Application of Bayesian variable selection methods and shrinkage priors to epidemiological data

Jingyan Fu, School of Medicine

McGill University, Montreal

December, 2022

A thesis submitted to McGill University in fulfillment of the requirements of the degree of

Master in Biostatistics

Abstract

First developed in the 1970s, variable selection plays a vital role in selecting the correct inclusion of variables in a prediction model. From forward selection to penalized regression, a great number of approaches have been constructed within the frequentist framework. More recently, Bayesian variable selection techniques have been developed and gained popularity because of their ability to learn from prior knowledge and construct credible intervals for the parameters without additional computations. However, the application of the Bayesian techniques on data with missingness or spatial information have been limited and the use of customizable programs such as `JAGS` and `RStan` are required.

In this thesis, we provide a comprehensive review of some commonly-used variable selection methods, especially the Bayesian priors, and the comparisons of these methods from past literature. We then apply the regularized Horseshoe prior to two epidemiological datasets to investigate: (1) the random effect of health care regions and social-material deprivation on the treatment decision for patients with aortic stenosis, and (2) the correlations between epidemiological factors and HIV status obtained from a HIV self-testing arm survey. The shrinkage prior is first applied to the AS treatment dataset from the Institut national de santé publique du Québec (INSPQ). Spatial information is contained in postal codes and treated as random effects with 0-1 adjacency matrix. We then analyze the HIV data obtained from a quasi-randomized control trial to explore the risk factors of human immunodeficiency virus (HIV) status. With missing values, we apply a Bayesian hierarchical model for imputation and the shrinkage prior to drop irrelevant predictors. At last, we provide our code and offer ideas for future research.

Abrégé

Datant des années 1970, la sélection de variables joue un rôle important dans le développement de modèles de prédiction. Un grand nombre d'approches ont été développé pour l'approche fréquentiste, tel que la régression pénalisée. Récemment, la sélection de variables sous le paradigme Bayésien a gagné en popularité, notamment dû à sa capacité à apprendre des informations connues à priori et à construire des intervalles de crédibilité sans calculs additionnels. Toutefois, lorsque les données sont corrélées spatialement ou manquantes, l'utilisation de programmes personnalisables tels que JAGS et RSTAN est requise pour analyser.

Dans cette thèse, une revue complète des distributions à priori couramment utilisées sous le paradigme Bayésien pour la sélection de variables est présentée, ainsi que des comparaisons de ces méthodes avec des études précédentes. Par la suite, elles sont appliquées lors d'une étude de cas épidémiologique où seront étudiés (1) l'effet de la région et de la déprivation matérielle et sociale sur la décision relative au traitement de patients atteint de sténose aortique (SA) et (2) les déterminant du statut de VIH dérivé d'une enquête d'autotests. La distribution de rétrécissement à priori est d'abord appliquée sur les données SA rendues disponibles par l'Institut national de santé publique du Québec. L'information spatiale est traitée à l'aide d'un effet aléatoire ayant une matrice de covariance à poids binaire. Subséquemment, les données sur le VIH sont analysées pour explorer les possibles déterminants du statut de VIH. Nous appliquons un modèle hiérarchique bayésien et le priori de rétrécissement afin de substituer les valeurs man-

quantes et d'abandonner les variables impertinentes. Conséquemment, les codes utilisés sont fournis ainsi que quelques idées de recherche future.

Acknowledgements

I want to thank my supervisor, Professor Sahir Bhatnagar, my co-supervisor, Professor James Brophy and Professor Shirin Golchi, for their effort in guiding me through this thesis project. I also would like to thank Professor Alexandra M. Schmidt for reviewing my thesis and offering suggestions for improvement. Thank you for reviewing and commenting on my writing time after time. I could not finish this work without your help.

Besides, I want to thank my mother and my friends who supported me during my whole thesis year. Your love and encouragement keep me from depression and stress.

This project was supported by a training grant from CIHR. This research uses cardiovascular data provided by Institut national de santé publique du Québec (INSPQ) and HIV self-testing data provided by Dr. Nitika Pant Pai in collaboration with the University of Cape Town. I thank Lyne Nadeau for collecting data from INSPQ, Dr. Nitika Pant Pai for authorizing the use of HIV data and Cindy Leung Soo for organizing it.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The idea of variable selection, first proposed in the 1970s, is to select the best subset of variables that can explain the behaviour of a response variable (Lu and Lou, 2021). Especially in analyzing a high-dimensional dataset, variable selection procedures can remove the redundant predictors and select only the relevant ones (Noorie and Afsari, 2020). Besides, even for the low to medium dimensions, many variable selection methods have been proven to successfully reduce the false positives caused by grouped effects or correlations among predictors. Because of its wide application, researchers have proposed numerous methods in recent years, from frequentist to Bayesian approaches.

Though various Bayesian variable selection priors are available, their application to datasets with specific features, such as missing values or spatial information, is unclear because of their complicated prior specification and coding process. The user-friendly packages available today, such as `brms`, impute missing values only through multiple imputations instead of a fully Bayesian imputation approach. Therefore, even with the benefits brought by Bayesian methods, researchers do not have clear guidance on applying them to real problems.

This thesis introduces current variable selection methods, addresses the steps in Bayesian variable selection, and summarizes comparisons made by past researchers between frequentist and Bayesian approaches from hyperparameter settings to model performances. After the summarization, we provide a transparent application process for Bayesian variable selection on real-world data. With an administrative dataset provided by the Institut national de santé publique du Québec (INSPQ) on patients with aortic stenosis, we want to identify the relationships between group-level nonclinical factors and patients' treatment decisions. For the data extracted from an HIV self-testing report submitted by people at risk of HIV infection in South Africa, we intend to identify predictors associated with HIV infection while imputing the missing values in the data.

1.2 Thesis structure

In chapter 2, we will go through some popular frequentist methods, including model selection methods for data in low to medium dimensions and penalization methods like the Ridge and the least absolute shrinkage and selection operator (Lasso) families. Following that, we will describe the different working mechanisms of Bayesian methods, including three types of priors: the spike-and-slab prior, the continuous shrinkage priors and a prior that works as a continuous and discrete mixture. As part of the Bayesian inference process, the Markov chain Monte Carlo, the most crucial computing method in the Bayesian world now, and the convergence diagnosis will be explained in detail. To end this chapter, we will discuss the difference between frequentist and Bayesian approaches in hyperparameter selection and summarize the comparison among variable selection methods performed by Lu and Lou (2021), Celeux et al. (2012) and Piironen and Vehtari (2017).

In Chapter 3, we will investigate the existence of social/economic inequality in the treatment decisions of aortic stenosis patients by modelling the correlation between the social/material status of a patient and his/her treatment received. Because the data are

obtained from administrative data with medical records information, not all predictors we extracted from it will be related to the treatment decisions. Therefore, with the belief that only part of the predictors is valuable, we will perform variable selection in our model. Based on the past literature, we will employ the shrinkage prior for variable selection purposes in real data analysis because of its computing and coding advantage compared to the spike-and-slab and ease of inference on the parameter of interest as a Bayesian method (Lu and Lou, 2021; Piironen and Vehtari, 2017). By adding random effects for a patient's health region, we incorporate the spatial information in the data and will explore whether inequality existed in treatment allocation due to regional differences or social-material levels.

