
Figure 6.7: Posterior mean across space of γ_j and β_j for Ag.



(a) Posterior mean for variable inclusion

(b) Posterior mean linear effect

Figure 6.8: Posterior mean across space of γ_j and β_j for Mixed_Forest.



(a) Posterior mean for variable inclusion

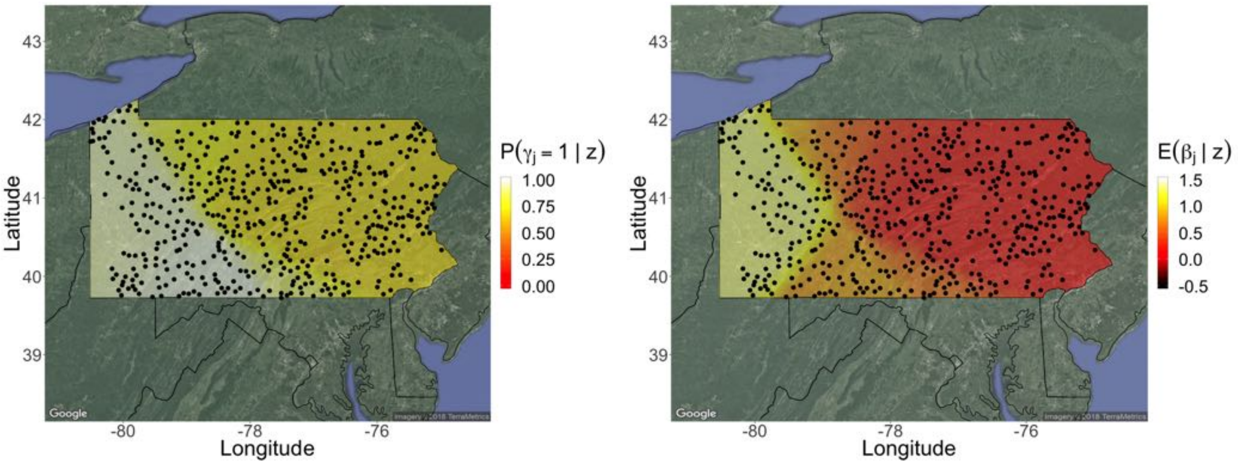
(b) Posterior mean linear effect

Figure 6.9: Posterior mean across space of γ_j and β_j for Log_Chem.

Covariates with high posterior probability of inclusion:

The two covariates that have yet to be addressed, Elevation and Total_Forest, are both highly important to the prediction of brook trout presence in Pennsylvania. In eastern Pennsylvania, the posterior probability of inclusion for Elevation is at its lowest (0.696) (Figure 6.10a) and grows to 0.993 in the south-central part of the state. There is almost 100% certainty that higher elevations predict higher brook trout presence probabilities in western and south-central Pennsylvania with posterior mean effects between 0.859 and 1.132 (Figure 6.10b). This effect is almost 0 in the east. For the covariate Total_Forest, the posterior probability of inclusion ranges from 0.649 in south-central Pennsylvania to 0.911 and 0.990 in western and eastern Pennsylvania, respectively (Figure 6.11a). The low inclusion probability in the south-central region manifests itself in a near-zero posterior mean effect (Figure 6.11b). Throughout the rest of the state, the Total_Forest effect is moderately positive, where higher Total_Forest would predict higher probability of brook trout presence.

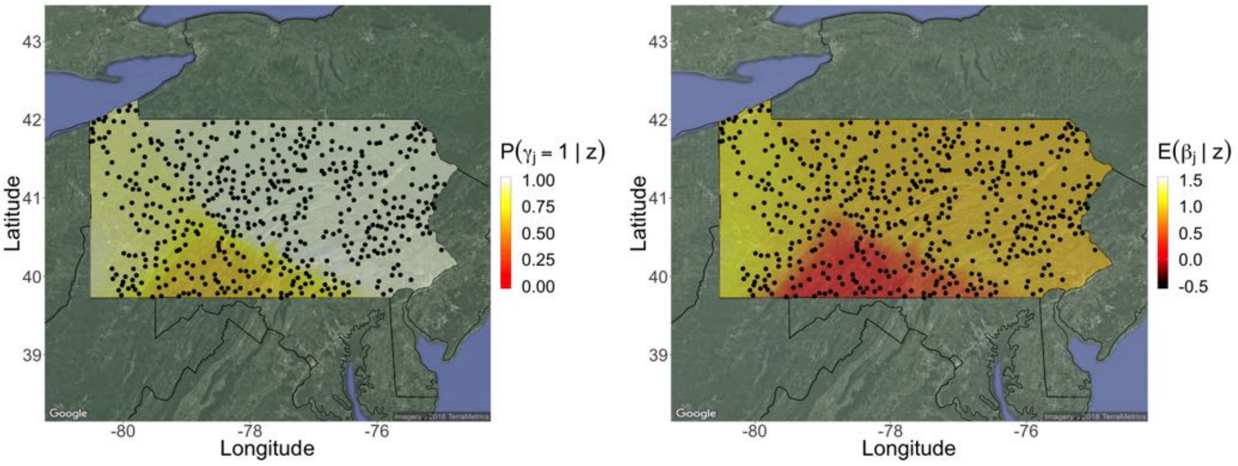
The shifting mean effects for Elevation and Total_Forest coincide with several geographical markers in Pennsylvania. For both Elevation and Total_Forest, we observe different behavior in the triangular region with vertices at $(39.72°, -79.49°)$, $(40.66°, -78.76°)$, (39.72°, −76, 89°). This region covers the Allegheny mountains in southern Pennsylvania. The behavior of these covariates changes depending on if the location is west of the mountains, in/near the mountains, or east of the mountains. Specifically, for the variable Elevation, we also observe a different effect in the north-central part of the state. This region is home to the Allegheny National Forest and the Moshannon, Sproul, Susquehannock, and Tioga State Forests.



(a) Posterior mean for variable inclusion

(b) Posterior mean linear effect

Figure 6.10: Posterior mean across space of γ_j and β_j for Elevation.



(a) Posterior mean for variable inclusion

(b) Posterior mean linear effect

Figure 6.11: Posterior mean across space of γ_j and β_j for Total_Forest.

While Figures 6.4-6.11 smooth over the cluster assignments, we can also consider the posterior distribution within each cluster. Consider the posterior "average" cluster assignments (assigning observations to most probable cluster) and corresponding convex hulls (Figure 6.12). Arbitrarily, we label these clusters as: southern Pennsylvania = cluster 1, western Pennsylvania = cluster 2, and central/eastern Pennsylvania = cluster 3. While these are dynamic assignments, this provides a context to the driving factor in the posterior heat maps (Figures 6.4-6.11). The posterior mean for variable inclusion indicators can be found in Table 6.5. In clusters 1 and 2, we observe high posterior probabilities for Elevation (0.993 and 0.962), while the probability is lower in cluster 3 (0.696). Similarly, the posterior probability of inclusion for Total_Forest is 0.911 and 0.990 in clusters 2 and 3, respectively. But the probability decreases greatly in cluster 1 (0.649). The posterior mean effect and 90% credible interval for each covariate and cluster, as well as intercepts and Gaussian process lengthscales, can be found in Table 6.6. The large effects for Elevation (in clusters 1 and 2) and Total_Forest (in clusters 2 and 3) are in bold.

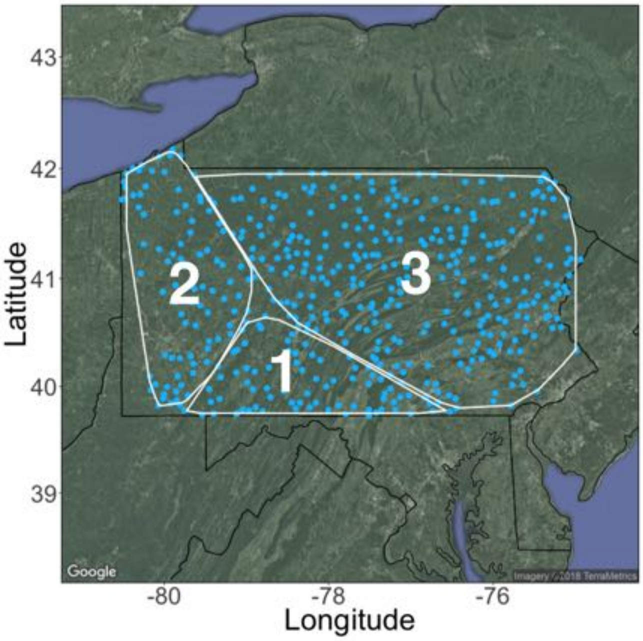


Figure 6.12: Posterior "average" cluster assignments with labeled clusters.

Table 6.5: Posterior mean for the variable inclusion indicators, $P(\gamma_j^{(r)} = 1|\boldsymbol{z})$. The variables Elevation and Total_Forest have the highest posterior probabilities across all clusters.

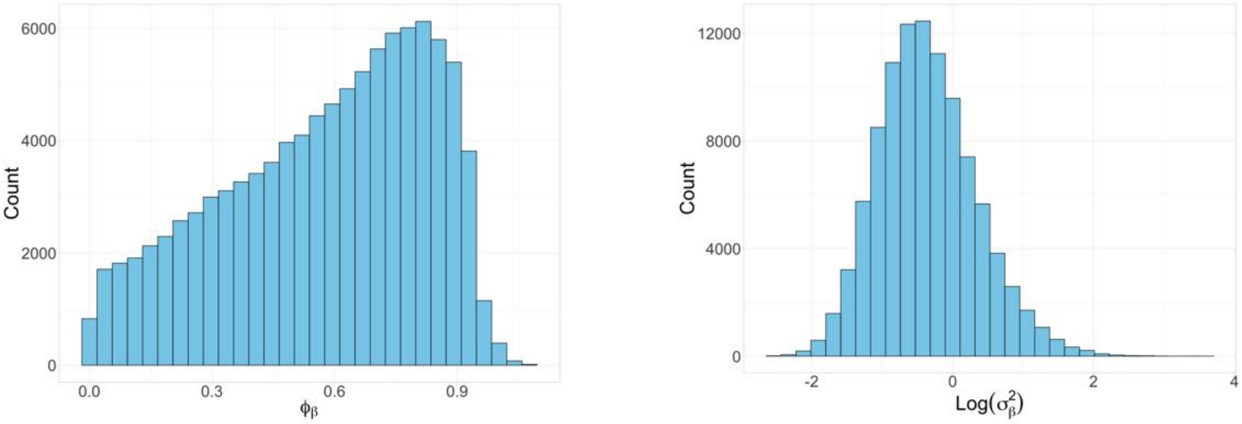
Variable	**Cluster 1**	**Cluster 2**	**Cluster 3**
Road_Density	0.302	0.369	0.179
Indust_Trans	0.107	0.344	0.185
Transitional	0.140	0.371	0.318
Ag	0.338	0.350	0.249
Mixed_Forest	0.341	0.265	0.271
Elevation	**0.993**	**0.962**	0.696
Total_Forest	0.649	**0.911**	**0.990**
Log_Chem	0.090	0.152	0.047

Table 6.6: Posterior mean and 90% credible intervals for the linear effects, $\beta_j^{(r)}$, the intercepts, $\mu^{(r)}$, and the Gaussian process lengthscales, $\phi^{(r)}$. Due to the skewness, some posterior means lie outside of the credible interval.

Variable	**Cluster 1**	**Cluster 2**	**Cluster 3**
Road_Density	-0.205 (-1.371,0.000)	-0.288 (-1.421,0.000)	-0.031 (-0.265,0.000)
Indust_Trans	0.073 (-0.004,0.777)	-0.012 (-0.388,0.287)	0.011 (0.000,0.080)
Transitional	0.001 (0.000,0.031)	0.264 (0.000,1.383)	0.035 (0.000,0.257)
Ag	0.210 (0.000,1.227)	-0.153 (-1.214,0.352)	0.032 (-0.161,0.404)
Mixed_Forest	0.181 (0.000,0.978)	0.065 (-0.070,0.525)	-0.056 (-0.344,0.000)
Elevation	**0.859** **(0.393,1.535)**	**1.132** **(0.000,2.772)**	0.194 (0.000,1.334)
Total_Forest	-0.013 (-0.783,0.916)	**1.066** **(0.000,2.208)**	**0.892** **(0.524,1.334)**
Log_Chem	0.003 (0.000,0.000)	-0.001 (-0.352,0.027)	0.002 (0.000,0.000)
Intercept	-0.345 (-1.081,0.175)	-0.933 (-1.515,-0.335)	0.730 (0.521,0.948)
ϕ (Gaussian process)	0.008 (0.001,0.027)	0.012 (0.001,0.041)	0.008 (0.002,0.014)

6.3.3 CAR and Ising parameters

In Section 6.3.2, we thoroughly discussed the inference associated with the parameters found within each cluster. Now, we discuss the parameters associated with the intra-cluster prior distributions: θ, the strength parameter of the Ising distribution (Section 3.2.2) and ϕ_β and σ^2_β, the spatial strength and scale parameters of the CAR model (Section 3.2.5). Higher values for both θ and ϕ_β suggest the presence of strong dependency between clusters. First, we consider the CAR model parameters. Histogram of $100,000$ draws from the marginal posterior distributions of ϕ_β and σ^2_β can be found in Figures 6.13a and 6.13b, respectively. When $\phi_\beta = 0$, the CAR model reduces to a diagonal matrix. With a posterior mode near $\phi_\beta = 0.8$, at least some spatial correlation exists between the nonzero $\beta^{(r)}_j$. To understand the scale of the dependency, we keep to also consider σ^2_β and the weight matrix $\boldsymbol{W}$. Together, the CAR covariance matrix is $\boldsymbol{K} = \sigma^2_\beta(\boldsymbol{I} - \phi_\beta \boldsymbol{W})^{-1}$.



