5-2017

A Bayesian Variable Selection Method with Applications to Spatial Data

Xiahan Tang
University of Arkansas, Fayetteville

A Bayesian Variable Selection Method with Applications to Spatial Data

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Statistics and Analytics

by

Xiahan Tang
University of Florida
Bachelor of Science in Statistics, 2015

May 2017
University of Arkansas

This thesis is approved for recommendation to the Graduate Council.

Dr. Avishek Chakraborty
Thesis Director

Dr. Qingyang Zhang
Committee Member

Dr. Giovanni Petris
Committee Member

Abstract

This thesis first describes the general idea behind Bayes Inference, various sampling methods based on Bayes theorem and many examples. Then a Bayes approach to model selection, called Stochastic Search Variable Selection (SSVS) is discussed. It was originally proposed by George and McCulloch (1993). In a normal regression model where the number of covariates is large, only a small subset tend to be significant most of the times. This Bayes procedure specifies a mixture prior for each of the unknown regression coefficient, the mixture prior was originally proposed by Geweke (1996). This mixture prior will be updated as data becomes available to generate a posterior distribution that assigns higher posterior probabilities to coefficients that are significant in explaining the response. Spatial modeling method is described in this thesis. Prior distribution for all unknown parameters and latent variables are specified. Simulated studies under different models have been implemented to test the efficiency of SSVS. A real dataset taken by choosing a small region from the Cape Floristic Region in South Africa is used to analyze the plants distribution in that region. The original multi-cateogory response is transformed into a presence and absence (binary) response for simpler analysis. First, SSVS is used on this dataset to select the subset of significant covariates. Then a spatial model is fitted using the chosen covariates and, post-estimation, predictive map of posterior probabilities of presence and absence are obtained for the study region. Posterior estimates for the true regression coefficients are also provided along with map for spatial random effects.

Acknowledgements

I would like to say thanks to the graduate school for improving my thesis format. Thanks to everyone at the math department for helping me during my thesis. Thanks to my family for encouraging me throughout my education. Thanks to Dr. Datta, Mohammed and Ghadeer for coming to my mock defense. Thanks to Dr. Luecking for editing my format. Thanks to my committee members for asking great questions during my defense. And special thanks to my advisor Dr. Chakraborty for helping me learn the materials. Finally, I want to acknowledge John Silander, Jr. and Anthony Rebelo for the CFR dataset that I used in my thesis for model implementation.

List of Abbreviations

IG	Inverse Gamma
MCMC	Markov Chain Monte Carlo
MVN	Multivariate Normal
N	Normal
SSVS	Stochastic Search Variable Selection
Unif	Uniform

Table of Contents

1 Bayesian Inference

Suppose, we are given an observed dataset D with n observations $x_1, x_2, \cdots, x_n$ from a density $f(x|\theta)$. The prior distribution of θ is $\pi(\theta)$. To find the posterior distribution of θ after observing the data. We simply multiply the likelihood of the data and prior distribution of the parameter of interest. This process represented as

$$\pi(\theta|D) \propto \prod_{i=1}^{n}(f(x_i|\theta)) * \pi(\theta) \tag{1}$$

This was proved using the classic Bayes formula

$$\pi(\theta|D,\eta) = \frac{f(D|\theta)g(\theta|\eta)}{\int f(D|u)g(u|\eta)du} \tag{2}$$

Where η is the hyper-parameter for θ, we assume it is a known constant. The denominator of Bayes formula gives the marginal distribution of D.

To explore properties of the parameter such as mean, variance, quantiles etc., we need to analyze its posterior distribution. For example, suppose a local government is interest in bringing manufacturing jobs into its city. But it is not sure what is a good place to build a manufacturing facility. The government send sociologists to do research on all zip codes around the city. And these sociologists send questionaries online and through mail to residents across all zip codes around the city. Suppose we let the percentage of people willing to work in manufacturing across all zip codes be our parameter of interest. We are interested in which zip code has the highest posterior percentage. The prior distribution in this example can be explained as, in areas where it is mostly non-college educated and highly unemployed, the percentage will be big. And in areas where it is mostly college graduates and upper middle class, the prior percentage will be small. After observing the response from questionnaires, we can get a posterior distribution of the percentage by combining our assumption and the actual questionnaire response.

We have discussed what prior and posterior each means above. Now we are interested in learning how to sample from the posterior. If the posterior is in standard form (a Gaussian or t-distribution for example), finding its analytical properties is relatively simpler. However, most of the time, it is in non-standard form, making those computations intractable. For example, suppose we have a posterior distribution

$$f(\theta|D) = e^{((-\theta^2)+2sin(\theta))/log(1+tan(\theta))} \tag{3}$$

In this case, the posterior distribution is not in any known form, it is analytically challenging to compute any integral involving this density. Ordinary Monte Carlo says that instead of computing such integrals by exact method, an alternative approach is to simulate N samples of θ from the posterior, $\theta_1, \cdots, \theta_N$, when our simulated sample size is large enough, meaning $N \to \infty$. The chain of simulated samples from this posterior will converge to our target posterior density, we can then make inference on the target posterior based on these simulated samples. For example, by take the average of these samples $1/N \sum_{i=1}^{N} \theta_i$,when N is large, it will converge to true target posterior mean based on law of large numbers. When it is hard to sample from the target posterior density, Ordinary Monte Carlo will become not useful. In this case, Monte Carlo Markov Chain(MCMC) suggests that an alternative approach is to sample from the target posterior one parameter at a time. The algorithm looks like this: Supposed we propose a θ_1 based on our target posterior, we want to generate $\theta_{(2)}, \cdots, \theta_{(N)}$ so that we can compute the empirical mean of the target posterior. The Markov Chain Monte Carlo method that we are going to discuss are M-H sampling and Gibbs Sampling methods

1.1 Metropolis-Hastings Method:

This method, abbreviated as MH, provides an acceptance ratio that can be used to decided whether θ^{new} can be used as $\theta^{(2)}$ when the chain of simulated samples is at state $\theta^{(1)}$ (Hastings, 1970). The general form for the acceptance ratio is written as following

$$p_a = \frac{g(\theta^c)}{g(\theta^o)} * \frac{q(\theta^o|\theta^c)}{q(\theta^c|\theta^o)} \wedge 1 \tag{4}$$

The g function in the $M-H$ acceptance ratio is the target posterior density, the density that we know how to derive and evaluate but do not know how to sample from. The q function is the proposed conditional density. It represents the probability that a candidate θ^c is going to be proposed given the current value θ^o. In $M-H$ sampling method, the proposal density is used to generate samples for the candidate parameters, one at a time. Since this algorithm won't start on itself, we need to come up with a starting candidate θ^0 so that it can be used to generate θ^1. The choice for θ^0 is flexible, under the law of large number the chain will converge to the true θ no matter what the initial choice is. However, a choice of θ_0 from the tail of the target distribution may cause the chain to converge slowly. This can be avoided by using some point estimate such as MLE based on the data as a starting value for θ_0. Eventually, we will have a Markov chain $\theta_0, ..., \theta_N$ where N represents the number of iterations and it is generally chosen to be large. After some initial iterations, the chain is expected to converge to its stationary distribution. We call the initial runs as burn in period. We will ignore these burn in samples. We will explore our interest among the rest of samples. There has been many discussions on choosing appropriate proposal density. Generally speaking, a good proposal density will significantly reduce the time for the chain to converge. We will typically choose a proposal such that it matches the shape of the target posterior as much as possible.

There are two common ways of choosing proposals: random walk and independence sampler. We will first discuss random walk MH. Suppose the support of θ is the entire real line, the proposal for θ is apprуниately assumed to come from a normal distribution. A random walk Metropolis Hastings proposal suggests that an appropriate mean for θ^c to follow is θ^o (Chib and Greenberg, 1995). In other words, the mean of each candidate is proposed to be the mean of the previous value

$$\pi(\theta^c) = N(\theta^o, \tau^2)$$

The variance for the proposal density τ^2 needs to be carefully chosen. When τ^2 is too small, the candidate parameter is going to be extremely close to the current value of the parameter on the chain. Thus the algorithm would take long time to traverse the entire parameter domain resulting in poor mixing and slow convergence. So, we will need to run a extremely large of iterations for the parameter to converge to the target distribution, which is extremely inefficient and time consuming. When c^2 is chosen to be too large, θ^c may drift far away from θ^o, thus potentially resulting the ratio between $g(\theta^c)$ and $g(\theta^o)$ to be too small, which lowers the acceptance rate. A good choice for variance is a value that is not either too small or too large and result in between 20% and 40% acceptance rate. One can tweak the proposal variance to achieve an acceptance rate within this region. When the proposal density is normally distributed, the second part of this acceptance ratio is always 1. This can be shown by the following mathematical proof

$$\begin{aligned}
\frac{q(\theta^o|\theta^c)}{q(\theta^c|\theta^o)} &= \frac{\exp\{-\frac{1}{2}\frac{(\theta^c-\theta^o)^2}{c^2}\}}{\exp\{-\frac{1}{2}\frac{(\theta^o-\theta^c)^2}{c^2}\}} \\
&= \frac{\exp\{-\frac{1}{2}\frac{(\theta^c-\theta^o)^2}{c^2}\}}{\exp\{-\frac{1}{2}\frac{(\theta^c-\theta^o)^2}{c^2}\}} \\
&= 1
\end{aligned}$$

Therefore, due to the insignificance of the second part of the ratio, whether p_a is big or small is highly dependent on the first part $\frac{g(\theta^c)}{g(\theta^o)}$. We call this importance ratio. The decision rule says that accept θ^c as $\theta^{(new)}$ when a uniform random number $u < p_a$. Otherwise, set $\theta^{(c)} = \theta^o$. When p_a is close to 1, chance of acceptance is high. When p_a is small, θ is more likely to not move from its previous position. Therefore, by its nature, the acceptance ratio encourages movement where $g(\theta^c)$ is high and discourages movement where $g(\theta^c)$ is low.

Another way of choosing proposal is independence sampler where $\pi(\theta^c|\theta^o) = \pi(\theta^c)$. In other words, the proposed distribution of the current state is independent of the distribution of previous state. Gibbs sampler is a special case of independence sampling. We will talk more on that in later sections.

We have discussed cases when the true θ ranges the entire real line. However, there are cases when the true θ ranges the positive real line, choosing a normal proposal density thus will not be appropriate. In that case, the proposal density is typically chosen to be gamma, inverse gamma or log normal distribution. For example, suppose the true θ ranges entire positive real line, then assuming assigning a gamma proposal would be appropriate

$$q(\theta^c|\theta^o) \sim \Gamma(\alpha, \lambda) \tag{5}$$

The mean of this gamma proposal is set to be θ^o using random walk Metropolis Hastings. The variance can be set to any pre-specified constant c using following equations:

$$\frac{\alpha}{\lambda} = \theta^o, \ \ \frac{\alpha}{\lambda^2} = c$$

Remember previously in the normal proposal the second part of the acceptance ratio is exactly one using random walk MH. It may not be the same for non-normal proposals. For example, in a log-normal proposal it would depend on the ratio of θ^c and θ^o as:

$$\frac{q(\theta^o|\theta^c)}{q(\theta^c|\theta^o)} \propto \frac{\theta^c}{\theta^o} \frac{\exp - \frac{(\theta^o - \theta^c)^2}{2c^2}}{\exp - \frac{(\theta^c - \theta^o)^2}{2c^2}} = \frac{\theta^c}{\theta^o}$$

1.2 Gibbs Sampling

Gibbs sampling method, or Gibbs Sampler which what most people would call it, is a special case of Metropolis Hastings sampling method. We will first discuss the history of Gibbs sampler and then we will look at the relationship between Gibbs sampler and the more general Metropolis

Hastings, finally we will discuss the its running algorithm.