Moreover, with data collected from the self-monitoring reports, we will estimate the correlation between biological information, education, financial situation, sexual behaviour and the HIV status of participants in South Africa in Chapter 4. Since the data contains missing values, we will impute them through a fully Bayesian approach. We will construct a Bayesian hierarchical model that simultaneously imputes missing data and fits the logistic regression model to estimate the log-odds ratios of predictors on HIV status. Similar to the administrative data in the first case, the report contains all possible personal information and behaviour questions that may relate to the spread of HIV. Therefore, among the predictors we collected, we assume the distribution of true predictors is sparse and assign a regularized Horseshoe prior to these parameters (Piironen and Vehtari, 2017). In the end, we will discuss the meaning of each effect based on estimated results.

Finally, in Chapter 5, we will conclude the contents of this thesis, summarize the results obtained from the analysis, and bring up ideas for potential research in the future.

Chapter 2

Literature Review

This chapter goes through some of the commonly-used variable selection methods in the frequentist and Bayesian approaches. We briefly review model selection methods and penalized likelihood methods, followed by more detailed descriptions of the Bayesian approaches. For a small to medium number of predictors, model selection can play the role of variable selection. As the dimension of datasets grows, methods that directly penalize a likelihood were developed and thrived. Unlike the frequentist ways, the Bayesian approach achieves variable selection by assigning certain priors to the regression coefficients with assumed sparsity level. We describe the structure of various priors that perform the variable selection and explain how we obtain the posterior distributions through the Markov chain Monte Carlo. In the last part of this chapter, we discuss the difference between frequentist and Bayesian approaches to selecting hyperparameter values and summarize the comparison of the performance of these methods made by Lu and Lou (2021) and Celeux et al. (2012).

For the following sections, we denote our model as

$$y_i = \alpha + \boldsymbol{X}_i\boldsymbol{\beta} + \epsilon_i$$

for the i^{th} observation, where y_i is the continuous outcome, the predictor $\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,p})$ is a vector of length p, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ a vector of length p corresponding to the effect on outcome both in the domain of $\mathbb{R}^p$, and α the intercept corresponding to the value of y_i

given all predictors in $\boldsymbol{X}_i$ equals zero. The error term $\epsilon_i \sim \mathcal{N}(0, \sigma_0^2)$ for $i = 1, \ldots, n$ and some positive σ_0, where n is the number of observations. For the binary outcome y_i, we assume it follows a Bernoulli distribution and the probability of $y_i = 1$ can be modelled through a logistic regression:

$$y_i \sim \text{Bernoulli}(\pi_i),$$

$$\text{logit}(\pi_i) = \alpha + \boldsymbol{X}_i\boldsymbol{\beta} + \epsilon_i,$$

with $\alpha, \boldsymbol{\beta}, \boldsymbol{X}_i, \epsilon_i$ defined the same as above.

2.1 Traditional methods

For samples with a small to a medium number of predictors, the variable selection problem can be solved through a model selection that exhaustively searches through the combinations of predictors and chooses the best one based on a measure of fit (Heinze et al., 2018). Some model selection algorithms include:

1. Forward Selection (FS): This method is named "forward" since it starts with the most significant parameter by ordering the p-values in each univariate regression model. Repeatedly adding parameters into the model, it keeps the new parameter if its p-value $p_{new} > \alpha_{criteria}$ for certain $\alpha_{criteria}$ corresponds to the desired confidence level until no parameters can be added in.

2. Backward Elimination (BE): Starting by including all predictors in the model, the backward elimination method repeatedly removes the most insignificant independent parameter and re-estimates the model until all of the parameters left in the model significantly contribute to the estimation. Though some statisticians prefer BE to FS, it can be impossible to operate in complex situations (Heinze et al., 2018).

3. Augmented Backward Elimination: Based on BE, the Augmented BE calculates the standardized change-in estimate, compares it to a fixed constant c_0, and removes the parameter if greater than c_0 (Heinze et al., 2018). Change-in-estimate is the change

in β_1, which equals to the change of the regression coefficient of $X_{i,1}$ by removing $X_{i,2}$ from the model $y_i = \alpha + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i$. Compared to BE, it selects a model with more predictors and less bias.

4. Stepwise Selection: The selection is called stepwise forward if it starts with either the null or stepwise backward if it starts with the full model. Stepwise selection iterates the FS and BE alternatively till no more parameters can be added or removed. Though the stepwise selection has computational advantages, it still has high variability and often gives the local optimal solutions rather than global ones (Zou, 2006).

5. Best Subset Selection (BS): The BS approach generates a total of 2^p candidate models for p parameters (Heinze et al., 2018). In the selection process, each candidate model is evaluated through estimation accuracy criteria like AIC, and the model with the lowest score is chosen as the best model (Yuan and Lin, 2006).

   The Akaike Information Criterion (AIC) is given by $-2\log L(\boldsymbol{y}|\hat{\boldsymbol{\beta}}) + 2k$, with $\boldsymbol{y}$ the vector of response variable, $\hat{\boldsymbol{\beta}}$ the estimated regression coefficients, $L(\cdot)$ representing the likelihood function with $\hat{\boldsymbol{\beta}}$ and k is the number of parameters in the model, is proposed to evaluate the "expectation of the cross-validated log-likelihood" in order to penalize the more complex model fit (Heinze et al., 2018). Unlike AIC, the Bayesian Information Criterion (BIC) further penalizes the model by $\log(n)$, where n is the sample size. Because of the additional penalization, BIC selects a model with fewer predictors than AIC.

6. Bayes Factor (BF): Bayes Factor, the ratio of the marginal likelihood for two competing models, calculates evidence in favouring of a model. Given two models labeled $\boldsymbol{\gamma}_s$ and $\boldsymbol{\gamma}_t$, we write the Bayes Factor, $B_{s,t}$, as

$$B_{s,t} = \frac{p(\boldsymbol{y}|\boldsymbol{\gamma}_s)}{p(\boldsymbol{y}|\boldsymbol{\gamma}_t)},$$

   with $p(\boldsymbol{y}|\boldsymbol{\gamma})$ the likelihood. By the Bayes Theorem, we have

$$\frac{p(\boldsymbol{\gamma}_s|\boldsymbol{y})}{p(\boldsymbol{\gamma}_t|\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\boldsymbol{\gamma}_s)}{p(\boldsymbol{y}|\boldsymbol{\gamma}_t)} \frac{P(\boldsymbol{\gamma}_s)}{P(\boldsymbol{\gamma}_t)} = B_{s,t} \frac{P(\boldsymbol{\gamma}_s)}{P(\boldsymbol{\gamma}_t)},$$

called the posterior odds. Different from the p-value, the Bayes factor does not reject any null hypothesis but favours or disinclines the alternatively incorporated subset of variables (Kass and Raftery, 1995). We generally interpret the level in favour of model s by the value of $B_{s,t}$. When $B_{s,t} > 10$, we say that we strongly favour model s (Kass and Raftery, 1995).

Though easily understood and implemented, the traditional variable selection methods are infeasible to compute when the number of predictors is large and impractical for even a moderate number of factors because candidate models grow exponentially. Moreover, they are highly variable because the evaluation criteria, for example, AIC, only approximates the out-of-sample error in prediction but does not measure the actual error (Gelman, 2022).

2.2 Penalized likelihood approaches

In order to apply variable selection in higher dimensions, statisticians developed methods that add penalization to each selected parameter to balance the bias and variance produced. Unlike the model selection methods that exclude undesirable predictors, the penalized approaches shrink these unnecessary ones close to or even equal zero. Because the intercept in penalized likelihood approaches receives no penalization, for simplification, we assume that the outcome variable in the regression model listed in this section is centred and results in an intercept of zero.