(a) $P(\phi_\beta|\boldsymbol{z})$ (b) $P(Log(\sigma^2_\beta)|\boldsymbol{z})$

Figure 6.13: Draws from the marginal posterior distributions of ϕ_β and σ^2_β.

Using the draws from the joint posterior distribution of ϕ_β and σ^2_β with the associated weight matrix $\boldsymbol{W}$, we compute the posterior distribution of the matrix $\boldsymbol{K}$. Using $\boldsymbol{K}$, we can calculate the correlation between clusters r and r' as $\rho_{r,r'} = K_{r,r'}/\sqrt{K_{r,r}K_{r',r'}}$. The posterior distribution $P(\rho_{r,r'}|\boldsymbol{z})$ for all pairs of clusters can be found in the boxplots of Figure 6.14. Posterior means and 90% credible intervals are found in Table 6.7. The distribution for $\rho_{r,r'}$ is similar for all pairs of clusters, with a posterior mean of $\rho_{r,r'} \approx 0.5$ with a large amount of variability. About 50% of the posterior draws for $\rho_{r,r'}$ lie between 0.25 and 0.75. We conclude that spatial correlation between nonzero linear effects is moderate and consistent between clusters.

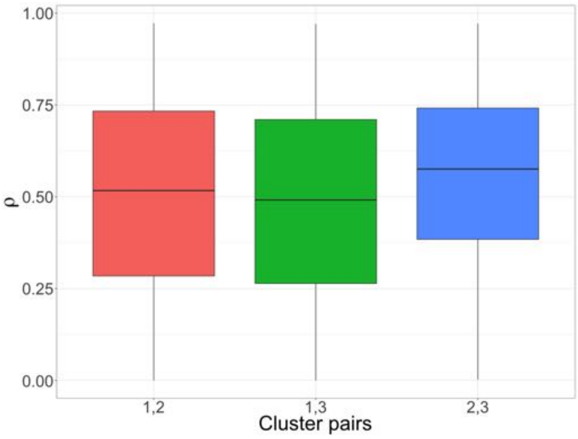


Figure 6.14: Posterior distribution of the pairwise correlation between clusters, $P(\rho_{r,r'}|\boldsymbol{z})$.

Table 6.7: Posterior mean and 90% credible intervals for the correlation between each pair of clusters.

Variable	**Mean**	**Lower Bound**	**Upper Bound**
$\rho_{1,2}$	0.509	0.070	0.932
$\rho_{1,3}$	0.490	0.062	0.922
$\rho_{2,3}$	0.557	0.162	0.890

The second method with which we embedded spatial correlation between clusters was the Ising distribution. For a given covariate, we jointly specified the prior distribution on $(\gamma_j^{(1)}, \ldots, \gamma_j^{(R)})' \sim Ising(\theta)$ (details in Section 3.2.2). Larger values for θ imply a stronger spatial dependency. The posterior distribution of θ is found in Figure 6.15. The posterior mode is near 0.25 and θ takes on values between 0.0001 and 6.788. The quantity of posterior mass concentrated on larger values suggests spatial dependence between the variable inclusion indicators.

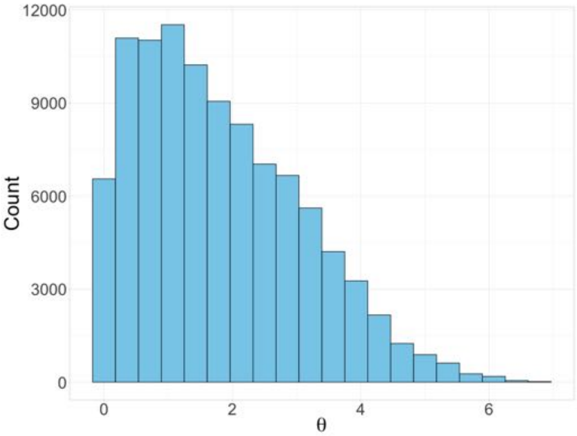


Figure 6.15: Posterior distribution of θ.

Much like the covariance under the CAR model, the effect that θ has on the posterior probability of the variable inclusion indicators depends on the weight matrix $\boldsymbol{W}$. To demonstrate the effect, we will consider the prior marginal probability $P(\gamma_j^{(r)} = 1|\boldsymbol{\gamma}_j^{-(r)} = 1)$, the maximum prior probability placed on $\gamma_j^{(r)} = 1$ under the condition that all other clusters include covariate j. A large value of θ will imply a higher prior probability of inclusion. Recall, the prior probability of variable inclusion is at most

$$P(\gamma_j^{(r)} = 1|\boldsymbol{\gamma}_j^{-(r)} = 1, \theta) = \left(1 + exp\{-\theta \sum_{r\sim r'} w_{r,r'}\}\right)^{-1}.$$

For further details, see Sections 3.2.2 and 5.1. In Figure 6.16, we show the distribution of $P(\gamma_j^{(r)} = 1|\boldsymbol{\gamma}_j^{-(r)} = 1)$ after marginalizing over the posterior draws of θ for clusters 1, 2, and 3. Since $\theta \geq 0$, the minimum for each cluster is 0.5. On average, the maximum prior probability of variable inclusion is between 0.7 and 0.95. Therefore, we conclude that the intra-cluster dependence from the variable inclusion indicators is present to a moderate or high degree.

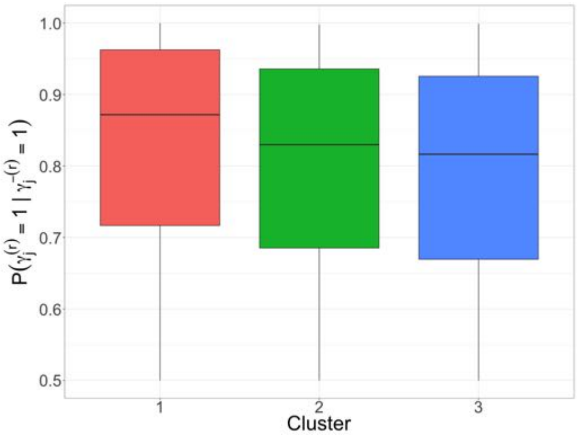
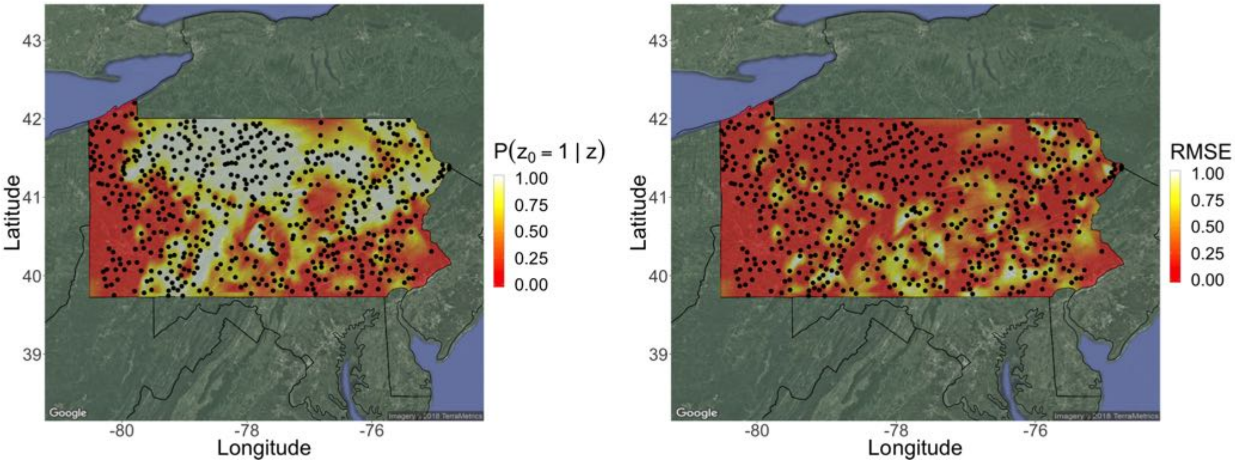


Figure 6.16: Distribution of $P(\gamma_j^{(r)} = 1|\boldsymbol{\gamma}_j^{-(r)})$ under posterior draws of θ.

6.3.4 Prediction across Pennsylvania

As discussed in Section 6.1, we used 500 randomly selected locations within Pennsylvania for the fitting of the HGP methodology. Using the draws from the posterior distribution, we are able to predict at the other 585 Pennsylvania locations. The geographical areas with the highest posterior probability of Brook trout presence lie in the Allegheny mountains, the northern forests, and the northeastern part of Pennsylvania (Figure 6.17a). Posterior probabilities are consistently low across western and the southeastern corner of the state.

The overall AFCCF for the remaining 585 locations is 0.779. The model is able to correctly classify true presence slighly better than true absence (presence $AFCCF = 0.798$, absence $AFCCF = 0.751$). The areas with the highest predictive RMSE lie in the southern half of Pennsylvania, largely aligning with cluster 1 (Figure 6.17b). The model also does not perform as well in some smaller pockets in the northeastern part of the state. Many areas with higher values of RMSE lie near the boundaries of the convex hulls. The observations near these boundaries (Figure 6.3a) have the largest amount of uncertainty, often jumping cluster assignments throughout the posterior sampling.



(a) Posterior prediction of Brook trout presence (b) Root mean-squared error for predictions

Figure 6.17: Posterior prediction and root mean-squared error for Brook trout presence.

6.3.5 Comparison to K-means

A more automatic, albeit less flexible, approach to modeling the EBTJV dataset would be to use a clustering technique on the spatial coordinates and then perform model selection in the pre-defined clusters. The performance of this technique depends on the relative position and density of the spatial coordinates. Unlike HGP, running K-means before modeling will result in cluster assignments that do not consider the response or covariates. To denomstrate the difference, we will set $K = 3$ for K-means, cluster the observations based on their coordinates, and fit a Gaussian process regression using Indepenent-SSVS.

Figure 6.18 shows the cluster assignments when using K-means on the coordinates with three centroids. Fitting the model to the 500 observations and then predicting at the remaining 585 locations, the overall AFCCF is 0.754. The AFCCF on the true brook trout presence and absence locations is 0.795 and 0.694, respectively. The performance on the sites with no brook trout is significantly worse than that of HGP ($AFCCF = 0.751$). In this particular example, K-means does fairly well overall since the clusters happen to run vertically; the western cluster closely matches cluster 2 from HGP. However, if the spatial layout of the sites changed, K-means could produce very different results.

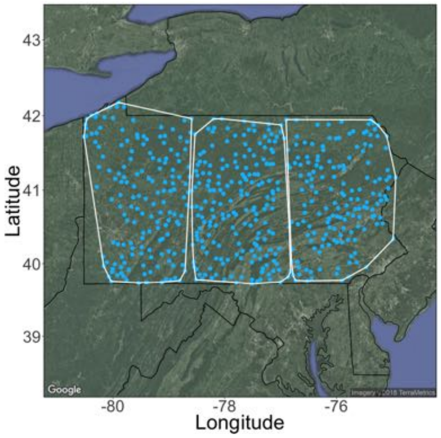


Figure 6.18: Cluster assignments using K-means with three centroids.

In a general setting, model-based clustering will allow locations to group together based on how prediction quality. As seen in Section 6.3.2, with the specified $R = 3$ clusters, the state was partitioned in a manner that created a western, southern, and central/eastern cluster. The assignments were made at each location based on which model predicted it the best. By clustering first, we can only use coordinates to make the clustering decision.

Chapter 7

Conclusions and future direction

7.1 Conclusions

In this work, we presented a method for performing model selection in the presence of a nonstationary (mean and covariance structure) spatial field. Within the HGP methodology, we develop a technique for 1) creating spatial clusters, each with its own stationary model, 2) embedding correlation in the estimated intercepts and nonzero linear effects between clusters, and 3) associating the variable inclusion indicators between clusters. We use non-overlapping convex hulls to form spatial cluster assignments for the observed locations, fit a Gaussian process within each cluster, and simultaneously use CAR and Ising models to infer the intra-cluster relationships.