The idea was originally used in physics using conditional probability to explain temperature and pressure. It hasn't entered mainstream statistics circle until the 1990s, after Gelfand and Smith 1990 paper (Gelfand and Smith, 1990). It was a revolutionary invention in the field of statistics and Gibbs sampler is frequently used today.

Gibbs sampler is a special case of the independence MH proposal we discussed previous section, which is part of the more general Metropolis Hastings sampling method. In Gibbs sampler, the proposal density is set to be equal to the target posterior of the parameter of interest. That is, $g(\theta^c) = q(\theta^c|\theta^o)$. The will result in the acceptance ratio always being one

$$\begin{aligned} p_a &= \frac{g(\theta^c)}{g(\theta^o)} * \frac{q(\theta^o|\theta^c)}{q(\theta^c|\theta^o)} \\ &= \frac{g(\theta^c)}{g(\theta^o)} * \frac{g(\theta^o)}{g(\theta^c)} \\ &= 1 \end{aligned}$$

Therefore, by using Gibbs sampler, all the candidate parameters are automatically accepted, which makes the algorithm simpler than Metropolis Hastings.

Suppose there are n unknown parameters, $\theta_1, \theta_2 \cdots \theta_n$, and D is the observed dataset. The posterior joint density for all n parameters given the observed data can be represented as $\pi(\theta_1 \cdots \theta_n|D)$. We are interested in various properties about this target posterior such as mean and quantiles. When its density is not in standard form, sampling from it would be intractable. Suppose we have a joint density that is in the following form

$$\pi(\theta_1, \theta_2|D) = \frac{2ysin(\theta_1)\exp\{-\theta_2^2\}}{(2sin(\theta_1) + 3\theta_2)}$$

There is no known way to sample from this joint density. However, we do konw how to sample from full conditional densities for each of $\theta_1 ... \theta_n$ individually by treating all other elements in the density, including all other parameters other than θ_i, where i is the index for the current parameter whose full conditional we are trying to sample from, and the observed data as fixed constants.So, we sequentially sample the following n posterior full conditionals

$$\begin{aligned}
\theta_1 &\sim \pi(\theta_1|\theta_2,\theta_3,...,\theta_n,D) \\
\theta_2 &\sim \pi(\theta_2|\theta_1,\theta_3,..,\theta_n,D) \\
\vdots & \qquad\quad \vdots \\
\theta_n &\sim \pi(\theta_n|\theta_1,\theta_2,..,\theta_{n-1},D)
\end{aligned}$$

As you notice that each time we are only sampling one unknown parameter, the name full conditional means everything else other than the parameter of interest is given.

1.3 An example for Gibbs Sampling Algorithm

Suppose there is a dataset D with n observations $X_1 \cdots X_n \sim N(\mu,\sigma^2)$, the prior distribution for μ is $\pi(\mu) \sim N(m,c^2)$, and the prior distribution for σ^2 is $\pi(\sigma^2) = IG(a,b)$, and we want to explore various properties on the posterior $\pi(\mu,\sigma^2|D)$. Finding the joint posterior density for μ and σ^2 is a straight forward process

$$\begin{aligned}
\pi(\mu,\sigma^2|D) &\propto L(D|\mu,\sigma^2)\pi(\sigma^2)\pi(\mu) \\
&= (\frac{1}{\sqrt{2\pi\sigma^2}})^n exp(\frac{-\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2})(\sigma^2)^{-a-1} exp(-\frac{b}{\sigma^2}) \\
exp(-\frac{(\mu-m)^2}{2c^2}) & \\
&\propto (\sigma^2)^{-a-1-\frac{n}{2}} exp(\frac{-\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2} - \frac{b}{\sigma^2} - \frac{(\mu-m)^2}{2c^2})
\end{aligned}$$

After finding the joint posterior density. We need to compute the full conditional density individually for μ and σ^2 so that we can easily draw samples from each of these 2 densities. We are going to start with σ^2, we do this by treating μ as a given constant

$$\begin{aligned}\pi(\sigma^2|\mu,D) &\propto (\sigma^2)^{-a-1-\frac{n}{2}} \times exp(\frac{-\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2} - \frac{b}{\sigma^2}) \\ &\propto (\sigma^2)^{-(a+1)-\frac{n}{2}} \times exp(-\frac{1}{\sigma^2}(\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2} + b)) \\ &\sim IG(a+\frac{n}{2}, \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2} + b)\end{aligned}$$

The visualization of the distribution of the simulated samples of σ^2 is given in figure 1.1

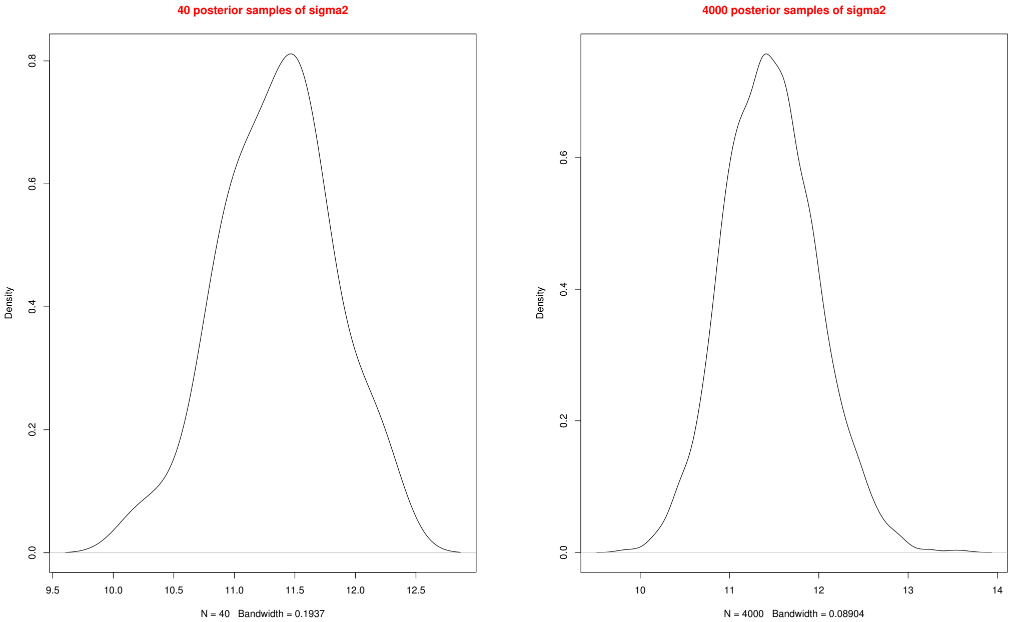

Figure 1.1: Posterior Density estimation for σ^2 with (L) 40 samples and (R) 4000 samples.

Now we find the full conditional density for μ, we do this by treating σ^2 as given constant.

$$\begin{aligned}
\pi(\mu|\sigma^2,D) &\propto exp(\frac{-\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}-\frac{(\mu-m)^2}{2c^2}) \\
&= exp-\frac{1}{2}(\frac{\sum_{i=1}^{n}(x_i^2+\mu^2-2x_i\mu)}{\sigma^2}+\frac{\mu^2+m^2-2mu}{c^2}) \\
&= exp-\frac{1}{2}(\frac{nx_i^2+n\mu^2-2\mu(n\bar{x})}{\sigma^2}+\frac{\mu^2+m^2-2mu}{c^2}) \\
&\propto exp-\frac{1}{2}(\frac{n\mu^2-2\mu(n\bar{x})}{\sigma^2}+\frac{\mu^2-2m\mu}{c^2}) \\
&= exp-\frac{1}{2}(\mu^2(\frac{n}{\sigma^2}+\frac{1}{c^2})-2\mu(\frac{n\bar{x}}{\sigma^2}+\frac{m}{c^2})) \\
&= exp(-\frac{1}{2}(\frac{n}{\sigma^2}+\frac{1}{c^2})(\mu^2-\frac{2\mu(\frac{n\bar{x}}{\sigma^2}+\frac{m}{c^2})}{\frac{n}{\sigma^2}+\frac{1}{c^2}})) \\
&\propto exp(-\frac{1}{2}(\frac{n}{\sigma^2}+\frac{1}{c^2})(\mu^2-\frac{2\mu(\frac{n\bar{x}}{\sigma^2}+\frac{m}{c^2})}{\frac{n}{\sigma^2}+\frac{1}{c^2}}+(\frac{\frac{n\bar{x}}{\sigma^2}+\frac{m}{c^2}}{\frac{n}{\sigma^2}+\frac{1}{c^2}}))) \\
&= exp(-\frac{1}{2}(\frac{n}{\sigma^2}+\frac{1}{c^2})(\mu-\frac{\frac{n\bar{x}}{\sigma^2}+\frac{m}{c^2}}{\frac{n}{\sigma^2}+\frac{1}{c^2}})^2) \\
&\sim N(\frac{\frac{n\bar{x}}{\sigma^2}+\frac{m}{c^2}}{\frac{n}{\sigma^2}+\frac{1}{c^2}},\frac{1}{\frac{n}{\sigma^2}+\frac{1}{c^2}})
\end{aligned}$$

The visualization of the distribution of the simulated samples of μ is given in figure 1.2

Our original goal was to explore various properties on the posterior $\pi(\mu,\sigma^2|D)$. We have found the full conditional densities for each of μ and σ^2, Monte Carlo method is going to be used to simulate large number of samples from each of these 2 distributions. As you can see from the graph above, as the number of simulations become larger, samples of σ^2 are looking more like an inverse gamma distribution, samples of μ are looking more like a normal distribution, and when the number of samples becomes extremely large, we expect them to converge fully to our target distribution.

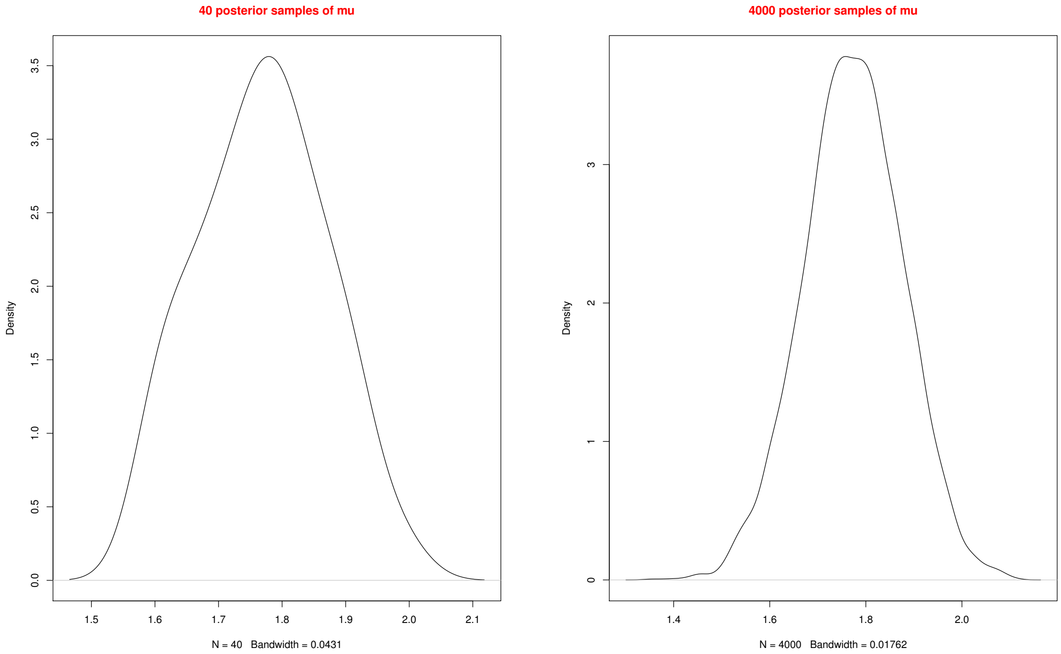

Figure 1.2: Posterior Density estimation for μ with (L) 300 simulations and (R) 30000 simulations.