2.2.1 The Ridge estimator

Derived from the least squares, the Ridge estimator implements a penalization based on the square of the coefficient's magnitude (NCSS, 2022). By adding a small positive value λ_2 to constrain the model from selecting redundant predictors, the Ridge estimator is

expressed as:

$$\hat{\boldsymbol{\beta}}_{Ridge} = \arg\min_{\boldsymbol{\beta}} \left(\sum_{i=1}^{n} (y_i - \boldsymbol{X}_i\boldsymbol{\beta})^2 + \lambda_2 \boldsymbol{\beta}'\boldsymbol{\beta} \right).$$

With the penalization, the Ridge estimator successfully generates estimates even if the matrix $\boldsymbol{X}'\boldsymbol{X}$ is singular or nearly singular caused by multi-collinearity among predictors (PSU, 2022). However, though it shrinks the estimates toward zero, none of the estimates will equal zero. Therefore, the Ridge estimator cannot produce a parsimonious model because it always includes all the predictors.

2.2.2 The Lasso estimator

Inheriting the idea of penalization, Tibshirani (1996) proposed a method called the Lasso estimator, which applies a penalty term to the absolute magnitude of the coefficients. It is defined as

$$\hat{\boldsymbol{\beta}}_{Lasso} = \arg\min_{\boldsymbol{\beta}} \left(\sum_{i=1}^{n} (y_i - \boldsymbol{X}_i\boldsymbol{\beta})^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 \right)$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$ the standard l_1 norm, which is also called the "L-1" penalty. We have λ_1 the non-negative regularization parameter that controls the amount of shrinkage. It shrinks the coefficients towards 0 as λ_1 increases, while shrinking coefficients to exact 0 if λ_1 is sufficiently large. As an improvement of the traditional methods and Ridge penalization, the Lasso obtains better prediction accuracy and simultaneously produces shrinkage on a continuous scale and variable selection (Zou, 2006; Zou and Hastie, 2005).

However, the Lasso estimator is not always consistent in selection under certain situations. Because it is designed for individual-level variables selection, it often selects more false-positive features (Zou and Hastie, 2005). Moreover, there are three specific scenarios where the Lasso selection is limited:

1. When $p > n$, the Lasso selects no more than n variables before saturation. Furthermore, the performance of the Lasso relies on the bound value on the L-1 norm: λ_1 needs to be smaller than a certain value in order to select more than n predictors (Zou and Hastie, 2005);

2. When the pairwise correlations among a group of variables are high, the Lasso tends to select only one variable from the group, but not all;

3. Even for $n > p$, if predictors are highly correlated, empirical cases showed that the prediction performance of the Lasso is even worse than the Ridge regression (Zou and Hastie, 2005). The performance depends on how the factors are orthonormalized. Different reparametrization will lead to a different set of selected variables (Yuan and Lin, 2006).

Since the introduction of the Lasso to replace the traditional methods, many extensions have been proposed to improve the Lasso estimator's behaviour with various data characteristics. Here we present three popular extensions to the original Lasso estimator.

2.2.3 The adaptive Lasso

To solve the inconsistency of the Lasso estimator, Zou (2006) proposed the "adaptive Lasso", which uses adaptive weights $\hat{w}_j = 1/|\beta_j|^\gamma$ for some $\gamma > 0$, to penalize coefficients differently in the L-1 penalty. The adaptive Lasso has similar algorithm as the Lasso, but uses a choice of λ_n varies with n:

$$\hat{\boldsymbol{\beta}}_{Adaptive} = \arg\min_{\boldsymbol{\beta}} \left(\sum_{i=1}^{n} (y_i - \boldsymbol{X}_i\boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j \, |\beta_j| \right).$$

Compared to the Lasso, the adaptive one ensures consistency in variable selection and makes the estimates asymptotically normal (Zou, 2006). Nevertheless, the adaptive Lasso is not perfect because it cannot provide consistent estimates when $p_n > n \to \infty$.

2.2.4 The elastic net

As another extension of the Lasso, the elastic net is proposed by Zou and Hastie (2005) to overcome the limitations mentioned above. Given some fixed (λ_1, λ_2), the elastic net estimates are:

$$\hat{\boldsymbol{\beta}}_{EN} = \arg\min_{\boldsymbol{\beta}} \left(\sum_{i=1}^{n} (y_i - \boldsymbol{X}_i\boldsymbol{\beta})^2 + \lambda_2 \boldsymbol{\beta}'\boldsymbol{\beta} + \lambda_1 \|\boldsymbol{\beta}\|_1 \right).$$

This estimator significantly improves prediction accuracy in high-dimensional cases and reduces prediction error compared to the Lasso.

Moreover, the elastic net is proven to be successfully applied to correlated data because the regression coefficients of a group of highly-correlated predictors tend to be close. Suppose that $\hat{\boldsymbol{\beta}}_{ENi}\hat{\boldsymbol{\beta}}_{ENj} > 0$ for some j and k given our choice of λ_1, λ_2, we define the difference between the coefficient paths of the j^{th} and k^{th} predictors, $\boldsymbol{X}_{,j} = (X_{1,j}, \ldots, X_{n,j})$ and $\boldsymbol{X}_{,k} = (X_{1,k}, \ldots, X_{n,k})$, as

$$D_{\lambda_1,\lambda_2}(j,k) = \frac{1}{n}\left|\hat{\beta}_{ENj} - \hat{\beta}_{ENk}\right|,$$

where $|\cdot|$ the absolute value. Then $D_{\lambda_1,\lambda_2}(j,k) \leq \frac{1}{\lambda_2}\sqrt{2(1-\rho)}$, with ρ the sample correlation. Therefore, when $\boldsymbol{X}_{,j}, \boldsymbol{X}_{,k}$ are highly correlated, $D_{\lambda_1,\lambda_2}(j,k)$ is close to 0 and results in a grouping effect during the variable selection.

2.2.5 The group Lasso

Suppose each sample has p predictors that can be grouped into J factors. Denoting $\boldsymbol{X}_{,j} = (\boldsymbol{X}_{,1}, \ldots, \boldsymbol{X}_{,p_j})$ an $n \times p_j$ matrix corresponding to the j^{th} factor and $\boldsymbol{\beta}_j$ a vector of length p_j with $\sum_{j=1}^{J} p_j = p$, Yuan and Lin (2006) proposed the group Lasso to overcome the Lasso estimator's dependence on how each predictor $\boldsymbol{X}_{,j}$ is orthonormalized. Instead, the group Lasso selects predictors based on the strength of groups of variables when some of the input predictors are highly correlated.

For a vector $\eta \in \mathbb{R}^d$, $d \geq 1$, and $\boldsymbol{K}$ a symmetric $d \times d$ positive definite matrix, we denote $\|\eta\|_{\boldsymbol{K}} = (\eta'\boldsymbol{K}\eta)^{1/2}$. Given $\boldsymbol{K}_1, \ldots, \boldsymbol{K}_J$ positive definite matrices, we define the group Lasso estimates as

$$\hat{\boldsymbol{\beta}}_{group} = \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{2}\sum_{j=1}^{J}\sum_{i=1}^{n}\left(y_i - \boldsymbol{X}_{i,j}\boldsymbol{\beta}_j\right)^2 + \lambda_g \sum_{j=1}^{J} \left|\boldsymbol{\beta}_j\right|_{\boldsymbol{K}_j} \right),$$

with λ_g the penalty parameter and $|\cdot|_{\boldsymbol{K}_j}$ denotes the $\boldsymbol{K}_j^{th}$ norm.