The HGP prior distributions are general enough that we are able to infer the level of spatial correlation between the clusters, rather than simply specifying a relationship. That is, if clusters are actually unrelated, we are able to infer that relationship. However, in scenarios where spatial correlation between the clusters does exist, the HGP methodology is beneficial. In Section 5.2, we show that when strong linear effects exist across clusters, we are better able to find the important covariates when using CAR-SSVS instead of Independent-

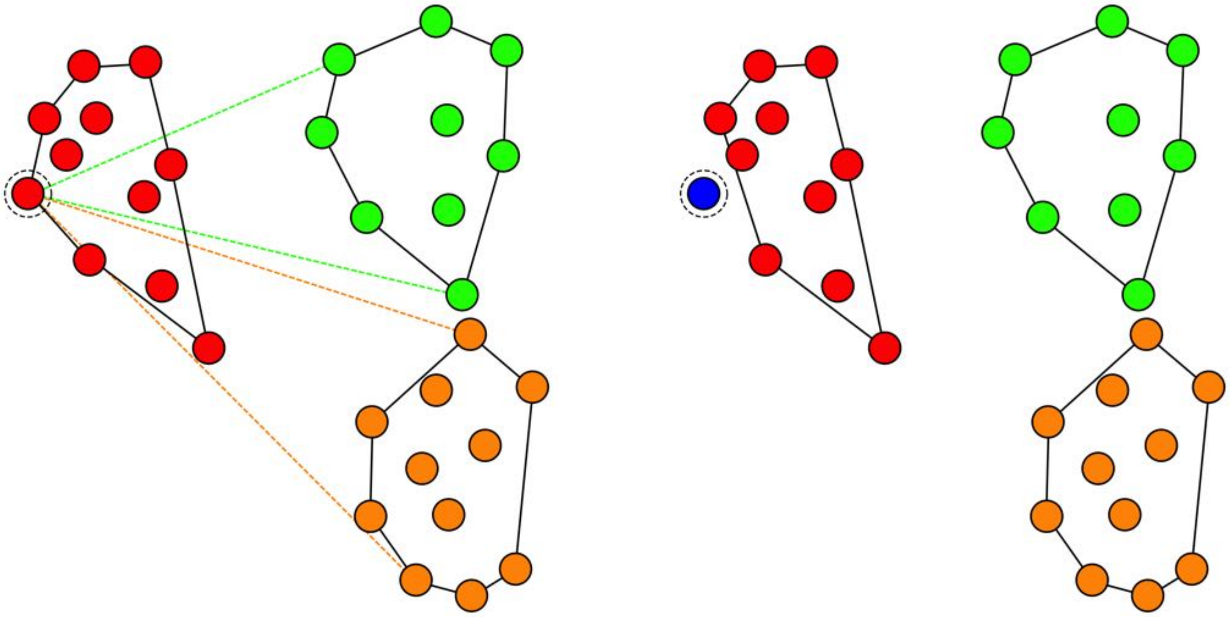
SSVS. The discrepancy in performance is larger when the nonzero effects become larger. Similarly, in Section 5.3, we demonstrate the improvement in model selection when inferring the spatial relationship between variable inclusion indicators. As the Ising parameter θ grows larger (relative to the weights), it is more important to learn its value instead of assuming independent models.

We applied the HGP methodology to a sample of 500 observations from the EBTJV dataset in Pennsylvania (Section 6.3.2). Finding $R = 3$ to be the optimal number of clusters (based on cross validation), we were able to model the nonstationary spatial process. We found two significant effects (Elevation and Total_Forest), where the importance of these effects varied depending on the location within Pennsylvania. Further, the two parameters that drive the intra-cluster dependency, θ and ϕ_β, were found to be nonzero. Therefore, we conclude that while the three clusters in Pennsylvania are best described with different models, the models are related (in terms of the linear effects and the variable inclusion indicators).

While the inference from the HGP methodology is useful for learning about nonstationary spatial processes, there are computational considerations that can make implementation difficult. In Section 4.2.2, we discussed the procedure for proposing cluster assignments for each spatial location. For locations that exist on the convex hull, we have to perform rank-one updates to each cluster's covariance matrix to compute the probability of a move. The computational burden is greatly reduced when compared to a clustering scheme without convex hulls, but if clusters are large the proposal can become demanding. The one-at-a-time cluster assignment proposals allow easy movement for a single point, however, making drastic changes in the cluster layout across space is difficult. In Section 4.3.1 we developed a cluster assignment proposal that allows large structural changes in the convex hulls. While there is an improvement in the MCMC mixing, the posterior sampling is challenging and requires a long run-time to ensure convergence.

7.2 Future research

The largest limitation to the HGP methodology is the MCMC mixing of the cluster assignments (and the associated non-overlapping convex hulls). Currently, we specify the number of clusters (R) and the non-overlapping convex hull constraint will always ensure that each point is in a cluster and that exactly R clusters exist. Points lying on a convex hull near the outside boundary of the entire dataset will not have an opportunity to move clusters unless the assignments shift around it, since its assignment to another cluster will inevitably lead to overlapping convex hulls (Figure 7.1a). Merging the current HGP methodology with the use of a Dirichlet process (Antoniak 1974) could improve the movement. Since Dirichlet process models are nonparametric, an observation on the convex hull could break away from its current cluster to form a new cluster (Figure 7.1a and 7.1b). However, since some of these new clusters could be small, special care would be necessary when specifying proper prior distributions for the other unknown parameters.



(a) Point ineligible for move

(b) Point forms own cluster

Figure 7.1: The circled point is on the outside of the dataset and is ineligible to move clusters. However, under a Dirichlet process the point make break off to form its own cluster.

There is also research to pursue with the weight matrix $\boldsymbol{W}$ that we utilize in both the CAR and Ising models. In Section 3.2.3 we chose $w_{r,r'} = 1/d_{r,r'}$ for $r \neq r'$, where $d_{r,r'}$ was based on the Euclidean distance between the cluster centroids. Since the cluster assignments change throughout the posterior sampling, $\boldsymbol{W}$ needs to be updated often. Choosing the centroid-to-centroid distance produced an easy-to-compute metric that produced a weight that decayed with distance. However, it could be advantageous to dissimilarity measures between clusters, as well as the relative position between clusters. For example, consider Figure 7.2. One cannot draw a line between any point in the red cluster to connect to the green cluster without intersecting the orange cluster. Given this position, it could be reasonable to set the weight to 0, implying a spatial neighborhood structure.

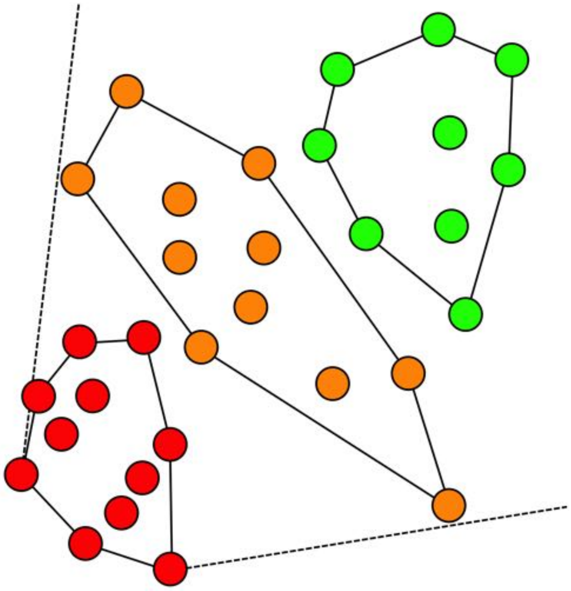


Figure 7.2: No line can be drawn to connect any point from the red cluster with the green cluster without intersecting the orange cluster.

Using spatial (longitude/latitude) trees would provide another opportunity to address neighborhoods structure and the number of clusters. Merging the HGP methodology with the treed GP of Gramacy & Lee (2008), we would have the flexibility to infer the number of clusters (tree nodes) via the posterior distribution while using the tree partitions as a first-order neighborhood structure (Figure 7.3). However, the posterior sampling of trees has its own difficulties that would need to be considered.

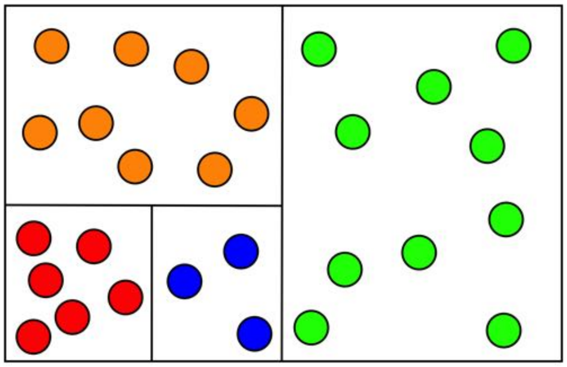


Figure 7.3: Example of two-dimensional tree (longitude/latitude). An adjacency matrix can be created using partitioning lines. For example, the red cluster will only have nonzero weights associated with the orange and blue clusters.

The HGP methodology presented in this thesis provides a framework for creating spatial clusters and modeling the intra-cluster dependency. For nonstationary spatial processes, we are able to learn how to cluster spatial locations together and to perform model selection within each cluster. While the MCMC mixing of the convex hulls can be challenging, the results of this methodology provide inference on how regression models are chosen and estimated in the presence of model dependency. The prior distributions used in HGP are highly modular and can be incorporated with many additional clustering schemes.

Appendix A

Full conditional distributions

The full conditional distribution for $(\tau^{(r)})^2$ is

$$\begin{aligned} P((\tau^{(r)})^2|\boldsymbol{y}^{(r)}, \mu^{(r)}, \boldsymbol{\beta}^{(r)}, \phi^{(r)}, g^{(r)}) &\propto \big((\tau^{(r)})^2\big)^{-n_r/2-1} exp\bigg\{-\frac{1}{(\tau^{(r)})^2}\frac{1}{2}(\tilde{\boldsymbol{y}}^{(r)})'(\boldsymbol{\rho}(\phi^{(r)})+g\boldsymbol{I})^{-1}(\tilde{\boldsymbol{y}}^{(r)})\bigg\} \\ &= InverseGamma\big(n_r/2, (\tilde{\boldsymbol{y}}^{(r)})'(\boldsymbol{\rho}(\phi^{(r)})+g\boldsymbol{I})^{-1}(\tilde{\boldsymbol{y}}^{(r)})/2\big), \end{aligned}$$

where $\tilde{\boldsymbol{y}}^{(r)} = \boldsymbol{y}^{(r)} - \mu^{(r)}\mathbf{1} - \boldsymbol{X}^{(r)}\boldsymbol{\beta}^{(r)}$ and $\boldsymbol{\rho}(\phi^{(r)})$ is the correlation matrix for observations in cluster r.

The joint full conditional distribution for $\phi^{(r)}$ and $g^{(r)}$ is

$$P(\phi^{(r)}, g^{(r)}|\boldsymbol{y}^{(r)}, \mu^{(r)}, \boldsymbol{\beta}^{(r)}, \tau^{(r)}) \propto |\boldsymbol{\Sigma}^{(r)}|^{-1/2} exp\bigg\{-\frac{1}{2}(\tilde{\boldsymbol{y}}^{(r)})'(\boldsymbol{\Sigma}^{(r)})^{-1}(\tilde{\boldsymbol{y}}^{(r)})\bigg\} P(\phi^{(r)}, g^{(r)}),$$

where $\tilde{\boldsymbol{y}}^{(r)} = \boldsymbol{y}^{(r)} - \mu^{(r)}\mathbf{1} - \boldsymbol{X}^{(r)}\boldsymbol{\beta}^{(r)}$, $\boldsymbol{\Sigma}^{(r)} = (\tau^{(r)})^2(\boldsymbol{\rho}(\phi^{(r)}) + g\boldsymbol{I})$, and $\boldsymbol{\rho}(\phi^{(r)})$ is the correlation matrix for observations in cluster r.

The full conditional distribution for $\mu^{(r)}$ is

$$P(\mu^{(r)}|\boldsymbol{y}^{(r)},\boldsymbol{\mu}^{-(r)},\boldsymbol{\beta}^{(r)},\boldsymbol{\Sigma}^{(r)},\boldsymbol{K}) \propto exp\Big\{-\frac{1}{2}\Big[(\mu^{(r)})^2(\mathbf{1}'(\boldsymbol{\Sigma}^{(r)})^{-1}\mathbf{1}+C^{-1}) - 2\mu^{(r)}(\mathbf{1}'(\boldsymbol{\Sigma}^{(r)})^{-1}\tilde{\boldsymbol{y}}^{(r)}+C^{-1}m)\Big]\Big\}$$
$$= Normal(E,V),$$

where $\tilde{\boldsymbol{y}}^{(r)} = \boldsymbol{y}^{(r)} - \boldsymbol{X}^{(r)}\boldsymbol{\beta}^{(r)}$, $\boldsymbol{\Sigma}^{(r)} = (\tau^{(r)})^2(\boldsymbol{\rho}(\phi^{(r)}) + g\boldsymbol{I})$, $\boldsymbol{K} = \sigma_\beta^2(\boldsymbol{I} - \phi_\beta\boldsymbol{W})^{-1}$, $C = \boldsymbol{K}_{r,r} - \boldsymbol{K}_{r,-r}\boldsymbol{K}_{-r,-r}^{-1}\boldsymbol{K}_{-r,r}$, $m = \boldsymbol{K}_{r,-r}\boldsymbol{K}_{-r,-r}^{-1}\boldsymbol{\mu}^{-(r)}$, $V = (\mathbf{1}'(\boldsymbol{\Sigma}^{(r)})^{-1}\mathbf{1} + C^{-1})^{-1}$, and $E = V(\mathbf{1}'(\boldsymbol{\Sigma}^{(r)})^{-1}\tilde{\boldsymbol{y}}^{(r)} + C^{-1}m)$.