2 A method for Bayesian variable selection

First of all, since adding covariates to a model will only have more of the response explained, why are we wanting to select covariates? The answer can be explained in a few points. First of all, generally in a large raw data set there will be a large number of covariates, and many of these covariates are correlated to each other. Collinearity will highly lower the model efficiency. Secondly, more covariates means that you need to spend money and time to observe them, which are unnecessary most of the time.

There are many examples on how model selection has benefited different areas of research. For example, In the famous Barley data (Tinker et al., 1996) which was taken from the North American Genome Mapping project. Statisticians was given a dataset with one response and 127 binary covariates. A classic data set is the pollution data set. It is studying mortality rates given 15 continuous covariates, including rainfall, January Temperature, July Temperature, population density etc.(O'Hara et al., 2009) In a paper about model selection with high dimensional data(Tadesse et al., 2005), the classic Iris dataset is being analyzed. This dataset contains 50 samples from three species, Iris setosa, Iris versicolor and Iris virginica. By applying Bayesian clustering model selection method, the samples are efficiently clustered into its own category. In the same paper, a data set is simulated with 15 responses and 20 covariates from 4 different normal densities. Additional noisy covariates are generated as well. The purpose was to study whether the covariates will stay important under increasing number of unimportant noisy covariates. In the end, a total number of around 35 covariates is reduced to 15-20 covariates using clustering and Bayesian variable selection(Tadesse et al., 2005). In a gene dataset with a binary response BRCA1 or BRAC2. Bayesian variable selection was used to identify strong significant genes for the classification of BRCA1 or BRCA2. Between 5-27 covariates are selected to be significant, with the most significant ones being keratin 8, TOB1 etc. The method significantly reduced the genes selected comparing with most methods used in other papers, which typically end up more than 55 significant genes (Lee et al., 2003). In the epidemiology study (Walter and Tiemeier, 2009), a regression model was fitted using a conditional logistic model, 17 categorical covariates were used. 2 or 3(depending on the interpretation) were selected to be influential.

The examples above have shown the power of variable selection. Many methods have been developed to find a way to select a subset of ”best” covariates from a larger pool of covariates. Since the rise of computing power in the 1990s, Bayesian variable selection methods have gradually come into popularity because sampling from the posterior distribution became easier than ever. In a Bayesian approach, instead of comparing several models and choose the best one, marginal posterior probability of each covariate that should be in the model is computed, which means we will choose the covariates with the highest posterior probability. More practically speaking, dummy variables are assigned to each of the potential covariates, and the ”best” covariates are those will yield high posterior probabilities of their corresponding dummy variables. Below, we start with MCMC scheme for linear regression and later modify it to accommodate variable selection.

2.1 MCMC for Linear Model

In a regression problem with n observations and p covariates, consider it to be a dataset with very large p. The response Y is often represented as $Y = X\beta + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2 I)$,
$Y \sim N(x^T\beta, \sigma^2 I)$ The unknown parameters of which we want to sample from are β and σ^2. By assigning a prior onto β and σ^2 individually, we will be able combine these priors with the data to produce posterior densities. The regression coefficients

$$\beta_1 \ldots \beta_p$$

are conveniently modeled to follow a multivariate normal distribution. Let $\beta \sim MVN(m, c^2)$, An appropriate distribution for the residual variance to follow is an Inverse Gamma distribution, $\sigma^2 \sim IG(a,b)$. First of all we are to find out the joint posterior of β and σ^2 given the observed data

$$\begin{aligned}
\pi(\beta, \sigma^2|D) &\propto (\sigma^2)^{-n/2} \exp\{-\frac{1}{2}\frac{(y-x\beta)^T(y-x\beta)}{\sigma^2}\}\frac{1}{(c^2)^{p/2}} \\
&\quad \times \exp\{-\frac{1}{2}\frac{(\beta-m)^T(\beta-m)}{c^2}\}(\sigma^2)^{-a-1}\exp-\frac{b}{\sigma^2} \\
&\propto (\sigma^2)^{-n/2-a-1}\exp(-\frac{1}{2}\frac{(y-x\beta)^T(y-x\beta)}{\sigma^2} - \frac{1}{2}\frac{(\beta-m)^T(\beta-m)}{c^2} - \frac{b}{\sigma^2})
\end{aligned}$$

To find out the posterior full conditional densities for them, all we need to do is assuming the parameter other than the one we want to sample from as constant

$$\begin{aligned}
\pi(\beta|\sigma^2,D) &\propto \exp-\frac{1}{2}(\frac{-2\beta^T x^T y+\beta^T x^T x\beta}{\sigma^2}+\frac{\beta^T\beta-2\beta^T m}{c^2}) \\
&= \exp-\frac{1}{2}(\beta^T(\frac{x^T x}{\sigma^2}+\frac{I}{c^2})\beta-2\beta^T(\frac{x^T y}{\sigma^2}+\frac{m}{c^2})) \\
&\propto \exp-\frac{1}{2}(\beta^T(\frac{x^T x}{\sigma^2}+\frac{I}{c^2})\beta-2\beta^T(\frac{x^T y}{\sigma^2}+\frac{m}{c^2})) \\
&\sim N((\frac{x^T x}{\sigma^2}+\frac{I}{c^2})^{-1}(\frac{x^T y}{\sigma^2}+\frac{m}{c^2}),(\frac{x^T x}{\sigma^2}+\frac{I}{c^2})^{-1})
\end{aligned}$$

And,

$$\begin{aligned}
\pi(\sigma^2|\beta,D) &\propto (\sigma^2)^{-n/2-a-1}\exp-\frac{1}{2}(\frac{(y-x\beta)^T(y-x\beta)+2b}{\sigma^2}) \\
&\sim IG(a+\frac{n}{2},\frac{(y-x\beta)^T(y-x\beta)+2b}{2})
\end{aligned}$$

After successfully finding posterior full conditional densities for β and σ^2, all we need to do is to simulate large amount samples from each density by applying Monte Carlo method. And figure 2.1 shows a scatterplot of 2000 simulated samples from full conditional density for β

2.2 Stochastic Search Variable Selection

In the last section, we introduced how to apply Bayesian method in a typical regression problem, where we are discovering the relationship between an observed response variable y and a set of covariates $x_1 \ldots x_p$. The problem arises when, since $\beta_1 \ldots \beta_p$ follows a multivariate normal with mean m and variance $\sigma^2 I_p$, where m is a $px1$ vector and $\sigma^2 I_p$ is a pxp matrix. That gives $p(\beta_j = 0) = 0$ where $j = 0, 1 \ldots p$ because the integral of β equal to any specific point is zero. That means the corresponding x_j has important effect on the response y. By following the same logic for each of the p parameters we would have to conclude that all x_s are important on y. Therefore, variable selection has no more meaning in this case. This problem can be solved in many different ways by assigning a different prior on β. In this paper, we do it by mixing

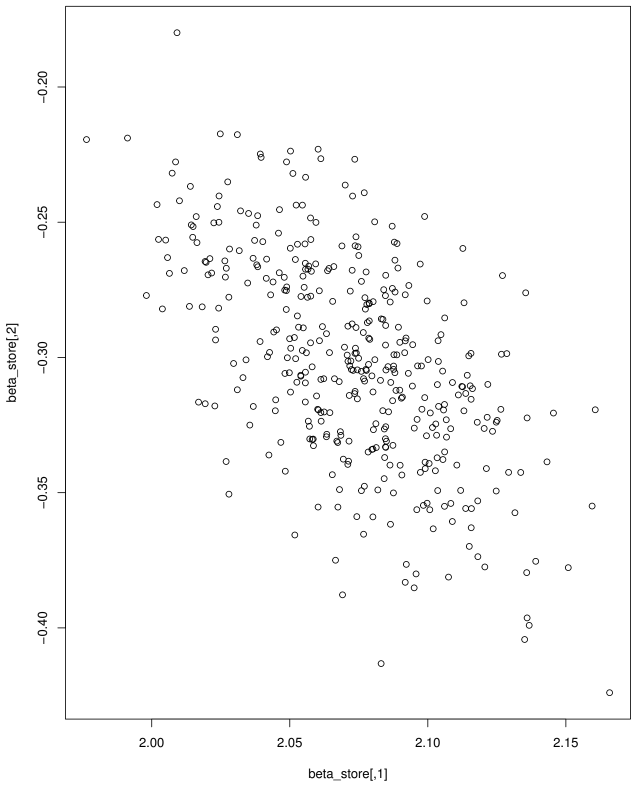

Figure 2.1: Plot of 2000 samples for full conditional density of β

probability measures p into the prior for β, where a new prior for β is formed (Geweke et al., 1996)

$$\pi(\beta_j) = \delta I_{\beta_j=0} + (1-\delta)N(m_j, c^2) = \sum_{j=1}^{j=2} p_j f_j(x)$$

A new prior density for β is introduced in hopes of assigning non-zero probability for it to be insignificant. In this new prior, the indicator has point mass at zero, when β_j is not equal to zero, it has a normal distribution with probability δ. Test this prior when $\beta_j = 0$,
$P(\beta_j = 0) = \delta * 1 + (1-\delta) * 0 = \delta$, which means not every x is important on the response. Goal of assigning non-zero probabilities for β_j is now accomplished.

Bayesian statistics is about discovering the properties of full conditional posterior of each unknown parameters, we did it for ordinary observations. The only difference is now we are doing in a variable selection setting. Suppose we have covariates $\{x_1 \cdots x_p\}$ each with n

observations $\{y_1 \cdots y_n\}$. The joint posterior for β and σ^2 is

$$\pi(\beta,\sigma^2|D) \quad \propto \quad \prod_{i=1}^{n} \pi(y_i|x_i^T\beta,\sigma^2) \prod_{j=1}^{p} \pi(\beta_j)\pi(\sigma^2)$$

The prior for β_j is in the form of a sum, thus making calculating $\prod_{j=1}^{p} \pi(\beta_j)$ extremely difficult. Therefore, dummy variables $z_1 \cdots z_p$ are introduced in hopes of getting rid of the sum, these dummy variables are appropriately set to follow a Bernoulli distribution

$$z_j = \begin{cases} 0 & \text{when } x_j \text{ is not important} \\ 1 & \text{when } x_j \text{ is important} \end{cases}$$

It can also be written in a more compact way

$$f(z) = \prod_{i=1}^{p} (1-\delta)^{z_j} \delta^{1-z_j} \tag{6}$$

Therefore,$z_j = 0$ implies $\beta_j = 0$,$z_j = 1$ implies β_j follows $N(m,c^2)$. Suppose there is one observation y with one covariate x, and $y = \beta x + \varepsilon$ with $\varepsilon \sim N(0,1)$. In the simplest case, to find the conditional posterior for β, we multiply the likelihood of the data and the new prior for β

$$\begin{aligned}
\pi(\beta|D,\sigma^2) &\propto L(D|\beta,\sigma^2)\pi(\beta) \\
&\propto e^{-\dfrac{1}{2}\sum_i (y_i - x_i^T\beta)^2} \prod_{j=1}^{p} [\delta I_{\beta_j=0} + (1-\delta)\frac{1}{\sqrt{2\pi c^2}} e^{-\dfrac{1}{2}\dfrac{(\beta_j - m)^2}{c^2}}] \\
&= \prod_{j=1}^{p} [\delta e^{-\dfrac{1}{2}\sum_i (y_i - x_i^T\beta)^2} I_{\beta_j=0} + (1-\delta) e^{-\dfrac{1}{2}\sum_i (y_i - x_i^T\beta)^2} \frac{1}{\sqrt{2\pi c^2}} e^{-\dfrac{1}{2}\dfrac{(\beta_j - m)^2}{c^2}}] \\
&= \delta_j' I_{\beta_j=0} + (1-\delta_j) N(m_j', {c_j'}^2)
\end{aligned}$$