The group Lasso estimate is reduced to the Lasso when all $\boldsymbol{K}_j = \boldsymbol{I}_{p_j}$, with $\boldsymbol{I}_{p_j}$ the identity matrix of p_j rows. Among the many choices for the kernel matrices $\boldsymbol{K}_j$s, Yuan

and Lin (2006) chose to have $\boldsymbol{K}_j = p_j \boldsymbol{I}_{p_j}$ due to the advantage that group Lasso does not depend on the form of predictors' orthonormalization.

Compared to the ordinary least squares estimate, backward stepwise selections and the Lasso, the group Lasso performs significantly better in terms of mean square error and number of false positives with acceptable computation cost through least angle selection (Efron et al., 2004) under different settings of interactions and co-linearity among the input variable (Yuan and Lin, 2006).

2.3 Bayesian variable selection approaches

In the frequentist approach, it is proven that a penalized estimator provides the best value even under the worst case with sparsity assumptions (Sridharan, 2018). However, as Bayesians, we use posterior distributions to provide a probabilistic measure of uncertainty. We would like to show that the entire posterior distribution concentrates on the optimal values, i.e., the posterior probability assigned to a shrinking neighbourhood of the true parameter value converges to 1 (Bhattacharya et al., 2015). Compared to the frequentist approaches, a Bayesian approach provides inference on the parameter of interest under a single framework.

2.3.1 Spike-and-slab priors

The spike-and-slab priors are used as a major method of Bayesian variable selection, with the prior constructed by two components: a spike concentrated around zero that shrinks minor effects towards zero, and a widespread slab that keeps the large values plausible. After being proposed, it is considered a "gold standard" for sparse Bayesian estimation because of its flexibility in the choice of priors (Agarwal, 2016).

The generic probability distribution function given by the spike-and-slab is in the form of

$$p(\boldsymbol{\beta}|\boldsymbol{\delta}) = \prod_{j=1}^{p} \left[\delta_j \Psi_1(\beta_j) + (1-\delta_j)\Psi_0(\beta_j)\right], \tag{2.1}$$

where the slab $\Psi_1(\cdot)$ represents some prior distribution and the spike $\Psi_0(\cdot)$ is narrow or fixed at zero (Bruinsma, 2019). The parameter vector $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)$ is the indicator variable of either 1 or 0. If β_j large, we have $\delta_j = 1$, representing β_j belongs to the slab, and if β_j equals zero or has very small magnitude, we have $\delta_j = 0$, indicating β_j is allocated to the spike. A common choice for the hierarchical structure of the indicator δ is $p(\delta_j|\omega) = \omega$, with $\omega \sim \text{Beta}(a_\omega, b_\omega)$ for some $a_\omega, b_\omega > 0$ or to have individual inclusion probability ω_j for each δ_j. The inclusion of a parameter is determined by the marginal posterior probability of inclusion (MPPI) of the estimated $\hat{\boldsymbol{\delta}}$. Therefore, variable j is selected if

$$P(\delta_j = 1|\hat{\boldsymbol{\beta}}) \geq 0.5.$$

Since we draw the posterior distribution through a Markov chain Monte Carlo (MCMC), we calculate $\hat{\delta}_j$ for variable j with (Ročková and George, 2015)

$$\hat{\delta}_j = \frac{1}{N}\sum_{k=1}^{N} \delta_j^{(k)},$$

where N stands for the number of iterations after burn-in in MCMC.

Among the numerous spike-and-slab priors modifications and variants, we introduce three methods that have been proved and widely applied.

1. Stochastic search variable selection

In the stochastic search variable selection (SSVS), the spikes and slabs can be seen as a mixture of normal distributions centred at zero (Malsiner-Walli and Wagner, 2011). For each $1 \leq j \leq p$,

$$\beta_j|\delta_j, \tau_j \sim \mathcal{N}(0, \Gamma(\delta_j)\tau_j^2),$$

where τ_j^2 is small but positive. Therefore, we have $\Gamma(\delta_j) = 1$ if $\delta_j = 1$, and $\Gamma(\delta_j) = r$ otherwise; r is the variance ratio between the spike and the slab, with

$$r = \frac{var_{spike}(\boldsymbol{\beta})}{var_{slab}(\boldsymbol{\beta})} = \frac{v_0}{v_1} \ll 1$$

as we denote the variance of the spike as v_0 and that of the slab as v_1.

2. Normal mixture of inverse-gamma

The SSVS is sensitive to the choice of hyperparameters r and $\boldsymbol{\tau}$, which are complex to tune and often data-dependent. As a modified version of SSVS, Ishwaran and Rao (2005) instead put the spike-and-slab prior on the variance of each Normal-distributed predictors, such that

$$\begin{aligned}\beta_j|\eta_j &\sim \mathcal{N}(0, \eta_j),\\ \eta_j|v_0, v_1, a_0, b_0 &\sim (1-\delta_j)\mathrm{IG}(a_0, \tfrac{v_0}{b_0}) + \delta_j \mathrm{IG}(a_0, \tfrac{v_1}{b_0}),\\ \delta_j|\omega_j &\sim \mathrm{Bernoulli}(\omega_j),\end{aligned}$$

is a mixture of inverse-gamma (IG) distributions for each j with some $a_0, b_0 > 0$ and some probability distribution $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)$. Due to the nature of spike-and-slab, we set v_0 small but positive and v_1 close to 1. The advantage of this normal mixture of inverse-gamma prior is to keep the hyperparameters fixed rather than tuning for each dataset.

3. The Expectation-Maximization approach

The stochastic search algorithm in the spike-and-slab has been developed rapidly. Nevertheless, its application on high-dimensional data still experiences difficulties. Ročková and George (2014) proposed the Expectation-Maximization (EM) variable selection (EMVS) to suit the high-dimensional settings, $p > n$, based on an EM algorithm that can find the posterior modes and the candidate models in a fraction of time compared to the stochastic search. Like the stochastic search variable selection process (SSVS), the EMVS introduced the latent indicator variable $\boldsymbol{\delta}$, while for each δ_j from 1 to p, $\delta_j \sim \mathrm{Bernoulli}(\theta_j)$ for some probability distribution $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$.

As described in the spike-and-slab priors, the variance v_0 in the spike distribution can be small but positive or zero in practice. The EM algorithm first applies a sequence of positive v_0 to find possible subsets and set $v_0 = 0$ to evaluate those submodels. The positive v_0 allows a closed-form EM algorithm while providing a tendency for sparser variable selection. For the variance of the slab, v_1, Ročková and George (2014) suggested

two options. It can either be fixed at a large value or treated as random to its prior $p(v_1)$ for a heavy-tailed slab as the double exponential distribution. Like the spike-and-slab variable selection, MPPI determines the inclusion of a variable.

Derived from the original SSVS prior, the EMVS can be extended to other priors with the heavy-tailed slab distributions, such as the Cauchy. The performance of the EMVS with the priors mentioned above heavily depends on the structure of the data. For example, in densely connected networks with sparse predictive variables, the logistic prior better balances the sparsity and group connection.

2.3.2 Shrinkage priors

Though the spike-and-slab prior is straightforward in interpreting variable selection, it has a couple of disadvantages: 1. the result is sensitive to prior choices; 2. the prior is computationally demanding. Even though the cost can be improved with expectation propagation or variational inference, it still requires a substantial increase of analytical work to derive the equations (Piironen and Vehtari, 2017).