$\beta_j^{(r)}$ and $\gamma_j^{(r)}$ are drawn from their joint full conditional distribution by first sampling the full conditional for $\gamma_j^{(r)}$ with $\beta_j^{(r)}$ integrated out. Then, $\beta_j^{(r)}$ is drawn from its full conditional distribution. The full conditional for $\gamma_j^{(r)}$ with $\beta_j^{(r)}$ integrated out is

$$P(\gamma_j^{(r)} = 1|\boldsymbol{y}^{(r)},\boldsymbol{\beta}_j^{-(r)},\boldsymbol{\beta}_{-j}^{(r)},\mu_j^{(r)},\boldsymbol{\Sigma}^{(r)},\boldsymbol{K},\theta,\boldsymbol{\gamma}_j^{-(r)}) = \left(1 + \frac{1-p_j^{(r)}}{p_j^{(r)}}\frac{\phi^*(E,V)}{\phi^*(m,C)}\right)^{-1}, \tag{A.1}$$

where $\phi^*(a,b)$ denotes the density of a Normal distribution at 0 with mean a and variance b, $C = \boldsymbol{K}_{r,r} - \boldsymbol{K}_{r,-r}\boldsymbol{K}_{-r,-r}^{-1}\boldsymbol{K}_{-r,r}$, $m = \boldsymbol{K}_{r,-r}\boldsymbol{K}_{-r,-r}^{-1}\boldsymbol{\beta}_j^{-(r)}$, $V = (\boldsymbol{x}_j'(\boldsymbol{\Sigma}^{(r)})^{-1}\boldsymbol{x}_j + C^{-1})^{-1}$, $E = V(\boldsymbol{x}_j'(\boldsymbol{\Sigma}^{(r)})^{-1}\tilde{\boldsymbol{y}}^{(r)} + C^{-1}m)$, $\tilde{\boldsymbol{y}}^{(r)} = \boldsymbol{y}^{(r)} - \mu^{(r)}\mathbf{1} - \boldsymbol{X}_{-j}^{(r)}\boldsymbol{\beta}_{-j}^{(r)}$, $\boldsymbol{x}_j$ is the j^{th} column of $\boldsymbol{X}^{(r)}$ and $p_j^{(r)} = \Big(1 + exp\Big\{\theta\sum_{r\sim r'} w_{r,r'}\big[\delta(\gamma_j^{(r')} = 0) - \delta(\gamma_j^{(r')} = 1)\big]\Big\}\Big)^{-1}$.

The full conditional distribution for $\beta_j^{(r)}$ is then

$$P(\beta_j^{(r)}|\boldsymbol{y}^{(r)},\boldsymbol{\beta}_j^{-(r)},\boldsymbol{\beta}_{-j}^{(r)},\mu_j^{(r)},\boldsymbol{\Sigma}^{(r)},\boldsymbol{K},\gamma_j^{(r)}) = \begin{cases}\delta(\beta_j^{(r)} = 0), & \text{if } \gamma_j^{(r)} = 0\\ Normal(E,V), & \text{if } \gamma_j^{(r)} = 1,\end{cases}$$

where E and V are as defined as in Equation A.1.

The full conditional distribution for τ^2_β is given by

$$\begin{aligned} P(\sigma^2_\beta|\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \phi_\beta) &\propto (\sigma^2_\beta)^{-(R+|\boldsymbol{\gamma}|)/2-1} exp\bigg\{ -\frac{1}{\tau^2_\beta}\frac{1}{2}\Big[\boldsymbol{\mu}'(\boldsymbol{I}-\phi_\beta \boldsymbol{W}_0)\boldsymbol{\mu} + \sum_{j=1}^{p} \boldsymbol{\beta}'_j(\boldsymbol{I}-\phi_\beta \boldsymbol{W}_j)\boldsymbol{\beta}_{\boldsymbol{j}}\Big]\bigg\} \\ &= InverseGamma\bigg((R+|\boldsymbol{\gamma}|)/2, \frac{1}{2}\Big[\boldsymbol{\mu}'(\boldsymbol{I}-\phi_\beta \boldsymbol{W}_0)\boldsymbol{\mu} + \sum_{j=1}^{p} \boldsymbol{\beta}'_j(\boldsymbol{I}-\phi_\beta \boldsymbol{W}_j)\boldsymbol{\beta}_{\boldsymbol{j}}\Big]\bigg) \end{aligned}$$

where $\boldsymbol{\mu} = (\mu^{(1)}, \dots, \mu^{(R)})'$, $\boldsymbol{\beta}_j$ is a vector of $\{\beta^{(r)}_j : \gamma^{(r)}_j = 1\}$, and $|\boldsymbol{\gamma}| = \sum_{j=1}^{p}\sum_{r=1}^{R} \gamma^{(r)}_j$.

The joint full conditional distribution for ϕ_β and g_β is

$$P(\phi_\beta|\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \tau_\beta) \propto \prod_{j=1}^{p} |\boldsymbol{K}_j|^{-1/2} exp\bigg\{ -\frac{1}{2}\Big[\boldsymbol{\mu}'\boldsymbol{K}_0^{-1}\boldsymbol{\mu} + \sum_{j=1}^{p} \boldsymbol{\beta}'_j \boldsymbol{K}_j^{-1}\boldsymbol{\beta}_{\boldsymbol{j}}\Big]\bigg\} P(\phi_\beta)$$

where $\boldsymbol{\mu} = (\mu^{(1)}, \dots, \mu^{(R)})'$ and $\boldsymbol{\beta}_j$ is a vector of $\{\beta^{(r)}_j : \gamma^{(r)}_j = 1\}$.

Appendix B

Proofs

Consider an n-dimensional realization from the Ising model where each location has weights $w_{r,r'} > 0$ and $y_r = 1$ for all r. Under the pseudo-likelihood model with flat prior distribution, the joint distribution for θ and $\boldsymbol{y}$ is

$$P(\theta, \boldsymbol{y}) = \prod_{r=1}^{n} \Big(1 + exp\{ - \theta \sum_{r \sim r'} w_{r,r'} \}\Big)^{-1}$$

and thus,

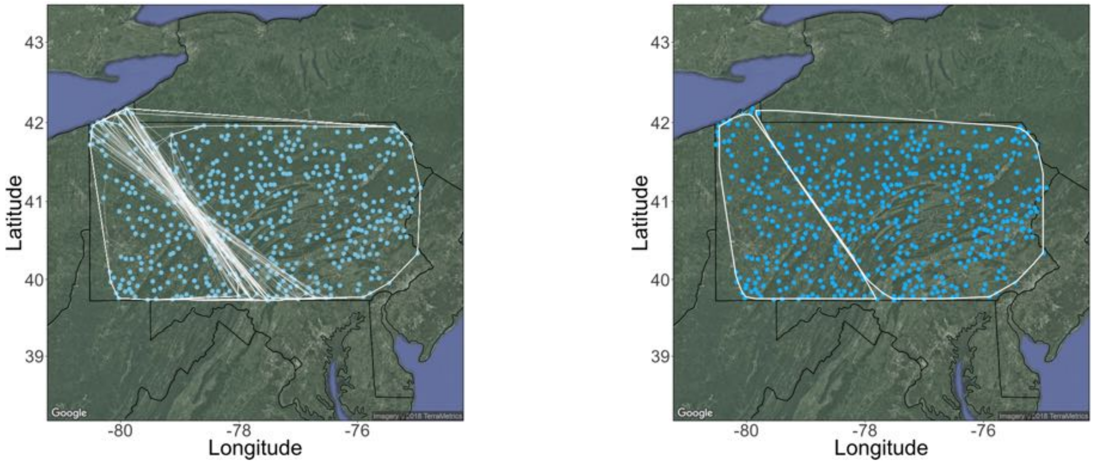
$$\begin{aligned} \int_0^\infty P(\theta, \boldsymbol{y}) d\theta &= \int_0^\infty \prod_{r=1}^{n} \Big(1 + exp\{ - \theta \sum_{r \sim r'} w_{r,r'} \}\Big)^{-1} d\theta \\ &\geq \int_0^\infty \prod_{r=1}^{n} \Big(1 + exp\{ - \theta \underset{r}{Max}\Big\{ \sum_{r \sim r'} w_{r,r'} \Big\}\}\Big)^{-1} d\theta \\ &= \int_0^\infty \Big(1 + exp\{ - \theta \underset{r}{Max}\Big\{ \sum_{r \sim r'} w_{r,r'} \Big\}\}\Big)^{-n} d\theta \end{aligned}$$

which diverges to ∞.

Appendix C

Additional EBTJV posterior analysis

C.1 Cluster assignments ($R = 2$)



(a) 250 draws from posterior distribution of cluster assignments

(b) Posterior "average" convex hull assignments

Figure C.1: Draws from the posterior distribution for cluster assignments when $R = 2$.

C.2 Linear effects and variable inclusion indicators

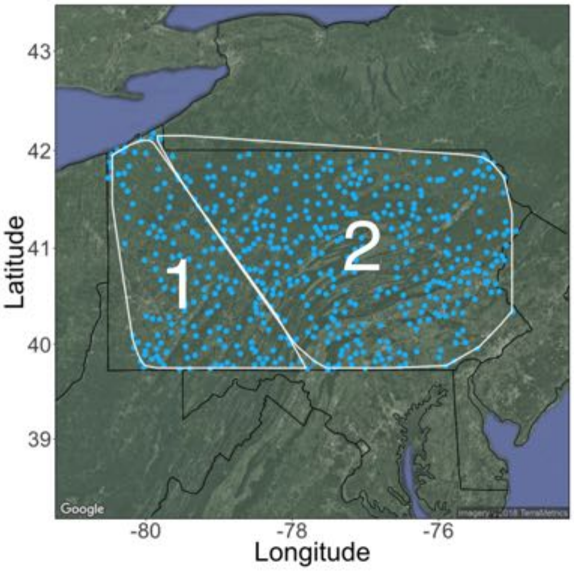


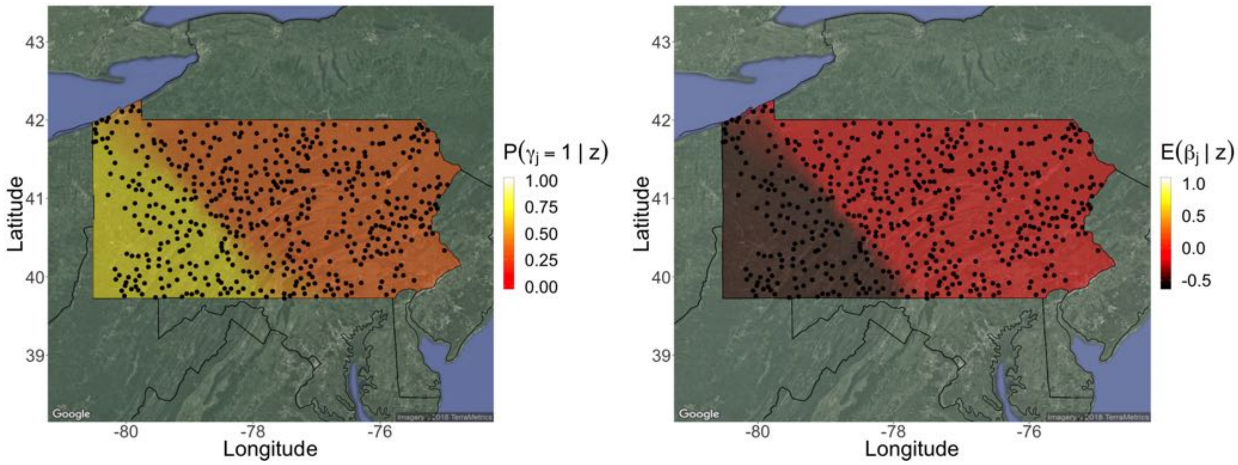
Figure C.2: Posterior "average" cluster assignments with labeled clusters.

Table C.1: Posterior mean for the variable inclusion indicators, $P(\gamma_j^{(r)} = 1|\boldsymbol{z})$ for $R = 2$.