From here it is obvious to see that the posterior and prior of β have exactly the same form except the details are different, thus it proves the conjugacy in Bayesian statistics.This kind of relationship between posterior and prior applies not just to β, but can be generalized to any parameters. Another interesting property which can be intuitively explained, is that when z is zero, β is also zero and it implies y has mean zero. When z is not zero,y has mean $x\beta$. This means when y is large, there is evidence against y with mean zero, therefore there is evidence against β equal to 0, which means $\pi(\beta = 0)$ is going smaller as y become larger. After some messy computation like we did in the non-regression posterior, we eventually have

$$\delta_j' = \frac{\delta}{\delta + (1-\delta)\kappa};\ (1-\delta_j') = \frac{(1-\delta)\kappa}{\delta + (1-\delta)\kappa},$$

with

$$\kappa = \frac{1}{\sqrt{c^2 A}} \exp[0.5 * (\frac{B^2}{A} - \frac{m^2}{c^2})],$$

where $A = \frac{1}{c^2} + \frac{\sum_i x_{ij}^2}{\sigma^2}$ and $B = \frac{m}{c^2} + \frac{\sum_i x_{ij} y_i^{(-j)}}{\sigma^2}$. We use $y_i^{(-j)}$ to denote the residual for i-th observation after subtracting effect of all covariates except j-th covariate. So, $y_i^{(-j)} = y_i - \beta_0 - \sum_{l \neq j} x_{il} \beta_l$.

The continuous component of posterior mixture distribution of β_j is given by $N(m_j' = \frac{B}{A}, c_j'^2 = \frac{1}{A})$. The posterior density for β_j can therefore be expressed as $\delta_j' I_{\beta_j = 0} + (1-\delta_j') N(m_j', c_j'^2)$ with $\delta_j', (1-\delta_j'), m_j', c_j'^2$ given above.

2.3 Other Methods

SSVS is among many other methods for variable selection. To name a few : Lasso and Ridge regression, where regression coefficients are shrinked to 0 to decrease the variance of estimates. A tuning parameter λ is added to penalize any large number of β. Indicator model selection, where β can either have some effects or no effect at all depending on how big the posterior probability of inclusion suggest; Adaptive Shrinkage, this is when the middleman I_j disappears and we directly choose a prior for $\beta_j | \tau^2$ where τ^2 is the variance for β_j. The word shrinkage is used because the goal is to choose a prior such that the probability of β is reduced to nearly zero when evidence suggests that it shouldn't be in the model; Model space approach, this is an entirely different method than above. In this approach, instead of putting priors on all possible covariates, we only put prior on those covariates that are already selected. Overall, there is no guarantee which model selection method is the best. Each has its own advantages and disadvantages depending on the data given. A lot of times it is useful to apply all variable selection methods and compare the results.

3 Spatial Modeling

Many data are presented in longitudinal and latitudinal forms, these types of geography referenced data are considered spatial data. For example, the most common type of representation of spatial data is heat index map. It gives different levels of heat index by presenting the map in various colors. A lot of times, spatial data maps are very colorful and better looking than typical black and white statistical graphs.

Most of the times a given spatial data set can be classified into one of the three basic types. (Banerjee et al., 2014) First type is point-referenced data, which is often referred to as geocoded or geostatistical data. This is a type of spatial data when the response at one point location $y(s)$ is the random variable. Given a point location s, where s is continuously varying among a bigger region, a value of response corresponds to it. The second type of spatial data is Areal data. This type of data is similar to point-referenced data in the sense that responses are the random variables. The difference is that Areal data responses corresponds to an area A among a bigger region, instead of continuous points. This means, given an area A, the response $y(A)$ specifically correspond to A. The third type of spatial data is point pattern data. Unlike the previous two spatial data types, here the location is the random variable. In other words, given a set of observed responses $y(s)$, we want to know what each response corresponds to which location.(Banerjee et al., 2014) Below we will see some examples for each type of spatial data.

Figure 3.1 is taken from (http://gothos.info/tag/maps/). This is a point-referenced spatial data map. It shows the distribution of all 16700 libraries in the US. The data was collected by the IMLS Public Library Survey in 2009. On the map, it is obvious to see that libraries are extremely concentrated in the Northeast and Midwest. It is not surprising that Northeast has the most concentrated libraries because in addition to the dense population in many Northeastern states, that is where the country started, and therefore culturally it is very rich. Contrary to the popular belief that dense population implies more libraies, the Midwest and the South is identical in terms of population density. However, the density of libraries in the Midwest overwhelms those in the

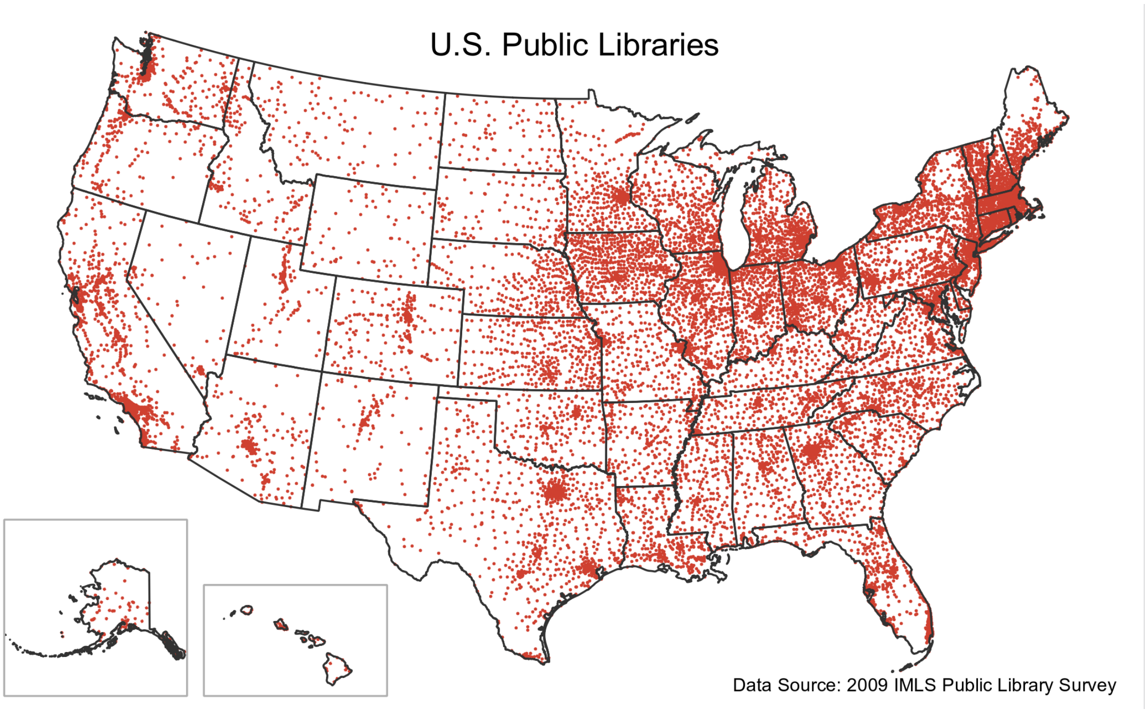

Figure 3.1: Distribution of all 16700 US libraries

South. It is also surprising that in Arkansas, libraries are almost uniformly distributed. We are expecting a skewed distribution where more libraries are concentrated in NWA and Little Rock, which in reality is not the case.

Figure 3.2 is from (https://www.maxmasnick.com/2011/11/15/obesity_by_county/). This is an areal spatial data map. It shows Age-Adjusted obesity rates by US county in 2008. The data was collected by CDC in 2008. On the map, we see a clear pattern that US South as a whole has the most counties with obese people. Overall, the map consolidates the popular belief that poverty leads to obese and wealth leads to a healthy lifestyle by showing that wealthy coastal counties, such as those in Massachusetts, New Jersey, Florida and California, are less influenced by obesity.

Figure 3.3 is taken from (EPA.gov). It was measured in 2007. This is a point referenced data map. The map shows ozone levels at monitoring stations measured across the US. From the map, it is obvious that the mountain region of the US has the high ozone levels compare to other states. The states with large populations such as the Northeast and California have lower background

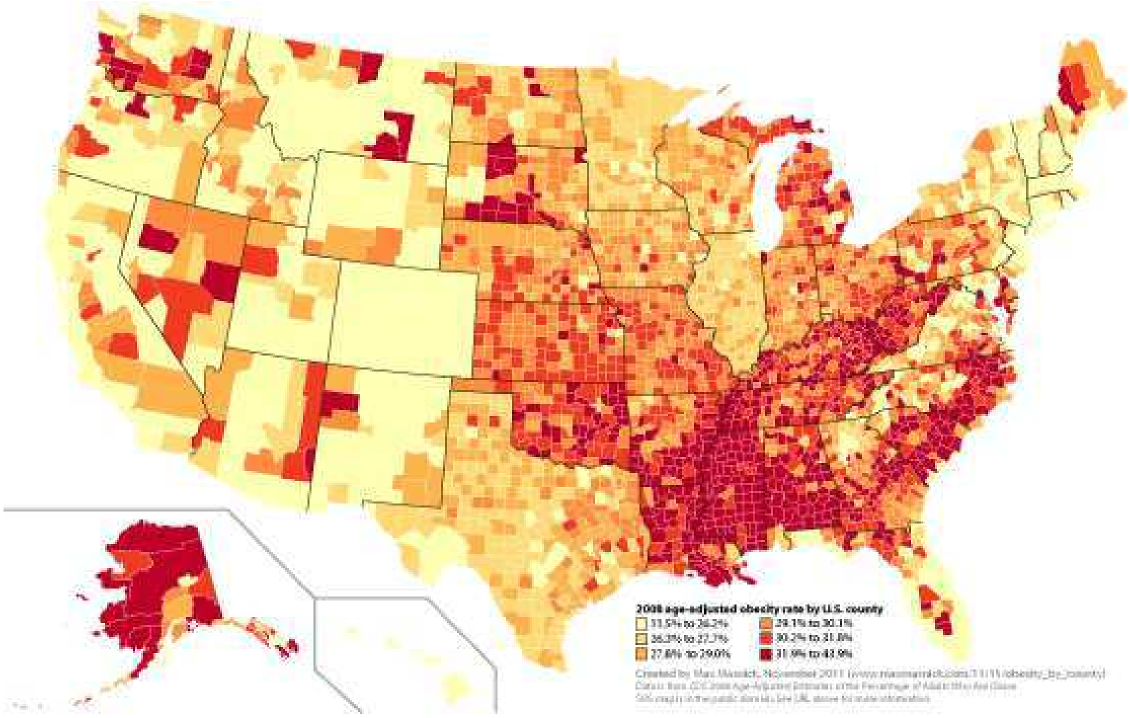
2008 age-adjusted obesity rate by U.S. county
11.5% to 26.2%
26.2% to 27.7%
27.8% to 29.0%
29.1% to 30.1%
30.2% to 31.8%
31.9% to 43.9%
Created by Max Masnick, November 2011 (www.maxmasnick.com/11/11/obesity_by_county)