A new category of priors, shrinkage priors, is proposed to overcome the shortcomings of the spike-and-slab prior. The continuous shrinkage priors are easy to implement, computationally efficient with tools like Stan, and provide the same or better results compared to the spike-and-slab(Piironen and Vehtari, 2017). Among the various available shrinkage priors, we present the following ones that have been proven with simulations and applied mainly to high-dimensional data analysis.

1. The Laplace (double-exponential) prior

The Laplace prior can be seen as the Bayesian analogue of the Lasso. Park (2008) developed it because of the intention to interpret the Lasso estimates through a designated prior distribution.

Let us denote the distribution of y_i as $y_i \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2)$. Through a Laplace prior conditioned on σ^2, the distribution of $\boldsymbol{\beta}$ can be expressed as (Tibshirani, 1996):

$$p(\boldsymbol{\beta} \mid \sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp(-\lambda \left|\beta_j\right| / \sqrt{\sigma^2}),$$

denoted as the Double Exponential (DE) distribution, with some $\lambda > 0$ to control the amount of penalization. For σ^2, we choose a non-informative scale-variant prior, such as $p(\sigma^2) = 1/\sigma^2$.

We can write the hierarchical representation of the model with Laplace prior as

$$y_i \mid \boldsymbol{X_i}, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2),$$
$$\beta_j \sim \mathrm{DE}(\mu, \sigma/\lambda),$$

for each observation i ($1 \leq i \leq n$) and predictor j ($1 \leq j \leq p$) with some $\sigma^2 > 0$. Similar to the reason mentioned in the penalization methods section, we omit the intercept. As an extension, Park (2008) mentioned the possibility to assign a Gamma prior on λ, i.e., $\lambda \sim \mathrm{Gamma}(a_l, b_l)$ with some $a_l, b_l > 0$, for adjusting the value of λ.

Compared to the Lasso penalization, the Laplace prior automatically provides credible intervals and other estimates from the posterior and obtains more stable estimates (Park, 2008). Though both methods are easy to implement, the Laplace prior is more computationally intensive. However, when we switch to non-linear models, the difference in computation cost between the Lasso penalization and the Bayesian Laplace will be reduced (Park, 2008).

2. The Horseshoe prior

Different from the Laplace prior that uses one parameter λ to determine the shrinkage level of all coefficients, the Horseshoe prior shrinks the $\boldsymbol{\beta}$ through 2 parameters: a global hyperparameter τ that shrinks all the coefficients towards zero, and local hyperparameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$ for j from 1 to p, which follows a heavy-tailed half-Cauchy prior, that allow some β_j to escape the shrinkage (Carvalho et al., 2010). We call this combination the global-local shrinkage. Therefore, the Horseshoe resembles the spike-and-slab with

an infinitely wide slab. Instead of assigning point mass at 0 or 1, the Horseshoe gives continuous priors for λ_j, such that

$$\beta_j \mid \lambda_j, \tau \sim \mathcal{N}(0, \tau^2\lambda_j^2),$$
$$\lambda_j \sim \mathcal{C}^+(0, s_j^2),$$

where $\mathcal{C}^+$ represents the half-Cauchy distribution and $s_1, \ldots, s_p > 0$. Here we assign τ to follow a half-Cauchy distribution with scale equals 1, i.e., $\tau \sim \mathcal{C}^+(0, 1)$. The sparsity in Horseshoe prior is controlled by τ: large τ leads to little shrinkage, while τ close to 0 brings all parameters towards 0.

The effect of the global shrinkage parameter τ can be further illustrated through the following equations. Given $\hat{\beta}_j$ the maximum likelihood solution, we can approximate β_j by $\tilde{\beta}_j$ through

$$\tilde{\beta}_j = (1 - \kappa_j)\hat{\beta}_j,$$
$$\kappa_j = \frac{1}{1+n\sigma^{-2}\tau^2 s_j^2 \lambda_j^2}.$$

Then we can see that, when $\kappa_j = 1$ ($\tau^2\lambda^2$ close to zero), we have complete shrinkage to all $\boldsymbol{\beta}$ and otherwise when $\kappa_j = 0$ ($\tau^2\lambda^2$ large), we apply no shrinkage at all.

However, the Horseshoe prior has several deficiencies. There is a lack of clear explanation on performing inference for τ, and the parameters far away from zero experience no shrinkage. Though the "no shrinkage" behaviour is often considered an advantage, it can be harmful when the data parameters are only weakly identified, especially in logistic regressions. Moreover, due to a Cauchy tail of the Horseshoe prior, the posterior means sampled for the regression coefficients may vanish (Piironen and Vehtari, 2017).

3. The regularized Horseshoe prior

Three changes were proposed to overcome the problems in Horseshoe priors: (Piironen and Vehtari, 2017)

1. introduce a concept of effective non-zero parameters: m_{eff}

2. propose a generalization of the Horseshoe prior that allows specification of shrinkage applied to the coefficients that are far away from zero

3. control the slab width by specifying the maximum effect of β_j we expect to see.

With these changes, Piironen and Vehtari (2017) modified the Horseshoe prior and proposed the regularized Horseshoe prior, which can be recognized as a continuous spike-and-slab prior with finite slab width.

Its hierarchical structure is given as follow

$$\begin{aligned}\beta_j \mid \lambda_j, \tau, c &\sim \mathcal{N}(0, \tau^2 \tilde{\lambda}_j^2), \\ \tilde{\lambda_j}^2 &= \frac{c^2 \lambda_j^2}{c^2 + \lambda_j^2 \tau^2}, \\ \lambda_j &\sim \mathcal{C}^+(0, 1)\end{aligned}$$

with c ≥ 0. When $\tau^2 \lambda^2 \ll c^2$, the regularized Horseshoe approaches the original Horseshoe. When $\tau^2 \lambda^2 \gg c^2$, the coefficient is far from zero and the prior of β_j approaches $\mathcal{N}(0, c^2)$.

We can retain the original Horseshoe by a slight modification on $\tilde{\lambda}_j^2$ as $\tilde{\lambda}_j^2 = \frac{c^2 \lambda_j^2}{\frac{\sigma^2}{n s_j^2} + c^2 + \lambda_j^2 \tau^2}$. However, unless n or the hyperparameter c^2 is very small, the additional term will be small compared to c^2. Therefore, it has little influence on the estimates. For the value of c, a reasonable choice is $c^2 \sim \text{IG}(a_r, b_r)$, with $a_r, b_r > 0$. With this choice, coefficients far from 0 follows a Student-t distribution $\text{t}_v(0, s^2)$, and further prevents the mass of $\tilde{\lambda}_j^2$ from accumulating near zero.

For the global parameter τ, Piironen and Vehtari (2017) and others recommend using a fully Bayesian inference rather than cross-validation. van der Pas et al. (2014) suggested that, assuming the number of true non-zero coefficients exists and is denoted as p_{true}, the optimal choice of τ is $\tau = \frac{p_{true}}{n}$. In general, we can assign τ to be fixed (=τ_0), or to follow a distribution ($\mathcal{N}^+(0, \tau_0^2)$, $\mathcal{C}^+(0, \tau_0^2)$, or $\mathcal{C}^+(0, 1)$), but the last prior will change the prior imposed on m_{eff} according to the choice of σ or n. For τ_0, a recommendation follows $\tau_0 = \frac{p_0}{p - p_0} \frac{\sigma}{n}$, where p_0 is the estimated number of non-zero predictors and usually set to $m_{eff} = \sum_{j=1}^{p} (1 - \kappa_j)$ in model fitting.