Variable	**Cluster 1**	**Cluster 2**
Road_Density	0.688	0.260
Indust_Trans	0.389	0.154
Transitional	0.138	0.213
Ag	0.316	0.265
Mixed_Forest	0.292	0.247
Elevation	1.000	0.605
Total_Forest	0.548	0.990
Log_Chem	0.125	0.073

Table C.2: Posterior mean and 90% credible intervals for the linear effects, $\beta_j^{(r)}$, the intercepts, $\mu^{(r)}$, and the Gaussian process lengthscales, $\phi^{(r)}$.

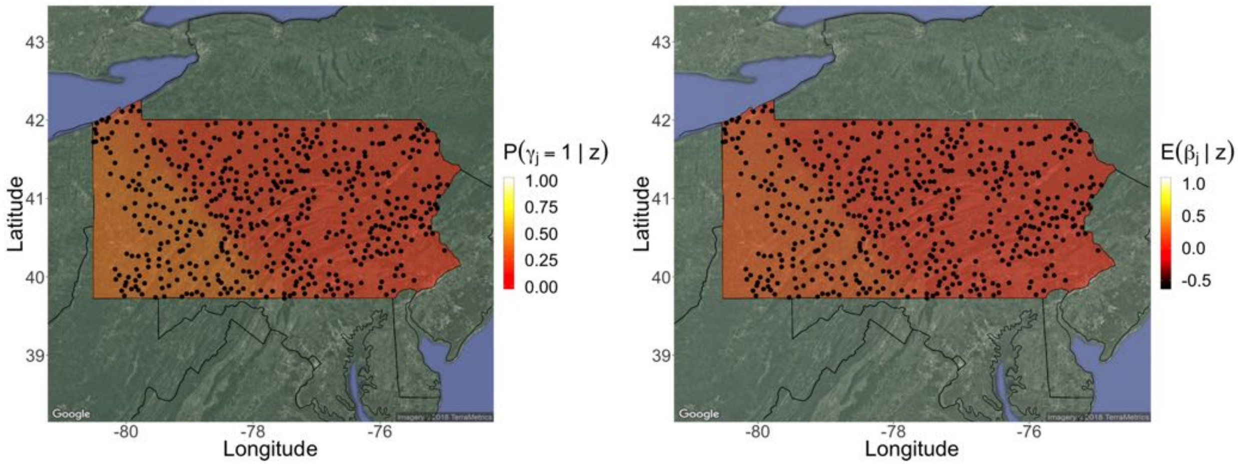
Variable	**Cluster 1**	**Cluster 2**
Road_Density	-0.555 (-1.518,0.000)	-0.036 (-0.267,0.000)
Indust_Trans	0.212 (0.000,1.011)	0.017 (0.000,0.172)
Transitional	0.005 (0.000,0.116)	0.037 (0.000,0.248)
Ag	0.136 (-0.072,0.966)	0.036 (-0.169,0.453)
Mixed_Forest	0.096 (0.000,0.578)	-0.047 (-0.304,0.000)
Elevation	1.028 (0.637,1.465)	0.156 (0.000,0.465)
Total_Forest	0.208 (-0.224,1.114)	0.917 (0.000,1.402)
Log_Chem	0.004 (-0.023,0.089)	0.003 (-0.352,0.000)
Intercept	-0.831 (-1.235,-0.450)	0.667 (0.486,0.854)
ϕ (Gaussian process)	0.010 (0.002,0.013)	0.007 (0.002,0.023)



(a) Posterior mean for variable inclusion

(b) Posterior mean linear effect

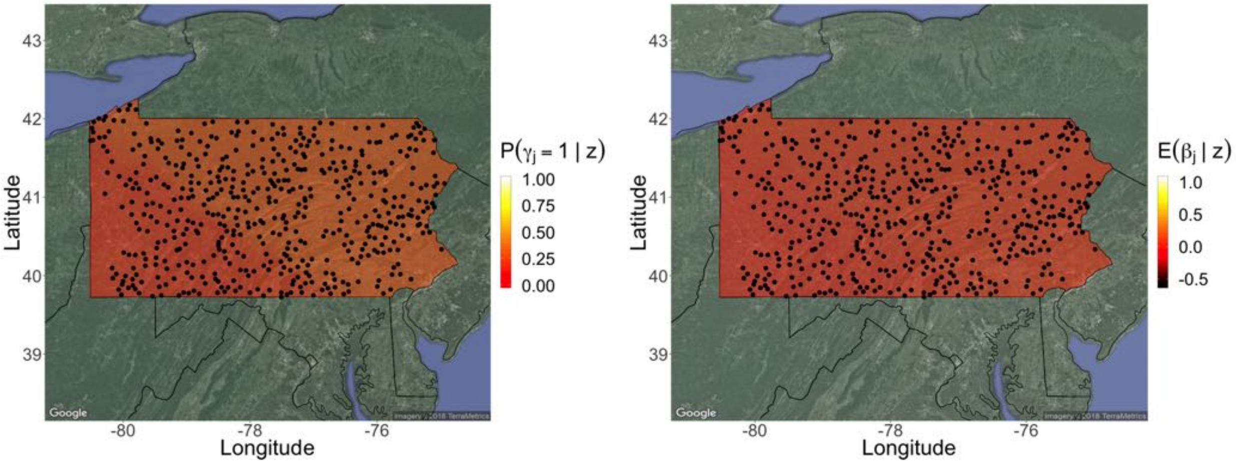
Figure C.3: Posterior mean across space of γ_j and β_j for Road_Density when $R = 2$.



(a) Posterior mean for variable inclusion

(b) Posterior mean linear effect

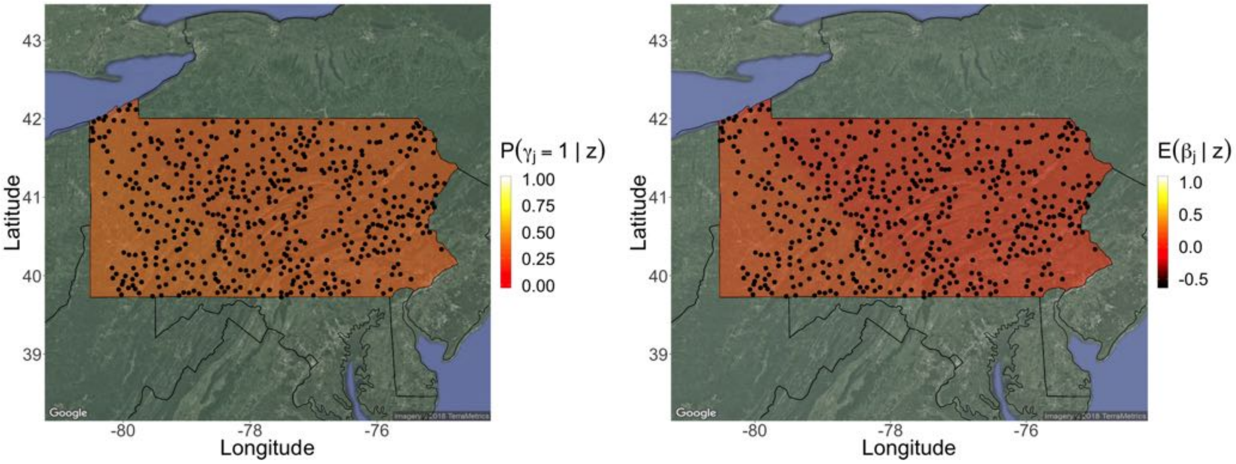
Figure C.4: Posterior mean across space of γ_j and β_j for Indust_Trans when $R = 2$.



(a) Posterior mean for variable inclusion

(b) Posterior mean linear effect

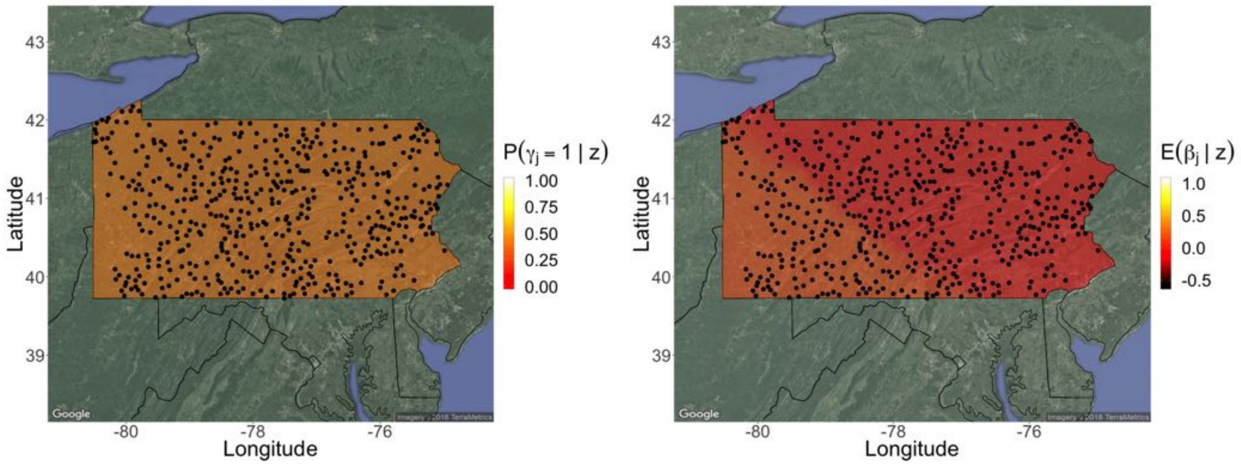
Figure C.5: Posterior mean across space of γ_j and β_j for Transitional when $R = 2$.



(a) Posterior mean for variable inclusion

(b) Posterior mean linear effect

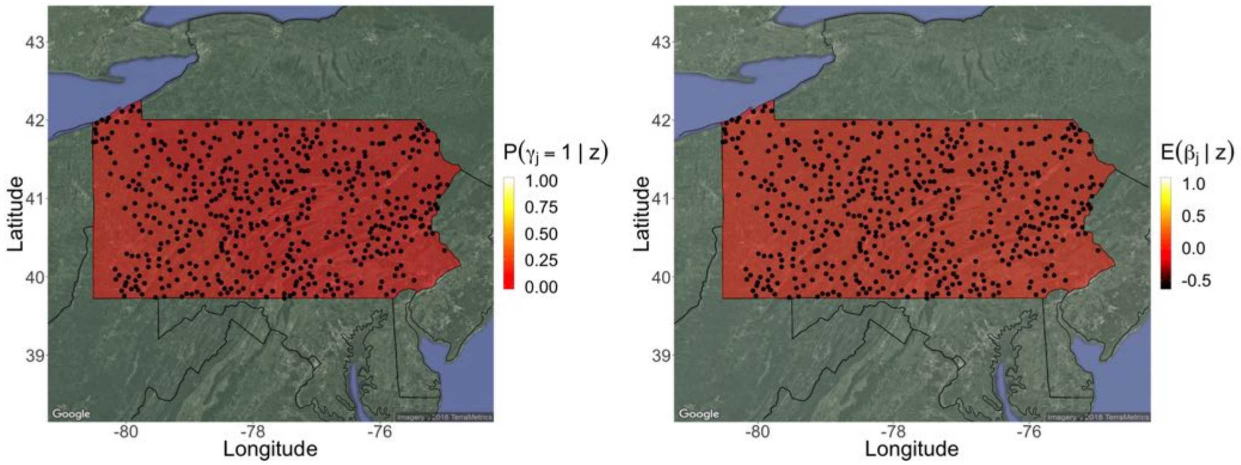
Figure C.6: Posterior mean across space of γ_j and β_j for Ag when $R = 2$.



(a) Posterior mean for variable inclusion

(b) Posterior mean linear effect

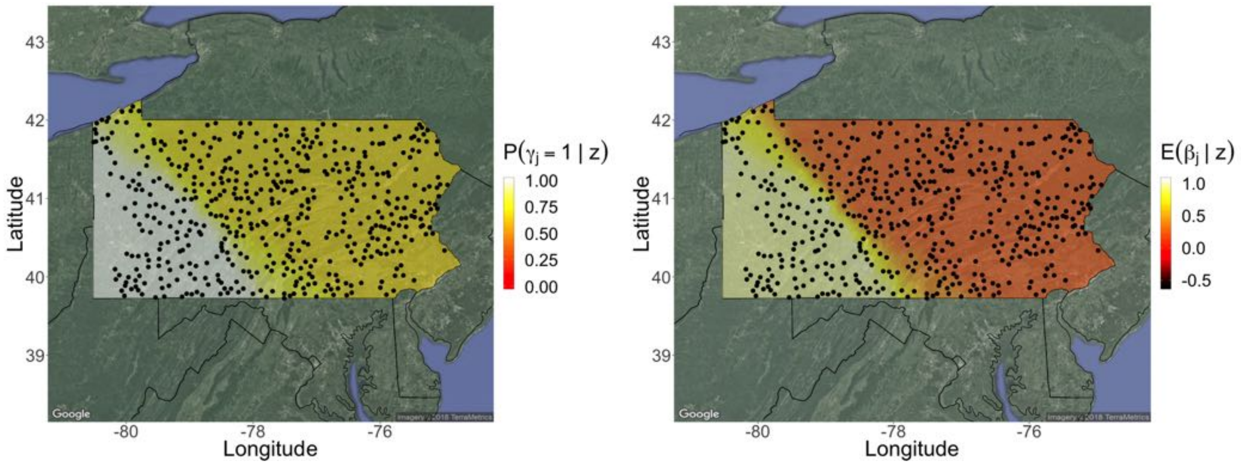
Figure C.7: Posterior mean across space of γ_j and β_j for Mixed_Forest when $R = 2$.