Figure 3.2: Age-Adjusted obesity rates by US county in 2008

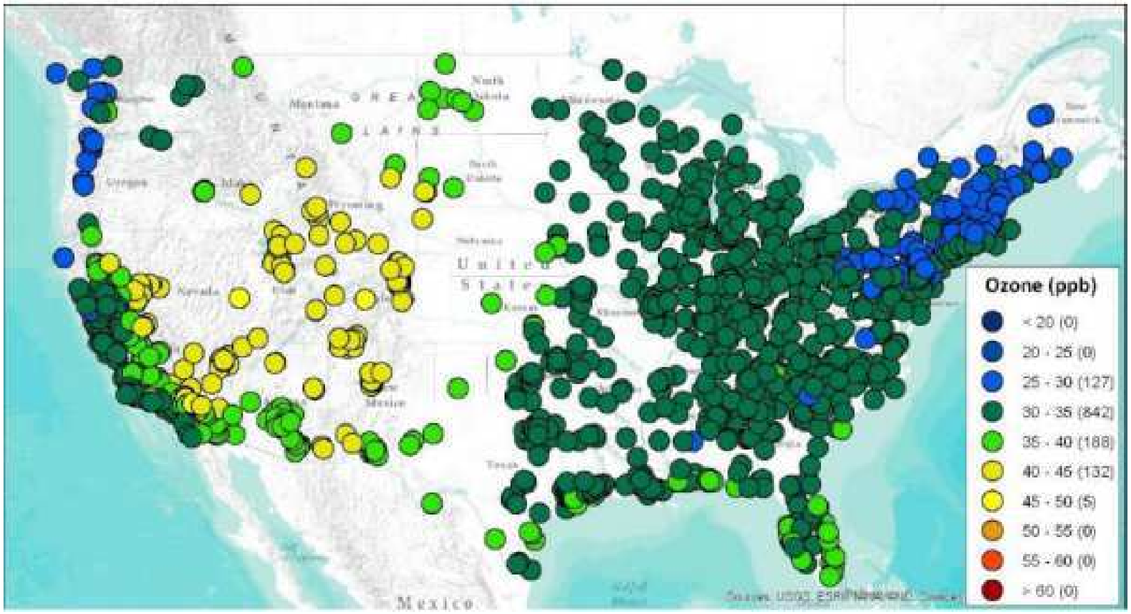

Figure 3.3: Average U.S. background ozone levels at monitoring stations across the U.S. in 2007

ozone levels. Which is not surprising because emissions are higher in those states and therefore the emission disrupt the background ozone level.

When building statistical model on a geography reference data set, standard regression often does not work well because the covariates are inefficient in explaining the response.Therefore a latent variable w is introduced in hopes of adding the information that is missed by using just covariates

$$y(s) = x(s)^T \beta + w(s) + \varepsilon(s)$$

This model resembles much of the standard regression model except a new latent variable $(w(s))$ is added. In this model, s is a symbol for space and $s = \{s_1 \cdots s_n\}$. Where n represents the number of space(locations) where we collected our samples from. $w(s)$ is column vector consists of random spatial effect at each location, $w(s) = \{w(s_1) \cdots w(s_n)\}$ and it is our main interest in spatial modeling. $\varepsilon(s)$ is a column vector of errors at each location s_j, $j = 1 \cdots n$. β is a column vector of covariate effects for all covariates of interest, $\beta = \{\beta_1 \cdots \beta_p\}$. Where p represents the number of covariates in the model. $y(s)$ is a column vector of response at each s_j, $j = 1 \cdots n$. We build our spatial model based on the following assumptions

$$\begin{aligned} \varepsilon(s_i) &\sim N(0,1) \\ w(s_i) &\sim MVN(0_n, \tau^2 r(\theta)) \\ \beta &\sim MVN(m, c^2 I_{p+1}) \\ \tau^2 \sim IG(a,b) & \\ \theta \sim uniform(\theta_{min}, \theta_{max}) & \end{aligned}$$

Recall

$$Corr(w(s_i), w(s_j)) = \frac{Cov(w(s_i), w(s_j))}{\sqrt{Var(w(s_i))Var(w(s_j))}}$$

This implies

$$\begin{aligned}Cov(w(s_i),w(s_j)) &= Corr(w(s_i),w(s_j)) * \sqrt{Var(w(s_i))Var(w(s_j))} \\ &= Corr(w(s_i),w(s_j)) * \tau^2\end{aligned}$$

Since τ^2 is a constant, we are essentially modeling $\exp - \parallel s_i - s_j \parallel \theta$ based on $Corr(w(s_i),w(s_j))$. It can be represented as

$$r(w(s_i),w(s_j)) = \exp - \parallel s_i - s_j \parallel \theta$$

m and $c^2 I_{p+1}$ are the mean and variance matrices for β_s which are given. $r(\theta)$ is a nxn correlation matrix for the spatial random effect variable $w(s)$. Commonly speaking, when the distance between two locations is close, the connections between these two locations will be stronger than when they are further apart. In mathematical terms, it means when $\parallel s_i - s_j \parallel$ is small, $r(\theta)$ is expected to be big. In other words, $r(\theta)_{ij}$ is said to be a decreasing function of $\parallel s_i - s_j \parallel$. θ is a parameter we use to represent this rate of decay. τ is a scalar.

To obtain the posterior of β, σ^2, use Gibbs-sampling

$$\begin{aligned}\pi(\beta|\sigma^2,\tau^2,\theta,D) &\propto \pi(\beta)L(D) \\ &\propto \frac{1}{(\sigma^2)^{n/2}|\Phi|^{1/2}} e^{-\frac{1}{2}\{\frac{(y-x\beta)^T\Phi^{-1}(y-x\beta)}{\sigma^2}+\frac{(\beta-m)^T(\beta-m)}{c^2}\}} \\ &\propto \frac{1}{|\Phi|^{1/2}} \exp -\frac{1}{2}\{\beta^T A\beta - 2\beta^T b\}\end{aligned}$$

Where $A = (c^2 I)^{-1} = \frac{x^T\Phi^{-1}x}{\sigma^2} + \frac{I}{c^2}$, and $b = (c^2 I)^{-1}m = \frac{x^T\Phi^{-1}y}{\sigma^2} + \frac{m}{c^2}$. After computation, the posterior of β is found to follow a normal distribution with mean A^{-1}b and variance A^{-1}

$$\begin{aligned}\pi(\sigma^2|\beta,\tau^2,\theta,D) &\propto \frac{1}{(\sigma^2)^{n/2}|\Phi|^{1/2}}\exp-\frac{1}{2}\{\frac{(y-x\beta)^T\Phi^{-1}(y-x\beta)}{\sigma^2}\}(\sigma^2)^{-a-1}e^{-\dfrac{b}{\sigma^2}}\\ &\propto (\sigma^2)^{-n/2-a-1}\frac{1}{|\Phi|^{1/2}}\exp-\frac{1}{2}\{\frac{(y-x\beta)^T\Phi^{-1}(y-x\beta)+2b}{\sigma^2}\}\end{aligned}$$

It is obvious to see that the posterior of σ^2 is an inverse gamma distribution with shape $a+\frac{n}{2}$, and scale $\frac{1}{2}[(y-x\beta)^T\Phi^{-1}(y-x\beta)+2b]$. Next, to find the posterior of τ, θ, M-H sampling method is preferred

$$\begin{aligned}\pi(\theta|\sigma^2,\tau^2,\beta,D) &\propto \pi(\theta)L(D)\\ &\propto \frac{1}{(\sigma^2)^{n/2}|\Phi|^{1/2}}\exp-\frac{1}{2}\{\frac{(y-x\beta)^T\Phi^{-1}(y-x\beta)}{\sigma^2}\}\frac{1}{\theta_{max}-\theta_{min}}\\ &I_[\theta_{min},\theta_{max}]\end{aligned}$$

At this point, it is very hard to tell which distribution does this posterior come from. Therefore, M-H sampling is to be used. To get started, take one sample from the prior of θ and say it is $\theta^{(0)}$, which is uniformly distributed between $(\theta_{min},\theta_{max})$. Then take a sample from the proposed density of θ and compare it to $\theta^{(0)}$, and θ^{new} can either be $\theta^{(0)}$ or the θ sampled from the proposed density. Do it until 10000 samples of θ_s is reached and use all of them to generate statistics about the true posterior of θ

The posterior of τ^2 is also in non standard form, therefore M-H is needed again to explore the properties of it. By looking at the acceptance ratio

$$P=\frac{\pi(\tau^{2(proposed)})}{\pi(\tau^{2(i)})}*\frac{q(\tau^{2(i)}|\tau^{2(proposed)})}{q(\tau^{2(proposed)}|\tau^{2(i)})}$$

Where $\tau^{2(i)}$ is the current sample of τ^2.Since the prior of τ^2 is restricted on the positive line, therefore it is inappropriate to assume $\tau^{2(proposed)}$ to come from a normal distribution. Let

$log(\tau^{2(proposed)}) \sim N(\tau^{2(i)}, \rho^2)$, then $\tau^{2(proposed)} \sim LogN(log(\tau^{2(i)}), \sigma^2)$. The proposed ratio of densities is equal to $\dfrac{\tau^{2(proposed)}}{\tau^{2(i)}}$ because

$$
\begin{aligned}
q(\tau^{2(i)}|\tau^{2(proposed)}) &\propto \frac{1}{\tau^{2(i)}} \exp -\frac{1}{2}\{\frac{(log(\tau^{2(i)}) - log(\tau^{2(proposed)}))^2}{\sigma^2}\} \\
q(\tau^{2(proposed)}|\tau^{2(i)}) &\propto \frac{1}{\tau^{2(proposed)}} \exp -\frac{1}{2}\{\frac{(log(\tau^{2(proposed)}) - log(\tau^{2(i)}))^2}{\sigma^2}\} \\
\frac{q(\tau^{2(i)}|\tau^{2(proposed)})}{q(\tau^{2(proposed)}|\tau^{2(i)})} &\propto \frac{\tau^{2(proposed)}}{\tau^{2(i)}} \exp -\frac{1}{2} \\
&\{\frac{[log(\tau^{2(i)}) - log(\tau^{2(proposed)})]^2 - [log(\tau^{2(proposed)}) - log(\tau^{2(i)})]^2}{\sigma^2}\} \\
&\propto \frac{\tau^{2(proposed)}}{\tau^{2(i)}}
\end{aligned}
$$

Now, we explore the posterior predictive distribution of the response at new locations. Let y be a vector of observations at locations $s_1 \cdots s_n$, and y^* be new responses that we are trying to predict at locations $s_1^* \cdots s_t^*$. The joint distribution of y, y^* can be written as

$$
\begin{bmatrix} y \\ y^* \end{bmatrix} \sim MVN(\begin{bmatrix} x \\ x^* \end{bmatrix} \beta, \sigma^2 \begin{bmatrix} \Phi & \Phi_c \\ \Phi_c^T & \Phi^* \end{bmatrix})
$$

Therefore, the marginal distribution of y^* given y is,

$$
y^*|y \sim MVN(\mu_{y^*|y}, \Sigma_{y^*|y})
$$

Where

$$
\begin{aligned}
\mu_{y^*|y} &= \mu y^* + \Phi_c^T \Phi^{-1}(y - \mu y) \\
\Sigma_{y^*|y} &= \Phi^* - \Phi_c^T \Phi^{-1} \Phi_c
\end{aligned}
$$