2.3.3 Continuous shrinkage and discrete mixture

We say that the spike-and-slab prior is a mixture of discrete functions to evaluate sparseness from a fully probabilistic point of view, and the penalization method optimizes constraints on a continuous scale. In order to benefit from both approaches, another group of prior, called the shrinkage and discrete mixture prior, is developed to incorporate both parts to simultaneously perform variable selection and parameter estimation. One representative of this family is the spike-and-slab Lasso (SSL) prior proposed by Ročková and George (2015).

To construct the SSL, Ročková and George (2015) replaced the slab $\Psi_1(\beta_j)$ for each j in (2.1) by

$$\Psi_1(\beta_j) = \frac{\lambda_1}{2} \exp(-\lambda_1|\beta_j|),$$

with λ_1 small, and the spike $\Psi_0(\beta_j)$ by

$$\Psi_0(\beta_j) = \frac{\lambda_0}{2} \exp(-\lambda_0|\beta_j|),$$

with λ_0 large. Moreover, $p(\boldsymbol{\delta})$ is assigned an exchangeable prior of the form

$$p(\boldsymbol{\delta}|\theta) = \prod_{j=1}^{p} \theta^{\delta_j}(1-\theta)^{1-\delta_j},$$

where $\theta = P(\delta_j = 1|\theta)$ is the expected fraction of large β_j prior to modelling. Finally, $\boldsymbol{\delta}$ is further marginalized such that the SSL prior can be treated as an independent product of the Lasso mixtures, i.e.

$$p(\boldsymbol{\beta}|\theta) = \prod_{j=1}^{p} \theta\Psi_1(\beta_j) + (1-\theta)\Psi_0(\beta_j).$$

Compared to the original spike-and-slab priors and continuous shrinkage methods like the Lasso (Tibshirani, 1996) and Horseshoe (Carvalho et al., 2010), these hierarchical mixtures stand out for producing adaptive posteriors to potential sparsity, performing automatic multiplicity adjustment and achieving Bayes factor consistency in $p \gg n$ cases. The simulation in the paper by Ročková and George (2015) shows that when θ follows a Beta distribution ($\mathrm{Beta}(1, p)$), which is close to the oracle value, the SSL prior can perform variable selection without false positives or false negatives (Ročková and George, 2015). Moreover, with this choice of θ, the model is outstanding in terms of true model discovery,

and its insensitivity to the choice of λ_0, λ_1 makes the variable selection performance even more encouraging.

2.3.4 Computation through MCMC

Computation difficulties for Bayesian methods have existed till late 80's. Thanks to the development of the MCMC, it is feasible to calculate medium to high dimensional integrals that support computing the posterior distribution (Green et al., 2015).

MCMC is a computational method that generates samples $\boldsymbol{\beta}^{(i)}$ in the i^{th} iteration while exploring the state space $\boldsymbol{\beta}$ using a Markov chain, such that it mimics the samples drawn from the target distribution (Andrieu et al., 2003). Specifically, the Markov chain is a stochastic process that the probability distribution of $\boldsymbol{\beta}$ at time i only depends on its distribution at time $(i-1)$ and is determined by an irreducible and aperiodic stochastic transition matrix T, i.e,

$$p\left(\boldsymbol{\beta}^{(i)}|\boldsymbol{\beta}^{(i-1)}, \ldots, \boldsymbol{\beta}^{(1)}\right) = T\left(\boldsymbol{\beta}^{(i)}|\boldsymbol{\beta}^{(i-1)}\right).$$

1. Metropolis-Hastings (MH)

The most popular MCMC implementation is through the Metropolis-Hastings algorithm. Denoting the invariant distribution as $p(\boldsymbol{\beta})$ and the proposal distribution as $q(\boldsymbol{\beta}^*|\boldsymbol{\beta})$, the MH algorithm samples a candidate value $\boldsymbol{\beta}^*$ from $q(\boldsymbol{\beta}^*|\boldsymbol{\beta})$ and then accepts the new estimate with a probability of $A(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = \min(1, [p(\boldsymbol{\beta})q(\boldsymbol{\beta}^*|\boldsymbol{\beta})]^{-1}p(\boldsymbol{\beta}^*)q(\boldsymbol{\beta}|\boldsymbol{\beta}^*))$.

Several samplers are used in the MH algorithm. For example, in the i^{th} iteration, the independent sampler assumes $q(\boldsymbol{\beta}^*|\boldsymbol{\beta}^{(i)}) = q(\boldsymbol{\beta}^*)$, and then the acceptance probability is simplified to

$$A\left(\boldsymbol{\beta}^{(i)}, \boldsymbol{\beta}^*\right) = \min\left(1, \frac{p\left(\boldsymbol{\beta}^*\right) q\left(\boldsymbol{\beta}^{(i)}\right)}{q\left(\boldsymbol{\beta}^*\right) p\left(\boldsymbol{\beta}^{(i)}\right)}\right).$$

As a beneficial property, the MH algorithm does not require normalizing the constant of the target distribution and is easy to simulate several chains in parallel. However, the algorithm's efficientcy largely depends on the choice of proposal distribution: different

choices of the proposal hyperparameters lead to different number of iterations required for convergence.

The Gibbs sampler: A special case of the Metropolis-Hastings

If we define the proposal distribution by the conditional distributions of the joint distribution assuming the conditional distribution in each iteration is tractable, we can rewrite the proposal distribution as $q\left(\boldsymbol{\beta}^*|\boldsymbol{\beta}^{(i)}\right) = p\left(\beta_j^*|\boldsymbol{\beta}_{-j}^{(i)}\right)$. Let us denote the vector of $\boldsymbol{\beta}$ without element β_j as $\boldsymbol{\beta}_{-j} = (\beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots \beta_p)$ for $j = 1, \ldots, p$. When $\boldsymbol{\beta}_{-j}^* = \boldsymbol{\beta}_{-j}^{(i)}$, the acceptance probability equals 1, and we call the sampler with the above assumption a Gibbs sampler. With initial values set as $\beta_1^0, \ldots, \beta_p^0$, the Gibbs sampler samples the variables iteratively in this algorithm: in iteration i,

$$\beta_j^{(i+1)} \sim p\left(\beta_j|\beta_1^{(i+1)}, \beta_2^{(i+1)}, \ldots, \beta_{j-1}^{(i+1)}, \beta_{j+1}^{(i)}, \ldots, \beta_p^{(i)}\right).$$

2. Hamiltonian Monte Carlo (HMC)

As an alternative of the Metropolis-Hastings algorithm, Hamiltonian Monte Carlo has gained its popularity in Bayesian computation for its efficiency in high-dimensional data. It is incorporated in Stan, a programming language that simplifies the inference for Bayesian models, as the sampling method (Green et al., 2015). As a member of the MCMC family, HMC uses the derivatives of the density function being sampled to generate efficient transitions spanning the posterior, to draw the Bayesian posterior (Stan, 2019).

Denote $\boldsymbol{\rho}$ the auxiliary momentum variables, the HMC draws from a joint density

$$p(\boldsymbol{\rho}, \boldsymbol{\beta}) = p(\boldsymbol{\rho}|\boldsymbol{\beta})p(\boldsymbol{\beta}),$$

where the auxiliary density is often chosen to be multivariate normal. i.e. $\boldsymbol{\rho} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}$ set as the identity matrix or estimated from warm-up draws and optionally restricted to a diagonal matrix in Stan.