(a) Posterior mean for variable inclusion

(b) Posterior mean linear effect

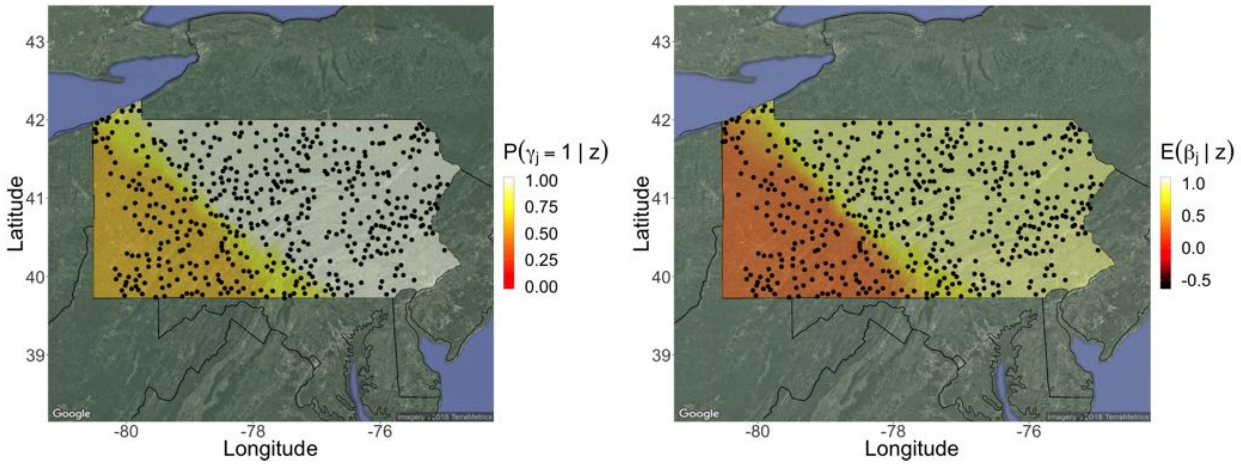
Figure C.8: Posterior mean across space of γ_j and β_j for Log_Chem when $R = 2$.



(a) Posterior mean for variable inclusion

(b) Posterior mean linear effect

Figure C.9: Posterior mean across space of γ_j and β_j for Elevation when $R = 2$.

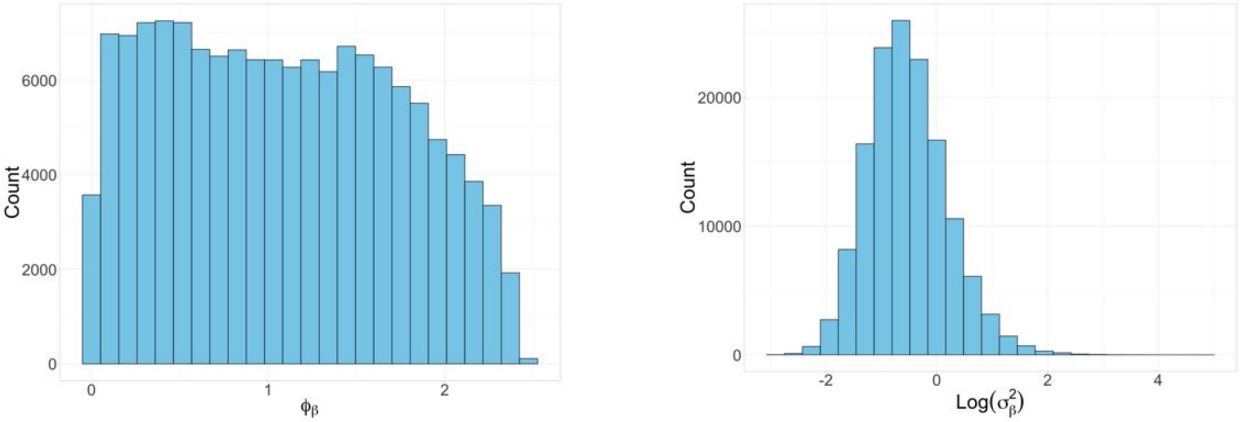


(a) Posterior mean for variable inclusion

(b) Posterior mean linear effect

Figure C.10: Posterior mean across space of γ_j and β_j for Total_Forest when $R = 2$.

C.3 CAR and Ising parameters



(a) $P(\phi_\beta|\boldsymbol{z})$

(b) $P(Log(\sigma^2_\beta)|\boldsymbol{z})$

Figure C.11: Draws from the marginal posterior distributions of ϕ_β and σ^2_β for $R = 2$

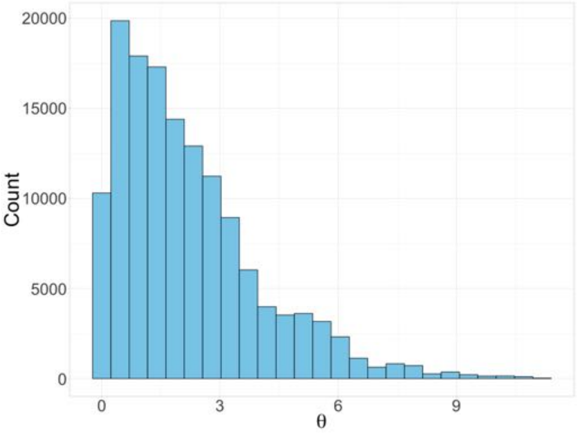
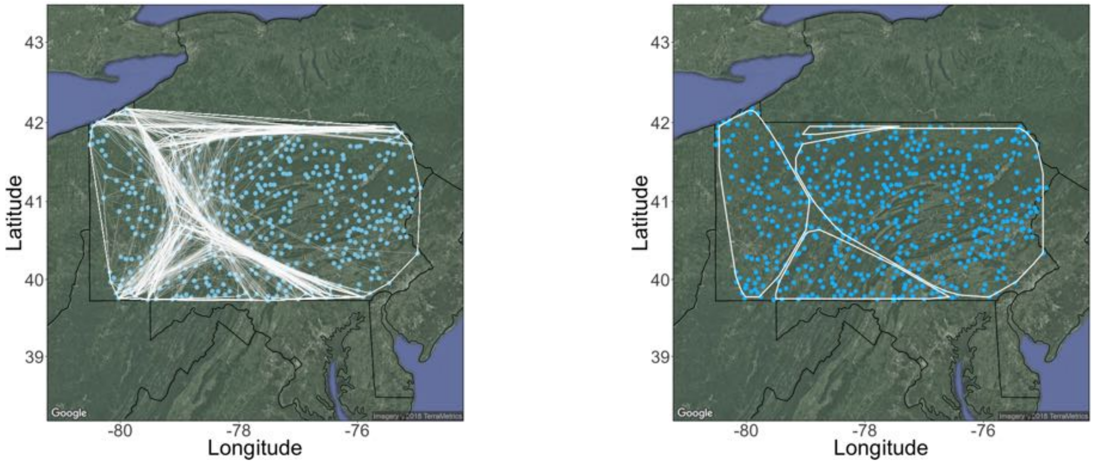


Figure C.12: Posterior distribution of θ for $R = 2$.

C.4 Cluster assignments ($R = 4$)



(a) 250 draws from posterior distribution of cluster assignments

(b) Posterior “average” convex hull assignments

Figure C.13: Draws from the posterior distribution for cluster assignments when $R = 4$.

C.5 Linear effects and variable inclusion indicators

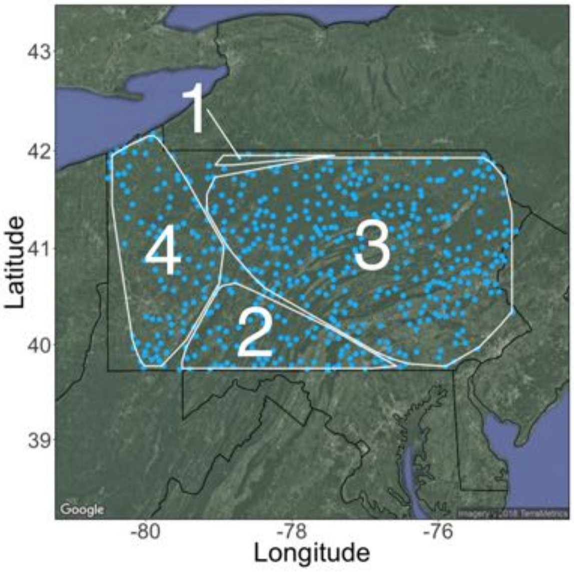


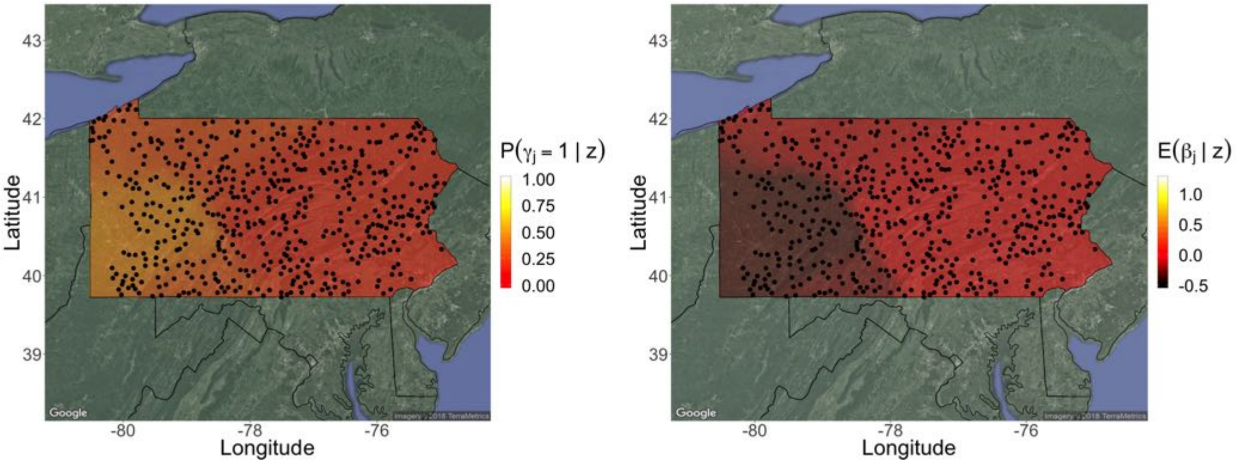
Figure C.14: Posterior “average” cluster assignments with labeled clusters.

Table C.3: Posterior mean for the variable inclusion indicators, $P(\gamma_j^{(r)} = 1|\boldsymbol{z})$ for $R = 4$.

Variable	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Road_Density	0.320	0.616	0.150	0.483
Indust_Trans	0.140	0.173	0.056	0.141
Transitional	0.169	0.048	0.093	0.151
Ag	0.536	0.589	0.283	0.455
Mixed_Forest	0.203	0.144	0.114	0.164
Elevation	0.921	0.994	0.479	0.954
Total_Forest	0.841	0.753	0.986	0.786
Log_Chem	0.170	0.087	0.043	0.114

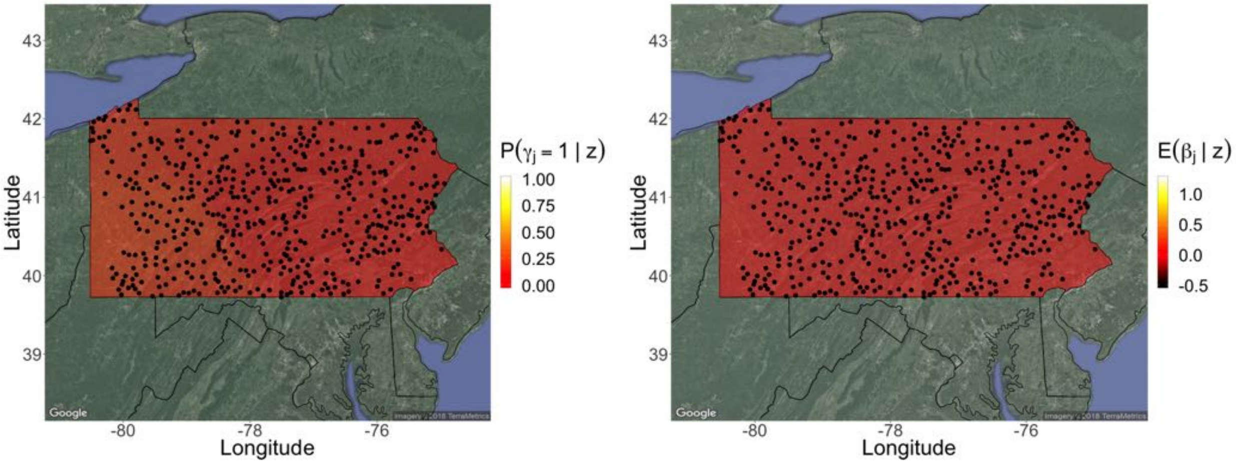
Table C.4: Posterior mean and 90% credible intervals for the linear effects, $\beta_j^{(r)}$, the intercepts, $\mu^{(r)}$, and the Gaussian process lengthscales, $\phi^{(r)}$ for $R = 4$.