4 Applications

4.1 An example of SSVS with simulated data

Before applying SSVS to real world dataset, an appropriate step to take is to try this method with simulated dataset. This way, we will actually know whether this method is capable of identifying the significance of each covariates without running any risk of getting wrong results for real world dataset. We will run the SSVS method on 4 different datasets to minimize inaccuracy from the results. This verifies our purpose of doing research, which is to explore whether a method works or not under different conditions. To start with, we first simulate 15 covariates independently each from a uniform distribution between 0 and 1. We can select any combination of these 15 covariates to use in our model. In the first model that we use, 6 covariates will be used

$$\begin{aligned}&Y = 2.5 + 0.464X_3 + 0.005X_4 + 0.298X_5 + 0.833X_7 + 0.475X_8 + 0.516X_{13} + \varepsilon,\\&\varepsilon \sim N(0, 0.15^2)\end{aligned}$$

In R, I ran the code with the first dataset to see if the code gives higher posterior probabilities on those covariate effects that are important. I ran the SSVS method with 20000 iterations thinned at every 5th sample to reduce correlation between simulated samples. The code took about 2 minutes and 30 seconds to run, and the output is given in Table 1

Table 1: Posterior Proportion of non-zero covariate effects

0.00025	0.00075	1	0.00050	1
0.00075	1	1	0.00025	0.00025
0.00025	0	1	0.00025	0.00025

In this table, the posterior proportions of non-zero covariate effects(β_s) are given for all 15 covariates. The true model has 6 covariates included, they are $X_3, X_4, X_5, X_7, X_8, X_{13}$. So, in the table, we are expecting to see that for these 6 covariates, the posterior proportion of their corresponding covariate effects is going to be high, and we are also expecting that for those 9 covariates that are not included, their corresponding posterior proportion of non-zero covariate effects is going to be low. After checking them one by one, we see that our SSVS method is doing a efficient job of identifying covariates with significant effects. For example, X_3 is in the true model, and its corresponding posterior proportion of non-zero covariate effect is 1. X_1 is not in the true model, and its corresponding posterior proportion of non-zero covariate effect is almost extremely small at 0.00025. The only thing in the table that doesn't match with the true model is X_4. This is not a big issue because we don't expect our SSVS method to be perfect, moreover, if we look at the coefficient for X_4 in the true model, it is extremely small at 0.005, this accounts for our SSVS method not being able to identify it as significant.

Next, for the ones that match with the true model, we want to know whether the posterior means is a good estimate for the true model coefficients. An appropriate way to set it up is to build a 95 percent credible interval for these posterior means and see whether the true model coefficient is included in the credible interval (Table 2)

Table 2: Point estimate and 95% credible intervals for posterior means of significant β_s

Quantiles	X_3	X_5	X_7	X_8	X_{13}
2.5%	0.44	0.26	0.82	0.46	0.5
50%	0.46	0.28	0.84	0.49	0.52
97.5%	0.49	0.31	0.86	0.51	0.54

We see that the posterior means are a good estimate for the true model coefficient. For example, the true coefficient for X_3 in the model is 0.464, and from the table, we can see that the point estimate for X_3 is 0.46, and 95 percent confidence interval for X_3 is between 0.44 and 0.49, where the true coefficient for X_3 is well included. Also, another example, the true model coefficient for X_7 is 0.833, and from the table, we see that the posterior mean from SSVS method for X_7 is 0.84, and 95 percent confidence interval for this mean is between 0.82 and 0.86, where 0.84 is well included. This proves that our SSVS method is not only capable of identifying the important covariates, but also it can provide an accurate estimate for the true effect of these covariates' impact on this specific model.

Now we present R plot result in figure 4.1

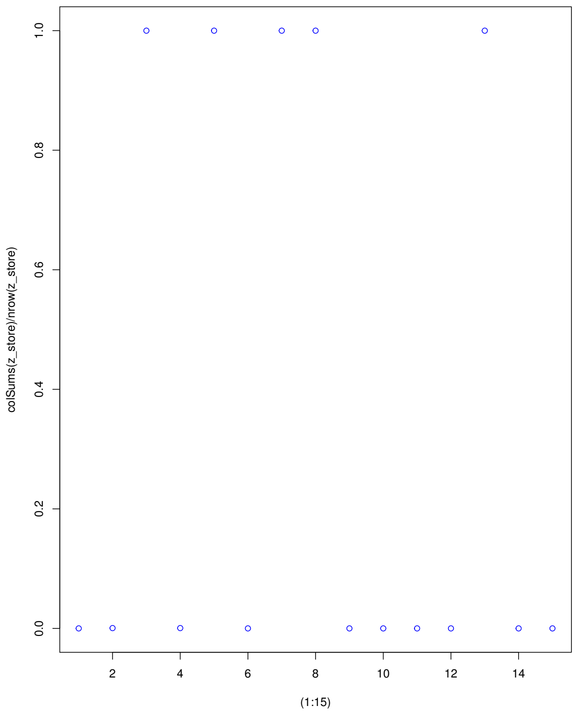

Figure 4.1: Plot of posterior proportions of non-zero β_s

From the plot, it is obvious to see that among the 15 covariates, 5 covariates stand out from the rest by having higher posterior proportions of non-zero covariate effects, they are X_3,X_5,X_7,X_8,X_{13}. Which again proves our findings previously to be right.

This time we run the algorithm on 4th dataset. The second model that we use is identical to the first model, except that we added more noise in the error term. In fact, we have increased the variance of error term 100 times comparing to the first model. Our goal is to see if the increase of noise in the model would have an impact our SSVS method's ability to identify the important covariates appropriately, the model is

$$Y = 2.5 + 0.464X_3 + 0.005X_4 + 0.298X_5 + 0.833X_7 + 0.475X_8 + 0.516X_{13} + \varepsilon,$$

$$\varepsilon \sim N(0, 1.5^2)$$

In R, we do the same thing like we did with the first model. Except we are using a different dataset. We first run the code with the fourth dataset to see if our SSVS method gives higher posterior proportions on those covariate effects that are important. I ran the SSVS method with 20000 iterations thinned at every 5th sample to reduce correlation between simulated samples. The code took about 2 minutes and 30 seconds to run, which is about the same time as the first model, and the proportions are given in Table 3

Table 3: Posterior Proportion of non-zero covariate effects

0.001	0.004	0.2045	0.0025	0.60775
0.00175	1	0.04225	0.00325	0.0005
0.00125	0.001	0.1225	0.00075	0.001

Not surprisingly, our SSVS method is not performing well in identifying the important covariates. We see that the significance of X_7 is still well captured because the posterior proportion of non-zero covariate effects is around 1. For X_5, the same proportion is roughly 0.6. For X_3 and $X_1$3, of which whose proportion should be high, it turns out that the proportion is very low. Therefore, our conclusion is that the ability of our SSVS method to detect significance of covariates is considerably reduced when the noise in the model is high, which is expected because in fact, any model fitting is inappropriate when there is a lot of noise in the model.

Another way we can explore the relationship between the proportions and the model coefficient is that, if the SSVS method works perfectly, it would end up with higher posterior proportions on coefficients whose corresponding covariate has greater true model coefficient. For example, even though X_5 is identified to be important, we observed that the coefficient for X_5 in the true model is smaller than that of X_3. This is strange because despite having higher true model coefficient than X_5, X_5 is selected and X_3 is not selected. This is again showing that SSVS method does not perform efficiently when noise is adding in the model. Another possible research on this topic that we are not going to discuss in this paper, however, which might be interesting, is that the relationship between efficiency of our SSVS method and the amount of noise added to the model. Our assumption is that the noise is added to the model, the less efficient our SSVS method would be.

Next, for the covariates that have posterior proportions that are big enough for us to consider to be significant, we want to know whether the posterior means is a good estimate for the true model coefficients. We will do it the same way as we did for the first model, that is to build a 95 percent credible interval for these posterior sample means of β_s. This time, however, we are not expecting the confidence intervals to accurately contain the true value of model coefficient because we already know that SSVS doesn't work well with noise in the model, the results are given in Table 4

Table 4: 95% credible intervals for posterior means of significant β_s

Quantiles	X_5	X_7
2.5%	0.22	0.54
50%	0.46	0.77
97.5%	0.69	1

We see that the credible intervals do in fact contain the true coefficients. However, they are extremely wide and can be not as informative as the confidence intervals from the previous example. Next we present plot from R in figure 4.2

Then we apply our SSVS method on the second dataset. The model corresponding to this dataset is a non-linear model, our goal is to explore whether SSVS method is capable of identifying significant covariates when the model is non-linear, the full model can be represented as

$$Y = 2.5 + 0.23X_2 + 0.15X_7 - 0.1log(X_15) + \varepsilon, \varepsilon \sim N(0, 0.15^2)$$

In this model, log function is applied to one of the covariates. We perform the SSVS method in R by running 20000 iterations thinned at every 5th sample. We burned the first 10000 samples before thinning. We ended up with 2000 samples for each coefficient of the covariates. The algorithm took about 2 minutes and 10 seconds to run, and the results are given in Table 5

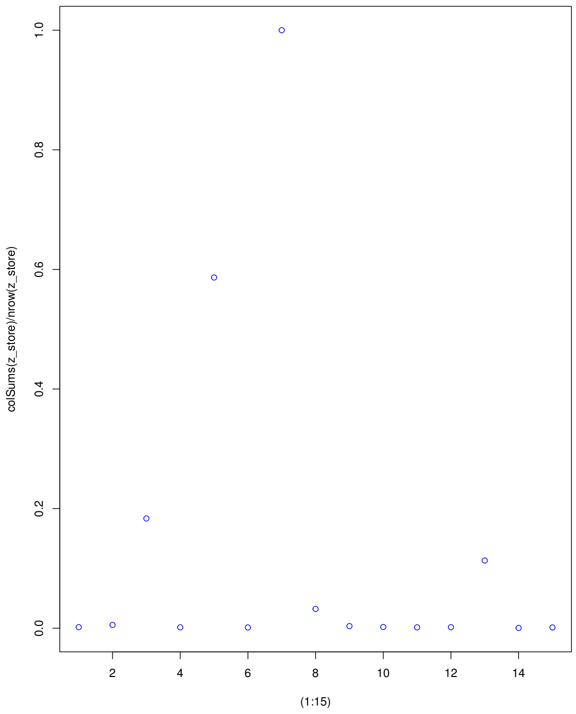

Figure 4.2: Plot of posterior proportions of non-zero β_s

Table 5: Posterior Proportion of non-zero covariate effects

0.000	1	0.00025	0.0000	0.0005
0.00025	1	0.00025	0.00025	0
0.0005	0.0005	0.00025	0.0000	1

By looking at the R output, we observed that X_2, X_7,X_{15} stand out from the rest by having higher posterior proportion of non-zero covariate effects. Then we look back at the original model to check and see if the same covariates were used. It turned out to be a perfect match. We then conclude that nonlinearity in a model doesn't affect SSVS method ability to select the significant covariates.

We then look at the other way like we did in the previous example. That is to explore the relationship between posterior proportions and true model coefficients for selected covariates. We observe that despite both being selected by the algorithm. X_2 has greater true model coefficient than $X_1$5. Then we go to the table and observe that both posterior proportions of the coefficient of these two covariates are 1. Which is not our ideal result. We could further discussing improving the algorithm to make it sensitive enough to show the difference of true model coefficients by giving different values of their corresponding posterior proportions of non-zero covariate effects. We will not discuss it in this paper.