Similar to the Metropolis-Hastings Algorithm, HMC applies a Metropolis acceptance step. The acceptance probability equals to $\min(1, \exp(\mathbf{H}(\boldsymbol{\rho}, \boldsymbol{\beta}) - \mathbf{H}(\boldsymbol{\rho}^*, \boldsymbol{\beta}^*)))$, if we define $\mathbf{H}(\boldsymbol{\rho}, \boldsymbol{\beta}) = -\log p(\boldsymbol{\rho}, \boldsymbol{\beta})$ and $\boldsymbol{\rho}^*, \boldsymbol{\beta}^*$ the resulting state as the end of each iteration. Com-

pared to the M-H, the HMC is more efficient since the distance between two consecutive steps are larger and the generated points are more likely to be accepted Neal (2012).

The HMC includes three hyperparameters: the discretization time ϵ, the mass matrix $\mathbf{\Sigma}^{-1}$, and the number of steps taken L. The sampling efficiency largely depends on these values. For example, if L is too small, the trajectory traced out in each iteration will be too short, and sampling will devolve to a random walk. When L is too large, the trajectory will be too long and the algorithm will more steps than needed on each iteration (Stan, 2019). Fortunately, Stan offers automatic hyperparameters tuning during the warm-up using the no-U-turn sampling (NUTS) algorithm (Hoffman and Gelman, 2014).

2.3.5 Convergence diagnostics

Convergence diagnosis is a critical step in Bayesian analysis. Here we briefly summarize some statistics that used to determine the convergence of a MCMC.

Statistics based on a single chain

Considering the MCMC chain as a related time series, Geweke (1991) proposed a statistic, denoted as Z_n, to measure the autocorrelation between two averages g_{n_a} and g_{n_b} based on the first n_a and last n_b observations (Roy, 2019). For a MCMC with n iterations, the Geweke statistic is expressed as (Geweke, 1991)

$$Z_n = (g_{n_a} - g_{n_b})\sqrt{\hat{S}_g(0)/n_a + \hat{S}_g(0)/n_b},$$

where $\hat{S}_g(0)$ is the estimated asymptotic variance of the average of all n observations. Suggested by (Geweke, 1991), we can set $n_a = 0.1n$ and $n_b = 0.5n$.

Statistics based on multiple chains

Trace plot, a common graphical method, shows the movement of each Markov chain around its state space (Roy, 2019). By looking at the traceplot, we consider the MCMC

has converged if the traces from multiple chains (at least two chains) mixed well. However, trace convergence sometimes relies on objective opinions and the observation can be exhausting for a large number of parameters.

To quantify the divergence between chains, Gelman and Rubin (1992) proposed the Gelman-Rubin diagnostic, denoted as R to measure whether the chains achieve stationarity by calculating the ratio of between-sequence variance, B/n, versus within-sequence variance W. For a MCMC with n iterations and m chains, these variances are defined as (Gelman and Rubin, 1992)

$$B/n = \frac{1}{m}\sum_{j=1}^{m}(\bar{\psi}_{j\cdot} - \bar{\psi}_{\cdot\cdot})^2,$$
$$W = \frac{1}{m(n-1)}\sum_{j=1}^{m}\sum_{t=1}^{n}(\psi_{jt} - \bar{\psi}_{j\cdot})^2,$$

where ψ_{jt} denotes the t^{th} value of parameter ψ of the n iterations in chain j. The ratio R is estimated by $\hat{R} = \hat{V}/W$, where $\hat{V} = \frac{n-1}{n}W + \frac{B}{n} + \frac{B}{mn}$, and then adjusted by multiplying $\frac{d}{d-2}$, with d the degree of freedom approximated by $\hat{V}/var\hat{V}$. Later, to correct the possibility of negative $d-2$ value, Brooks and Gelman (1998) changed the adjusting factor to $\frac{d+3}{d+1}$.

Based on the Gelman-Rubin diagnostic, Brooks and Gelman (1998) proposed the multivariate form, and labeled it as R_p, estimated by

$$\hat{R}_p = \arg\max_{\boldsymbol{a}} \frac{a'\hat{V}^* a}{a'W^* a},$$

with $\hat{V}^*$ and W^* the covariance matrix form of $\hat{V}$ and W in the univariate measure. When the MCMC reaches convergence, $\hat{R}$ or $\hat{R}_p$ is close to one, and we say the chains are divergent if $\hat{R}$ or $\hat{R}_p$ large. Some commonly-used thresholds are 1.01, 1.05 and 1.1 (Roy, 2019).

2.3.6 Model evaluation and comparison

1. Predictive performance

In order to make comparison between models, we evaluate the predictive performance of the models. For binary, it is common to compute the area under curve (AUC) for

the discrimination ability of the model. AUC measures the area under the Receiver Operating Characteristics (ROC) curve that accounts for the relative change of the true positive rate corresponding to the false positive rate, or vice-versa. In general, the closer the AUC is to 1, the better the model's discrimination. A perfect model, which is idealistic, gives an AUC of 1, while the random guess comes with an AUC of 0.5.

For continuous outcome, the most popular measure is the mean square error (MSE) . MSE calculates the distance between the predictive and the true values (Hyndman and Koehler, 2006). For a sample with sample size n, MSE can be expressed as

$$\text{MSE} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 / n,$$

with $\hat{y}_i$ the predicted value of sample i and y_i the observed value. Different from AUC, RMSE measures from the prospective of calibration.

2. WAIC

Besides the predictive ability, we can compare models based on how well they fit the input data. One of the popular measure of model's goodness of fit is the Watanabe-Akaike Information Criterion (WAIC) (Watanabe, 2010). It estimates the predictive loss through the expression

$$\text{WAIC} = -2 \sum_{i=1}^{n} \left(\log \int p(y_i|\boldsymbol{\beta}) p_{post}(\beta) d\boldsymbol{\beta} - var_{post}(\log p(y_i|\boldsymbol{\beta})) \right)$$

where $\sum_{i=1}^{n} \log \int p(y_i|\boldsymbol{\beta}) p_{post}(\boldsymbol{\beta}) d\boldsymbol{\beta}$ is the log pointwise predictive density and can be approximate by $\sum_{i=1}^{n} \left(\frac{1}{N} \sum_{j=1}^{N} p(y_j|\beta^{(j)}) \right)$, N the number of iterations, through Monte Carlo integration. The component $var_{post}(\log p(y_i|\boldsymbol{\beta}))$ is the variance of the individual log predictive densities, which can be computed by the $V_{j=1}^{N} \log p(y_i|\beta^{(j)})$, with $V_{j=1}^{N}$ the sample variance (Gelman et al., 2014). This calculation can be easily done once a sample from the posterior distribution is available, even when the posterior distribution is not close to Normal (Watanabe, 2018). In model comparisons, the model with smaller WAIC is better and a difference as small as 3 to 5 is considered "significant".

Other than WAIC, we can also compare Bayesian models through the Bayes Factor, which has been introduced in 2.1 for model selection purpose.

2.4 Comparisons between variable selection methods

2.4.1 Hyperparameter tuning

As a key component for all variable selection methods, the values of the hyperparameter in the frequentist approach affect the selection performance of the models. In contrast, hyperparameter values in the priors of Bayesian models affect how fast the MCMC reaches convergence. In frequentist approaches, the amount of penalization is assigned by the researcher and fixed. However, from the Bayesian perspective, we adjust hyperparameter values in the distribution of the shrinkage term. Their values reflect our ideal penalization or belief based on prior knowledge.