Variable	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Road_Density	-0.090	-1.467	-0.015	-0.496
	(-1.107,0.495)	(-4.017,0.108)	(-0.163,0.000)	(-1.956, 0.000)
Indust_Trans	-0.010	0.125	0.007	0.018
	(-0.384,0.183)	(0.000,1.103)	(0.000,0.000)	(-0.131,0.398)
Transitional	0.137	0.002	0.016	0.082
	(0.000,1.071)	(0.000,0.000)	(0.000,0.160)	(0.000,0.812)
Ag	-0.792	1.091	0.033	-0.204
	(-3.654,0.500)	(-0.048,2.632)	(-0.188,0.430)	(-1.600,0.652)
Mixed_Forest	-0.083	0.053	0.007	0.079
	(-0.735,0.000)	(0.000,0.525)	(-0.011,0.079)	(-0.129,0.476)
Elevation	2.259	2.223	0.102	1.320
	(0.000,5.385)	(0.474,4.179)	(-0.030,0.452)	(0.000,2.806)
Total_Forest	1.203	-0.772	0.905	0.583
	(-0.072,3.231)	(-2.358,0.907)	(0.520,1.369)	(-0.960,2.306)
Log_Chem	0.128	0.127	0.003	0.014
	(-0.072,1.250)	(0.000,0.172)	(0.000,0.000)	(-0.056,0.158)
Intercept	-0.167	-0.381	0.635	-1.043
	(-1.766,1.528)	(-1.360,0.727)	(0.379,0.902)	(-1.822,0.026)
ϕ (Gaussian process)	0.023	0.202	0.009	0.013
	(0.001,0.077)	(0.001,0.841)	(0.002,0.017)	(0.001,0.048)



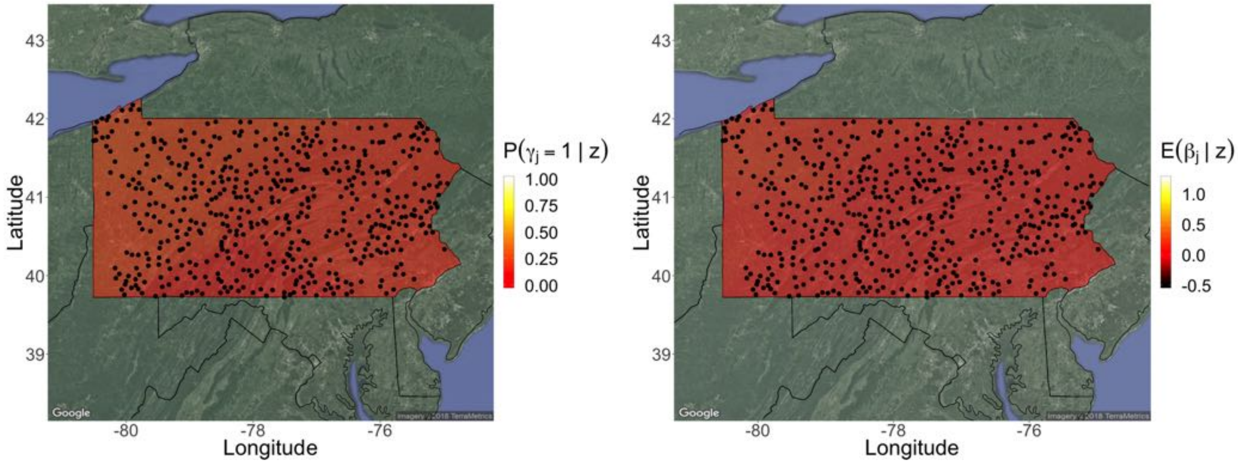
(a) Posterior mean for variable inclusion (b) Posterior mean linear effect

Figure C.15: Posterior mean across space of γ_j and β_j for Road_Density when $R = 4$.



(a) Posterior mean for variable inclusion (b) Posterior mean linear effect

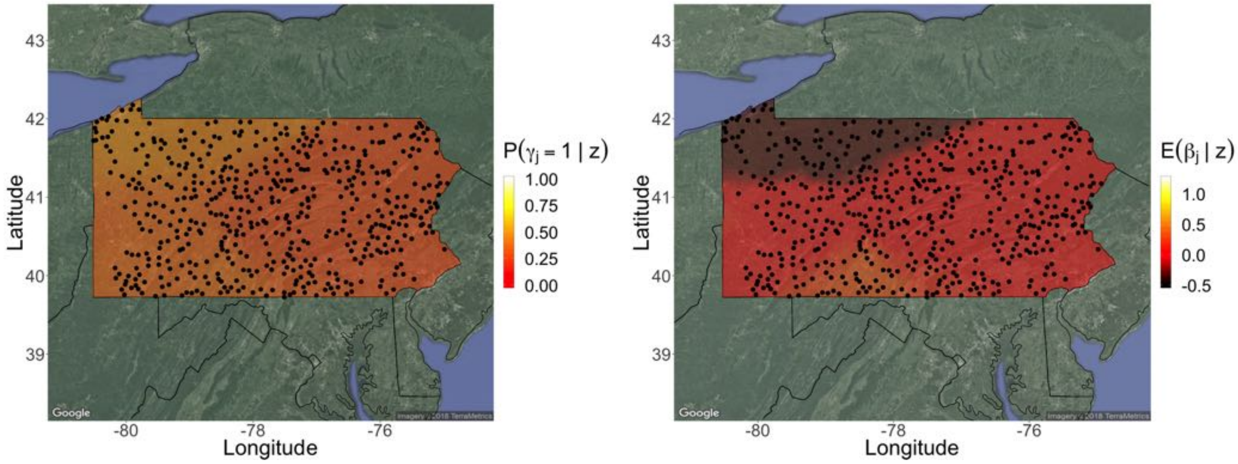
Figure C.16: Posterior mean across space of γ_j and β_j for Indust_Trans when $R = 4$.



(a) Posterior mean for variable inclusion

(b) Posterior mean linear effect

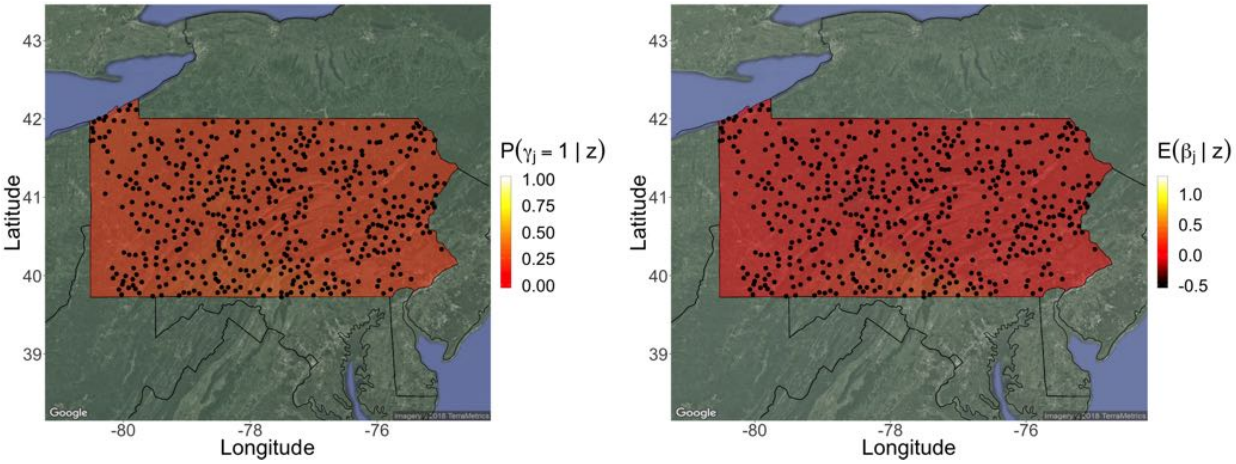
Figure C.17: Posterior mean across space of γ_j and β_j for Transitional when $R = 4$.



(a) Posterior mean for variable inclusion

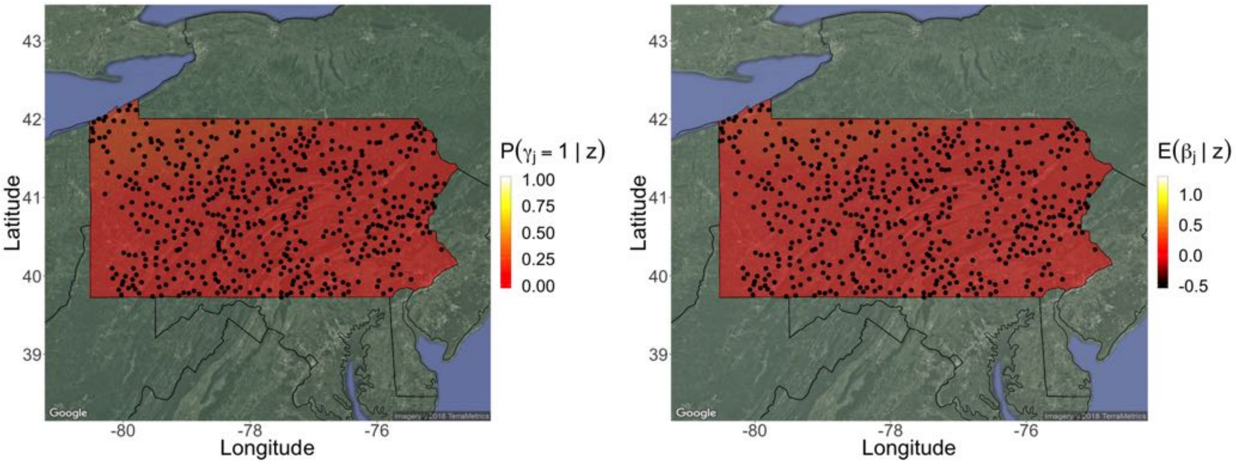
(b) Posterior mean linear effect

Figure C.18: Posterior mean across space of γ_j and β_j for Ag when $R = 4$.



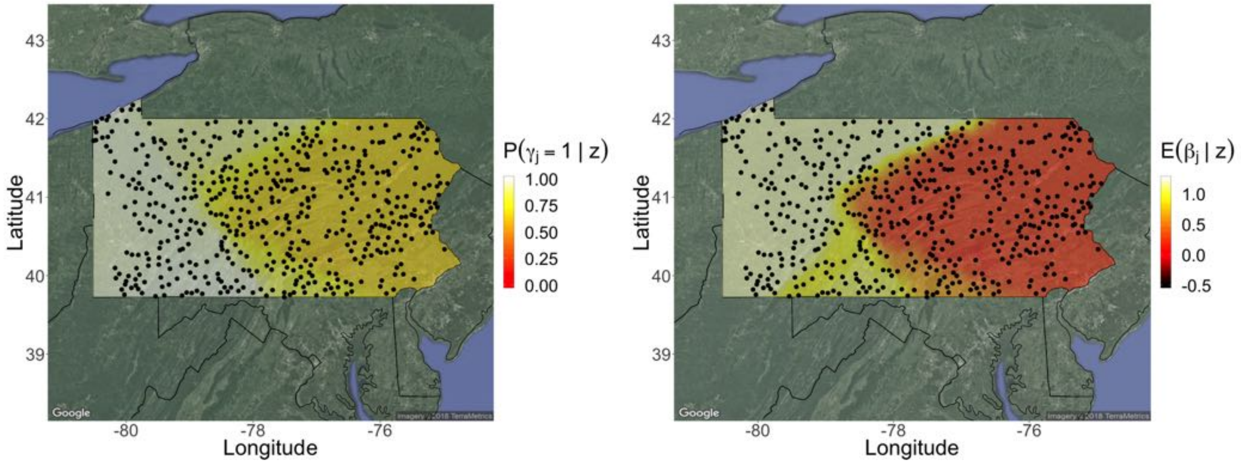
(a) Posterior mean for variable inclusion (b) Posterior mean linear effect

Figure C.19: Posterior mean across space of γ_j and β_j for Mixed_Forest when $R = 4$.