Next, we are going to look at 95% credible intervals for the posterior means to see if they contain the true covariate coefficients, based on the previous conclusion. We incline to think that the credible interval is going to contain the true covariate coefficients,they are given in Table 6

Table 6: 95% credible intervals for posterior means of significant β_s

Quantiles	X_2	X_7	X_{15}
2.5%	0.21	0.12	-0.34
50%	0.23	0.14	-0.32
97.5%	0.26	0.17	-0.29

To check the accuracy of the table, we go back to the original model and look at true model coefficients for these 3 covariates and see if they are included in the confidence intervals. It turns out X_2 and X_7 are matches but X_{15} is not. So we conclude that despite our SSVS method is unaffected when it comes to identify the important covariates when nonlinearity exists, its ability to find the covariate effect for the nonlinear term is affected. The scatterplot is given in figure 4.3

In the R plot, X_2, X_7,X_{15} stand out from the rest just like what happened in the table. This validated our conclusion that SSVS works well in identifying important variables when nonlinearity is present in the model.

Next we apply our SSVS method on the third dataset. The model corresponding to this dataset has an interaction effect between variables X_5 and X_8. Our goal is to explore whether SSVS method is suitable for models with interaction terms, the full model is represented as

$$Y = 2.0 - 0.23X_2 + 0.15X_5 + 1.06X_8 + 0.45X_5X_8 + \varepsilon, \varepsilon \sim N(0, 0.15^2)$$

We ran the SSVS algorithm in R on the third dataset for 30000 iterations and burn the first 10000.

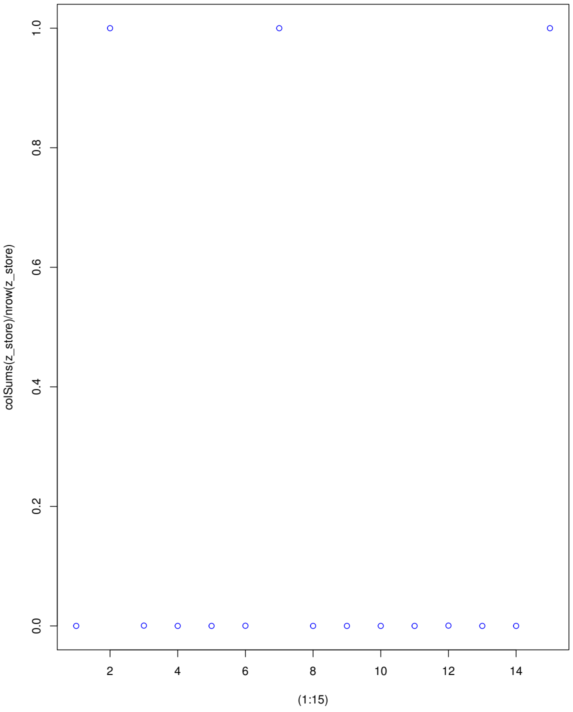

Figure 4.3: Plot of posterior proportions of non-zero β_s

We thinned at every 5th sample after the burning. The algorithm took about 5 minutes to run, which is slightly longer than the previous experiments. One possibility is that the multiplication effect between 2 covariates consumes more time in R. We ended up with 4000 samples for each of the 15 covariate effects. The R output is given in table 7

Table 7: Posterior Proportion of non-zero covariate effects

0.0025	0.0000	0.0000	0.0000	1.0000
0.0000	0.0000	1.0000	0.0000	0.0000
0.000	0.000	0.0000	0.0000	0.0000

We observed that the only high posterior proportion comes from covariate effect for X_5 and X_8. If our SSVS method was efficient, it would identify all covariates used in the model including X_2. This implies that our SSVS method does not perform efficiently when interaction effect is present in the model.

The R plot for chosen covariates given in figure 4.4 From the R plot of posterior proportions

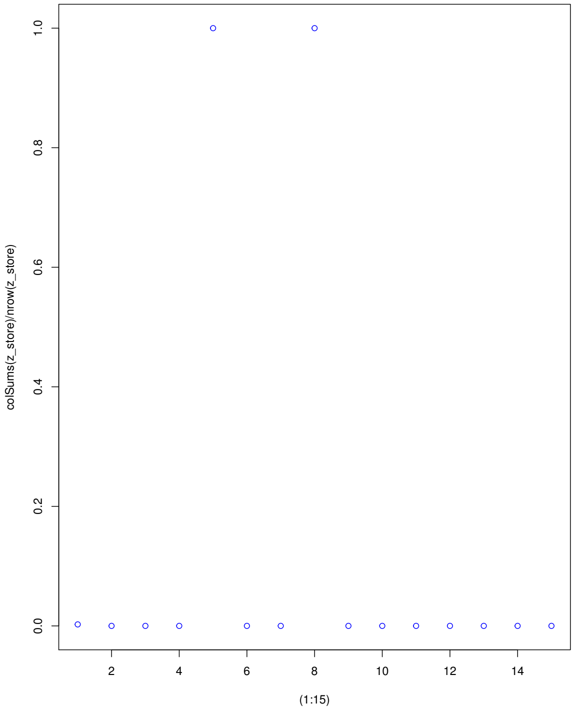

Figure 4.4: Plot of posterior proportions of non-zero β_s

of non-zero covariate effects, we clearly observed only X_5 and X_8 stands out to have high posterior proportion of non-zero β_s, while X_2 missing.

4.2 Real data analysis

4.2.1 Data discription

The spatial dataset that we are using is Protea Atlas dataset. The data is taken from a very biodiverse region in South Africa. It is called Cape Florist Region (CFR). Figure 4.5 displays the map of the region as taken from Chakraborty et al. (2010).

The region is in the most southern end of South Africa and encompasses about 90000 km^2 in area. It is considered to be one of the most biodiverse place in the world. 69 % of species found here exclusive to this region. Thus, with cooperated effort from ecologists in this local region, we

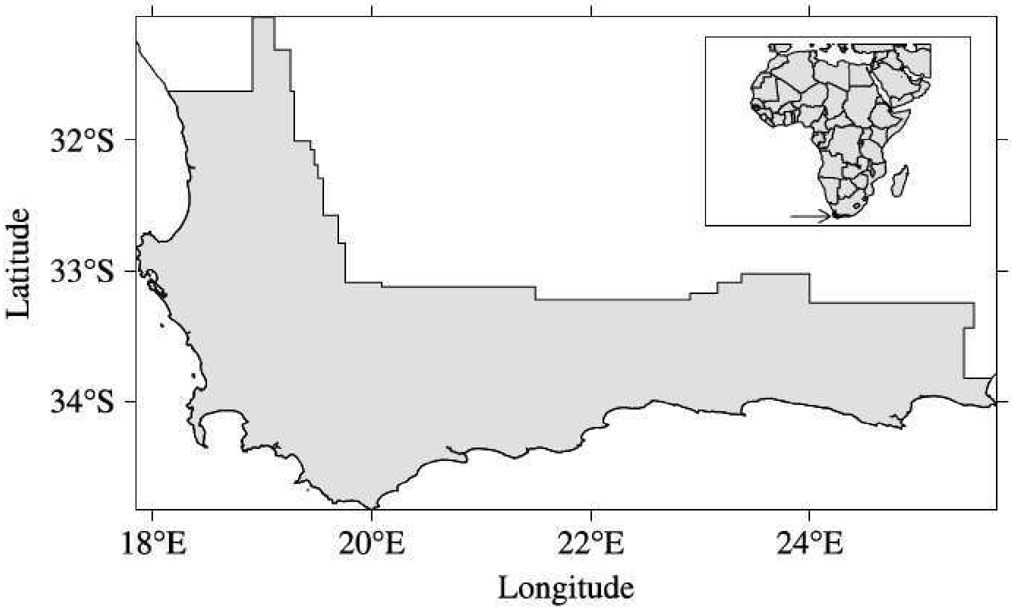

Figure 4.5: Map of Cape Florist Region(In shaded gray)

will be able to discover a lot of opportunities to develop models to identity patterns between biodiversity and environmental factors. For the convenience of statistical analysis, the entire CFR region is uniformed divided into 36,907 1 minute grid cells for the convenience of modeling, we will treat each cell as one observation so our entire dataset will have 36,907 rows. The ecologists have only taken samples from less than a third of total number of cells. That means our training data is very small comparing to the testing data. Due to the high complexity of building a spatial model with too large a number of cells, we will only use 1440 cells in [19°E, 20°E] × [33.6°S, 34°S]. There were 602 cells where the surveyors visited at least once and remaining 838 cells were never visited. So, the goal of the project is to predict at those unsampled cells.

The response is prevalence of *Protea repens*. For each grid cell, its abundance is categorized into 5 categories(0-4). And each category represents a certain number of this species, in ascending order. Table 8 details the categories:

For a cell which was not visited, all five categories of the response are 0. For cells which are visited, but no plants were observed, category 0 will be non-zero, all other 4 categories will be zero. To make this dataset possible for analyzing, we conveniently change this ordinal categorical

Table 8: Categories of response variable

Category	# of Plants
0	0
1	$1-10$
2	$10-100$
3	$100-1000$
4	1000 or more

data into binary data by allocating 1 to any cell where at least 1 site was found in category $1-4$, and allocating 0 if only the 0 category is non-zero. Hence, after this transformation, the data resembles a presence/absence dataset for plants. Hence, the response $y(s)$ will become binary. To make this transition possible, we use the latent variable approach for probit regression from Albert and Chib (1993) and introduce $z(s_i)$ such that $z(s_i)$ is 1 when $y(s_i)$ is positive and $z(s_i)$ is 0 when $y(s_i)$ is negative.

$$z(s_i) = I_{(y(s_i)>0)}$$

We included 18 covariates which reflect environmental and soil related information for that region. The description of these covariates is provided in Table 9

Soil texture level 1 indicates fine soil texture, level 4 indicates coarse soil texture. Soil fertility level 1 indicates low level fertility, level 3 indicates high level fertility. Soil pH level 1 indicates low Alkalinity concentration in soil, level 3 indicates high Alkalinity concentration in soil.

4.2.2 SSVS on real data

We apply the SSVS algorithm on this data set in hopes of selecting a few important covariates out of a large number of covariates. We ran the algorithm with 30000 iterations and burn the first 10000 iterations. We thinned them at everything 5th iteration and we finally ended up with 4000 samples. The code took about 40 minutes to run and the result of the output is given in the Table 10

The table gives posterior proportions of non-zero variable effects for all 18 potential

Table 9: Description of covariates

Variable	Description
$FROST.DURT$	Frost duration
MIN07	Minimum temp in July
HTUNT	Heat units
$MEAN.AN.PR$	Mean annual precipitation
MAX01	Maximum temp in January
NDVI	Enhanced vegetation index
$RAIN.CONCE$	Rainfall concentration
FERT1	Soil fertility level 1
SMDSUM	Summer soil moisture days
FERT2	Soil fertility level 2
SMDWIN	Winter soil moisture days
TEXT1	Soil texture level 1
TEXT3	Soil texture level 3
TEXT4	Soil texture level 4
FERT3	Soil fertility level 3
pH1	Soil pH level 1
pH3	Soil pH level 3
PPTCV	Inner annual coefficient of variation of precipitation

variables. We choose the variables for which the effects were chosen at least 50% of the time. These variables are given in Table 11

Next we want to learn about the point estimates and 95% credible intervals from the posterior distributions of each covariate effect. They are given in Table 12

We first look at the variable $FROST.DURT$, the 50% quantile (median) for its posterior samples is 1, and the same 95% credible interval is between 0.83 and 1.18. So it is appropriate to conclude the variable effect for $FROST.DURT$ to be 1. Following this logic, it is appropriate to set the variable effect for each of the 8 selected variables to be its posterior median of β_s.