(a) Posterior mean for variable inclusion (b) Posterior mean linear effect

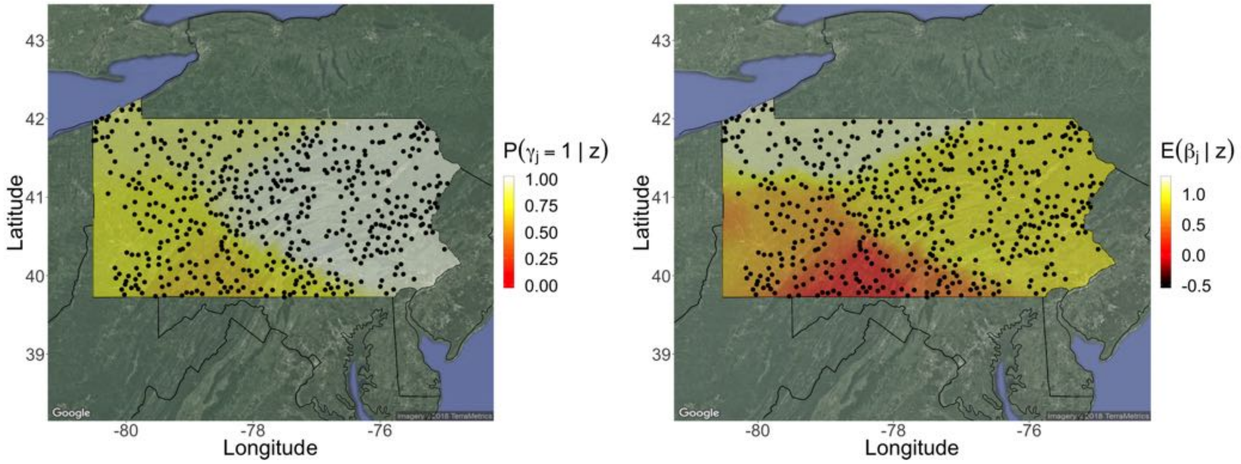
Figure C.20: Posterior mean across space of γ_j and β_j for Log_Chem when $R = 4$.



(a) Posterior mean for variable inclusion

(b) Posterior mean linear effect

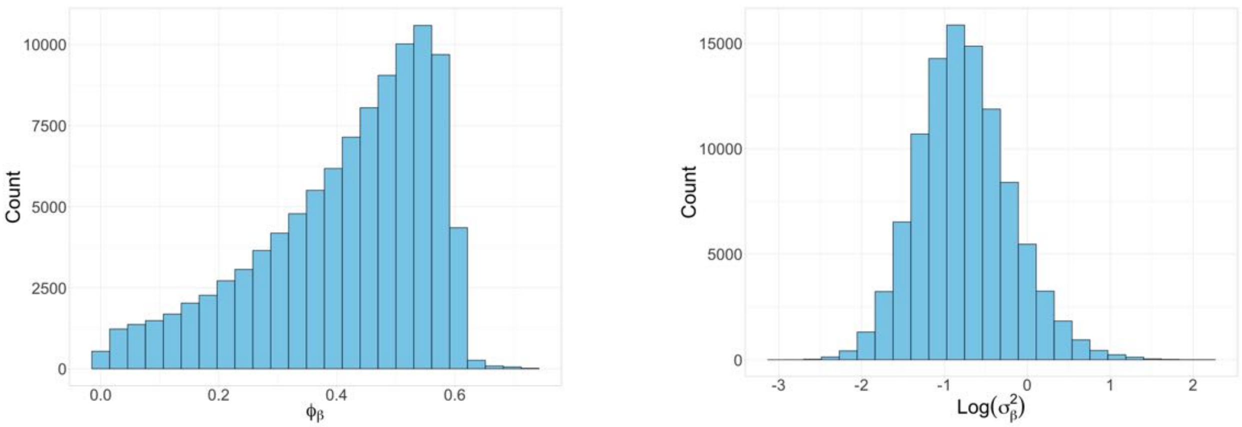
Figure C.21: Posterior mean across space of γ_j and β_j for Elevation when $R = 4$.



(a) Posterior mean for variable inclusion

(b) Posterior mean linear effect

Figure C.22: Posterior mean across space of γ_j and β_j for Total_Forest when $R = 4$.



(a) $P(\phi_\beta|\boldsymbol{z})$

(b) $P(Log(\sigma^2_\beta)|\boldsymbol{z})$

Figure C.23: Draws from the marginal posterior distributions of ϕ_β and σ^2_β for $R = 4$

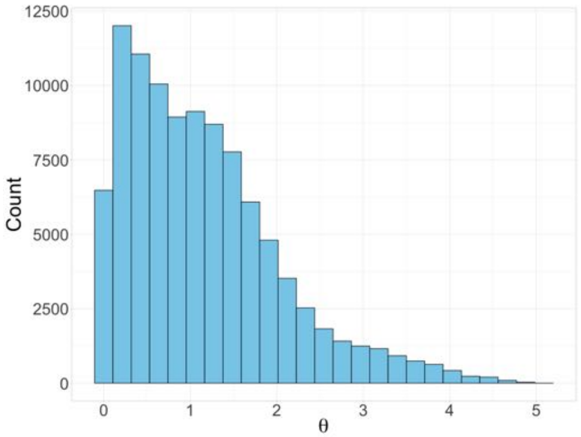


Figure C.24: Posterior distribution of θ for $R = 4$.

Bibliography

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control* **19**(6), 716–723.

Albert, J. H. & Chib, S. (1993), 'Bayesian analysis of binary and polychotomous response data', *Journal of the American Statistical Association* **88**(422), 669–679.

Antoniak, C. E. (1974), 'Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems', *The Annals of Statistics* pp. 1152–1174.

Banerjee, S., Carlin, B. P. & Gelfand, A. E. (2014), *Hierarchical Modeling and Analysis for Spatial Data*, CRC Press.

Berger, J. O., De Oliveira, V. & Sansó, B. (2001), 'Objective Bayesian analysis of spatially correlated data', *Journal of the American Statistical Association* **96**(456), 1361–1374.

Besag, J. (1974), 'Spatial interaction and the statistical analysis of lattice systems', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 192–236.

Besag, J. (1975), 'Statistical analysis of non-lattice data', *The Statistician* pp. 179–195.

Carvalho, C. M., Polson, N. G. & Scott, J. G. (2010), 'The horseshoe estimator for sparse signals', *Biometrika* **97**(2), 465–480.

Casella, G. & Berger, R. L. (2002), *Statistical Inference*, Vol. 2, Duxbury Pacific Grove, CA.

Claeskens, G., Hjort, N. L. et al. (2008), *Model Selection and Model Averaging*, Vol. 330, Cambridge University Press Cambridge.

Cressie, N. (2015), *Statistics for Spatial Data*, John Wiley & Sons.

De Oliveira, V. (2000), 'Bayesian prediction of clipped Gaussian random fields', *Computational Statistics & Data Analysis* **34**(3), 299–314.

De Oliveira, V. (2012), 'Bayesian analysis of conditional autoregressive models', *Annals of the Institute of Statistical Mathematics* **64**(1), 107–133.

De Oliveira, V., Kedem, B. & Short, D. A. (1997), 'Bayesian prediction of transformed Gaussian random fields', *Journal of the American Statistical Association* **92**(440), 1422–1433.

East Brook Trout Joint Venture (2018), `http://easternbrooktrout.org/`.

Ecker, M. D. & Gelfand, A. E. (1999), 'Bayesian modeling and inference for geometrically anisotropic spatial data', *Mathematical Geology* **31**(1), 67–83.

Friedman, J., Hastie, T. & Tibshirani, R. (2001), *The Elements of Statistical Learning*, Vol. 1, Springer Series in Statistics New York.

Gelman, A. & Meng, X.-L. (1998), 'Simulating normalizing constants: From importance sampling to bridge sampling to path sampling', *Statistical Science* pp. 163–185.

George, E. I. & McCulloch, R. E. (1993), 'Variable selection via Gibbs sampling', *Journal of the American Statistical Association* **88**(423), 881–889.

Geweke, J. (1996), 'Variable selection and model comparison in regression', *In Bayesian Statistics 5* .

Geyer, C. J. & Thompson, E. A. (1992), 'Constrained Monte Carlo maximum likelihood for dependent data', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 657–699.

Golub, G. H., Heath, M. & Wahba, G. (1979), ‘Generalized cross-validation as a method for choosing a good ridge parameter’, *Technometrics* **21**(2), 215–223.

Good, I. J. (1983), *Good Thinking: The Foundations of Probability and its Applications*, U of Minnesota Press.

Gramacy, R. B. & Apley, D. W. (2015), ‘Local Gaussian process approximation for large computer experiments’, *Journal of Computational and Graphical Statistics* **24**(2), 561–578.

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

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

Hastings, W. K. (1970), ‘Monte Carlo sampling methods using Markov chains and their applications’, *Biometrika* **57**(1), 97–109.

Higdon, D. (1998), ‘A process-convolution approach to modelling temperatures in the North Atlantic Ocean’, *Environmental and Ecological Statistics* **5**(2), 173–190.

Higdon, D. M. (1995), Spatial applications of Markov chain Monte Carlo for Bayesian inference, PhD thesis, University of Washington.

Higdon, D. et al. (2002), ‘Space and space-time modeling using process convolutions’, *Quantitative methods for current environmental issues* **3754**, 37–56.

Hoegh, A., Leman, S., Saraf, P. & Ramakrishnan, N. (2015), ‘Bayesian model fusion for forecasting civil unrest’, *Technometrics* **57**(3), 332–340.

Hoerl, A. E. & Kennard, R. W. (1970), ‘Ridge regression: biased estimation for nonorthogonal problems’, *Technometrics* **12**(1), 55–67.

Hudy, M., Thieling, T. M., Gillespie, N. & Smith, E. P. (2008), 'Distribution, status, and land use characteristics of subwatersheds within the native range of brook trout in the eastern United States', *North American Journal of Fisheries Management* **28**(4), 1069–1085.

Jeffreys, H. (1935), Some tests of significance, treated by the theory of probability, *in* 'Mathematical Proceedings of the Cambridge Philosophical Society', Vol. 31, Cambridge University Press, pp. 203–222.

Jeffreys, H. (1961), *Theory of probability (3rd edt.) Oxford University press*, Oxford University Press.

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

Kim, H.-M., Mallick, B. K. & Holmes, C. (2005), 'Analyzing nonstationary spatial data using piecewise Gaussian processes', *Journal of the American Statistical Association* **100**(470), 653–668.

Kullback, S. & Leibler, R. A. (1951), 'On information and sufficiency', *The Annals of Mathematical Statistics* **22**(1), 79–86.

Li, Q., Lin, N. et al. (2010), 'The Bayesian elastic net', *Bayesian analysis* **5**(1), 151–170.

Matérn, B. (2013), *Spatial variation*, Vol. 36, Springer Science & Business Media.

McCullagh, P. & Nelder, J. A. (1989), 'Generalized Linear Models, no. 37 in Monograph on Statistics and Applied Probability'.

Møller, J., Pettitt, A. N., Reeves, R. & Berthelsen, K. K. (2006), 'An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants', *Biometrika* **93**(2), 451–458.

Paciorek, C. J. & Schervish, M. J. (2006), 'Spatial modelling using a new class of nonstationary covariance functions', *Environmetrics* **17**(5), 483–506.

Park, T. & Casella, G. (2008), 'The Bayesian lasso', *Journal of the American Statistical Association* **103**(482), 681–686.

Prado, R. & West, M. (2010), *Time series: modeling, computation, and inference*, CRC Press.

Raftery, A. E. (1995), 'Bayesian model selection in social research', *Sociological methodology* pp. 111–163.

Rasmussen, C. E. & Ghahramani, Z. (2002), Infinite mixtures of Gaussian process experts, *in* 'Advances in Neural Information Processing Systems', pp. 881–888.

Ren, C. & Sun, D. (2013), 'Objective Bayesian analysis for CAR models', *Annals of the Institute of Statistical Mathematics* **65**(3), 457–472.

Schabenberger, O. & Pierce, F. J. (2001), *Contemporary statistical models for the plant and soil sciences*, CRC press.

Schwarz, G. et al. (1978), 'Estimating the dimension of a model', *The annals of statistics* **6**(2), 461–464.

Thieling, T. M. (2006), Assessment and predictive model for brook trout (Salvelinus fontinalis) population status in the eastern United States, Master's thesis, James Madison University.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

Velasco-Cruz, C. (2012), Spatially Correlated Model Selection (SCOMS), PhD thesis, Virginia Tech.

Whittle, P. (1954), 'On stationary processes in the plane', *Biometrika* pp. 434–449.

Wilkinson, L. (1999), 'SYSTAT 9', *SPSS, Chicago* .

Yaglom, A. M. (2012), *Correlation theory of stationary and related random functions: Supplementary notes and references*, Springer Science & Business Media.

Zhang, H., Thieling, T., Prins, S. C. B., Smith, E. P. & Hudy, M. (2008), 'Model-based clustering in a brook trout classification study within the eastern United States', *Transactions of the American Fisheries Society* **137**(3), 841–851.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2), 301–320.