Then we generate a scatterplot (figure 4.6) to have a visualization of the selected variables. In the plot, we see that 8 covariate effects stand out from the rest to have higher proportions of non-zero values being selected. And the 7th variable(FERT3) appears to have lower such proportions than the other 7 variables.

Table 10: Posterior Proportion of non-zero covariate effects

PPTCV	FROST.DURT	HTUNT	MAX01	MIN07
0.0095	1.0000	1.0000	1.0000	0.0020
NDVI	RAIN.CONCE	SMDSUM	SMDWIN	FERT1
0.0005	1.0000	1.0000	1.0000	0.0005
FERT3	TEXT1	TEXT3	TEXT4	PH1
0.7295	.00025	0.2260	.01425	1.0000
MEAN.AN.PR	FERT2	PH3		
.00375	0.0010	0.0025		

Table 11: Covariates Selected from SSVS

FROST.DURT	HTUNT	MAX01	RAIN.CONCE
SMDSUM	SMDWIN	FERT3	PH1

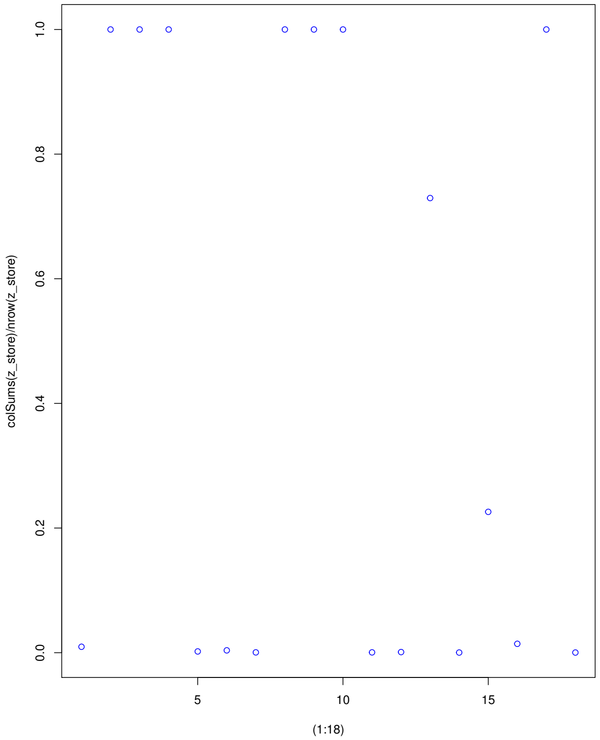

Figure 4.6: Plot of posterior proportions of non-zero β_s

Table 12: 95% credible intervals for posterior means of significant β_s

Quantiles	2.5%	50%	97.5%
$FROST.DURT$	0.8320627	1.0094178	1.1795773
HTUNT	-2.155948	-1.725745	-1.324397
MAX01	0.7902833	1.2747306	1.8328562
$RAIN.CONCE$	-4.523954	-4.228110	-3.932714
SMDSUM	-4.115252	-3.746278	-3.405547
SMDWIN	4.315590	4.620328	4.935613
FERT3	0.1753409	0.3377581	0.4879755
PH1	0.3106712	0.4437057	0.5444190

4.2.3 Spatial Model Building

In the previous section, we applied SSVS method in identifying the important covariates. In this section, we will use these covariates to build a spatial model for this dataset. The goal is for each grid cell, we shall obtain a posterior probability of presence and absence of plants across that region. For the convenience of analysis, we will not take samples from the entire CFR, instead we take a small portion of it. Since the researchers haven't been to all grids cells, we will be able to predict presence/absence of plants for each grid cell based on the posterior probabilities of the responses produced by the spatial model. The general form for this model can be seen in the last section. We run the algorithm for 7000 iterations and burned the first 2000, and we thinned the rest of 5000 samples at every 5th sample, the code took over 6 hours to run. And we ended up with 1000 samples for each of β, τ^2, θ, $w(s)$, $y(s)$. First, we present the posterior summaries of covariate effects in Table 13. We see from the table that SMDWIN(Winter Soil Moisture Days) has the most positive significant impact on the response(Presence/Absence of plants). And HUNT(Heat Unit) has the most negative impact on the response. We will also construct 95% credible intervals for these β_s. The intervals are given in Table 13.

We observed from the results that the 50% quantile(median) of these 8 covariates are not good point estimates for the true model coefficients. For example, the true model coefficient for the covariate $FROST.DURT$ is 7.31, it is included in the credible interval for this covariate,

Table 13: 95% credible intervals for posterior means of 8 covariates

Quantiles	2.5%	50%	97.5%
FROST.DURT	-6.7208304	-0.6486006	13.5733754
HTUNT	-20.371268	-4.879984	1.223927
MAX01	-12.5849263	0.7025249	28.9037163
RAIN.CONCE	-21.468516	-11.704067	-4.809063
SMDSUM	-28.814310	-13.017411	-3.343195
SMDWIN	6.988709	16.180462	38.269132
FERT3	1.470543	5.483995	13.291347
PH1	0.4330384	2.9543393	7.6421285

however, the posterior median of samples for it is -0.648, which has significant difference than the true model coefficient 7.31.

Next, the presence and absence maps(the visual version of average of 1000 simulated samples of response y(s) in each of 1440 grid cells) produced by R are shown in figure 4.7 There is an inverse relationship between probability of presence and probability of absence since $P[presence] = 1 - P[absence]$. We see this relationship clearly from the map when, the black region in the presence map corresponds to the red region in the absence map, and vice versa.

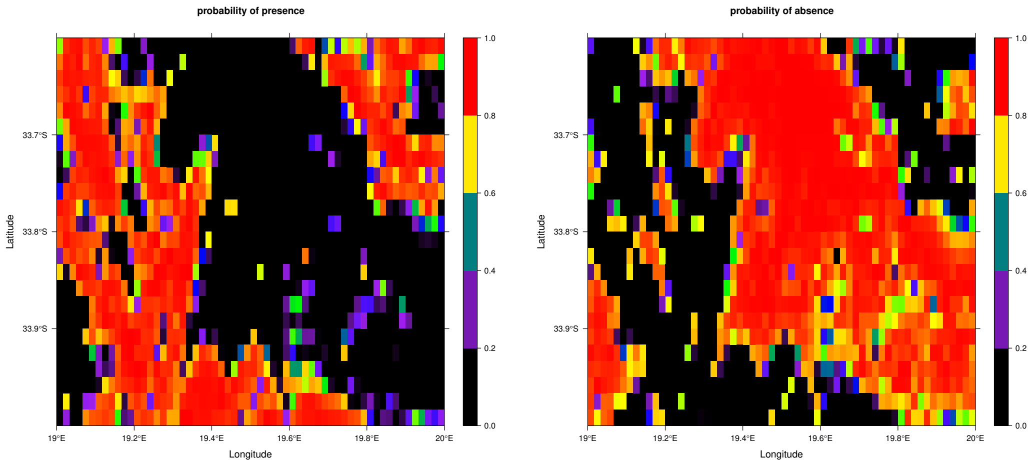

Figure 4.7: Map of probablity of presence (L) and absence (R) of plants in the region

Since we have 1440 grid cells in total, which means $w(s)$ will be a $1440x1$ column vector, we won't list all of them here. Instead , we will look at the spatial effect map generated in R provided in figure 4.8

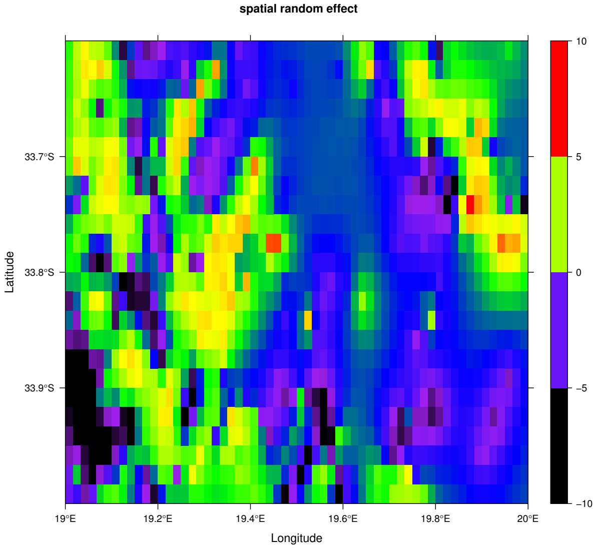

Figure 4.8: Map for spatial random effect

On the map, each grid cell takes the average of 1000 posterior samples of $w(s)$. Since we only have 1440 grid cells, we can still see the rough square patterns on the map. In the region where the spatial effect $(w(s))$ is positive, it implies the probability of presence is observed to be higher than what we could predict only using those 8 covariates. Similarly, in the region where $w(s)$ is negative, probability of absence is observed to be higher than we could predict only using those covariates. The effect for each of the 8 covariates (posterior means of β_s) is given in the following table.

5 Future Work

In this thesis, I described and applied SSVS treating many hyperparameters as fixed. A lot of times we can get more precise results by assigning a probability distribution to these hyperparameters. For example, a lot of literature can be discussed in choosing a and b, the hyperparameters for variance. I will also develop more simulation studies, cover a wider range of models so we can see clearly whether SSVS is efficient or not in identifying the significant covariates under those models. For example, I can increase the variance of the error term and research on how will each unit increase in variance influence the SSVS efficiency. And also, there are many other variable selection methods out there besides SSVS: Shrinkage based methods (Lasso, ridge), Bayesian Lasso, Adaptive Shrinkage etc. I will learn how other variable selection methods work and apply them on this CFR dataset, and many other datasets. Then I will be able to compare under what kind of dataset should a particular method be used. For example, in a high dimensional case, if we use SSVS we would have 2^p ways of choosing the set of regression coefficients. That is very inefficient in terms of computation. Last but not least, I am planning on developing a spatial model that gives more information about the distribution of plants in CFR. That is, instead of transforming the response into a presence and absence (binary) one. I will let the response stay multicategory. That way we will be able to know how the density of plant prevalence varies throughout the region.

References

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

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014), *Hierarchical modeling and analysis for spatial data*, Crc Press.

Chakraborty, A., Gelfand, A. E., Wilson, A. M., Latimer, A. M., and Silander Jr, J. A. (2010), "Modeling large scale species abundance with latent spatial processes," *The Annals of Applied Statistics*, pp. 1403–1429.

Chib, S. and Greenberg, E. (1995), "Understanding the metropolis-hastings algorithm," *The american statistician*, 49, 327–335.

Gelfand, A. E. and Smith, A. F. (1990), "Sampling-based approaches to calculating marginal densities," *Journal of the American statistical association*, 85, 398–409.

Geweke, J. et al. (1996), "Variable selection and model comparison in regression," *Bayesian statistics*, 5, 609–620.

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

Lee, K. E., Sha, N., Dougherty, E. R., Vannucci, M., and Mallick, B. K. (2003), "Gene selection: a Bayesian variable selection approach," *Bioinformatics*, 19, 90–97.

O'Hara, R. B., Sillanpää, M. J., et al. (2009), "A review of Bayesian variable selection methods: what, how and which," *Bayesian analysis*, 4, 85–117.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005), "Bayesian variable selection in clustering high-dimensional data," *Journal of the American Statistical Association*, 100, 602–617.

Tinker, N., Mather, D., Rossnagel, B., Kasha, K., Kleinhofs, A., Hayes, P., Falk, D., Ferguson, T., Shugar, L., Legge, W., et al. (1996), "Regions of the genome that affect agronomic performance in two-row barley," *Crop Science*, 36, 1053–1062.

Walter, S. and Tiemeier, H. (2009), "Variable selection: current practice in epidemiological studies," *European journal of epidemiology*, 24, 733.