Hyperparameter tuning is the process of adjusting hyperparameter values for the best performance using sample splitting. For hyperparameter tuning in the penalization model, we search through the possible hyperparameter space and evaluate the model's performance through cross-validation (Pedregosa et al., 2011). Cross-validation (CV) is the most commonly used method for model evaluation. It is the process of fitting the model with part of the data and evaluating its performance with the data left. Data samples are generally split into a preset number of groups, denoted as k groups. Each time one group of data is taken away from the population for evaluation, with the average of the k performance scores as the final output (Brownlee, 2018). When k equals the sample size n, we call it the leave-one-out CV. Though not preferred for large data because of its computation cost, it presents better prediction results, especially when the model itself is misspecified (Shalizi, 2015). The model with the highest prediction accuracy or the smallest mean squared error will be selected depending on the evaluation criteria.

This tuning method is called "grid search" because it searches through all the possible combinations of hyperparameter values (Brownlee, 2018). For example, in the elastic net method, with two hyperparameters λ_1 and λ_2, the researcher picks a relatively small grid of values of λ_2. For each λ_2, the other parameter (λ_1) is tuned by tenfold CV (Zou and

Hastie, 2005). Finally, the λ_2 giving the smallest CV error is chosen. Even in the $p \gg n$ setting, the computation cost grows linearly as p increases and is manageable.

As Bayesians, we assign values to hyperparameters in each prior based on our belief in the sparsity level. A conservative strategy is to choose a set of hyperparameters to construct a non-informative prior. Instead of searching for the "optimal" values for evaluating the hyperparameter values, Bayesian variable selection methods are often examined through a sensitivity analysis. It evaluates the model's performance concerning the changes in hyperparameter values if our prior belief vastly differs from the predictors' actual sparsity level.

Furthermore, sensitivity analysis, also called the prior robustness diagnostics, has global and local approaches (Roos et al., 2015). Like the grid search, local sensitivity analysis computes the posterior results for all options that fit the prior information, including possible extreme values. The local approach focuses on a smaller scale but calculates the rate of posterior change corresponding to the prior change. Because the global approach is often impractical in applications, Gustafson and Wasserman (1995) recommended applying the local sensitivity analysis on Bayesian models.

Several frameworks of Bayesian robustness diagnostics have been developed. For example, McCulloch (1989) studied the worst-case sensitivity by the principal eigenvalues and Gustafson and Wasserman (1995) designed approaches that vary according to posterior results. A recent formal approach proposed by Roos et al. (2015) handles both the circular and worst-case analysis for complex Bayesian hierarchical models that are formed by an appropriate generated grid for priors, and the values in the grid are used as inputs for computing the marginal posterior density.

Though the formal sensitivity analysis has appreciable theoretical advances, because of its complicated algorithms and obscure practices, many researchers apply the informal approach in real-world cases. Steps of the informal approach include repetitively refitting the model with varied hyperparameter inputs and evaluating the posteriors. For example, when constructing a model with a spike-and-slab prior, we assign different sets

of a_ω, b_ω values to adjust the assumed sparsity and $\boldsymbol{\tau}^2$ values to change the slab width. Similarly, when applying the Horseshoe prior, we try different values of $\boldsymbol{s^2}$ and τ_0 to control the shrinkage level and see the performance of the model under these values. For the Laplace prior, when a Gamma prior is assigned to λ, we adjust the values of a_l, b_l. Otherwise, we can estimate the λ value that optimizes the model's performance through marginal maximum likelihood. We say the prior is robust if we do not observe extreme differences from the posteriors. However, this strategy may be costly in time and not reproducible.

2.4.2 Variable selection performance

Besides the predictive performance and goodness of fit, the performance of variable selection can also be evaluated from the following metrics: the probability of selecting correct variables, which includes the sensitivity and specificity, stability of performance under multi-collinearity, computation efficiency, and coverage, which is the probability that the posteriors' 95% credible intervals cover true effects.

Past researchers summarized the selection ability of both the frequentist and the Bayesian methods through simulations. Celeux et al. (2012) compared the spike-and-slab priors to the frequentist approaches like the Lasso estimator, the elastic net estimator and the most traditional AIC and BIC selection criteria with limited training data. By looking at the specificity, the spike-and-slab priors successfully avoided overfitting. Moreover, they gave better prediction accuracy for sparse data with similar root mean squared error (RMSE) for all methods tested.

In addition, Lu and Lou (2021) compared selected methods under the three variable selection strategies categories: the stepwise selection, the spike-and-slab priors (the SSVS and the normal mixture of inverse-gamma), and the shrinkage priors (the Bayesian Lasso, the Horseshoe and its extension). Two settings were proposed by Lu and Lou (2021): the sparse condition with 30% non-zero covariates and the extremely sparse condition with

10% non-zero covariates. In each condition, the correlation ρ among covariates is set to be 0 (independent covariates), 0.5 and 0.9. The dimension p varies from 20 to 60.

Under the low-dimension setting, all models have similar RMSE. However, the stepwise selection tends to select more zero-effect variables than the Bayesian models. The SSVS is sensitive to multi-collinearity within the spike-and-slab priors if dimensions are large, especially when covariates are highly correlated ($\rho = 0.9$ scenario). Among all scenarios, the shrinkage priors are robust to the change of ρ and require less computation time than the NMIG prior. Therefore, we will apply shrinkage prior to later case analyses because of its interpretability, stability and efficiency. Moreover, among the Laplace, the Horseshoe, and the regularized Horseshoe prior, Carvalho et al. (2010) showed the ability of the Horseshoe prior to avoid undershrinking noise and overshrinking signals over the Laplace in sparse conditions. Furthermore, Piironen and Vehtari (2017) illustrated the benefit of the regularized Horseshoe prior over the original Horseshoe prior through simulations: it gives the most satisfactory performance in recovering the true values of large coefficients while shrinking the irrelevant ones.

Chapter 3

Investigating the correlation between non-clinical factors and aortic stenosis treatments across health regions in Quebec

3.1 Background

Aortic stenosis (AS) is a cardiovascular disease that occurs when the heart's aortic valve narrows (Clinic, 2021a). Patients diagnosed with severe AS need to repair or replace the valve. Surgical aortic valve replacement (SAVR) is the historical and most prevalent treatment plan among treatments. Though the techniques are mature, certain risks like infection and stroke accompany this open heart surgery (Johns Hopkins, 2022). Therefore, an innovative transcatheter aortic valve replacement (TAVR) was developed in 2002 and standardized in 2004 to expand the pool of potential candidates benefiting from this procedure (Cribier, 2012). Instead of performing a conventional surgical procedure, TAVR generally uses a percutaneous approach and potentially decreases morbidity and mortality rates, expanding the number of potential candidates (Clinic, 2021b). However, as

with all advanced cardiac procedures, only a limited number of hospital cardio centers are capable of this costly operation. Compared to the number of the traditional SAVR performed each year, the capacity of TAVR operations all over Canada is still low (Asgar et al., 2019).

In 2010, the Canadian Cardiovascular Society (CCS) began to develop a national quality reporting system, known as the Quality project, to reflect the quality of cardiac care in Canada (Asgar et al., 2019). The quality of TAVR surgeries is included as part of its mission. Despite the expansion of TAVR centers, the committee emphasized the transparency of TAVR allocation with the same effort. With data collected from the National Institute of Public Health, CSS attempts to ensure that the allocation of TAVR surgeries is fair and that patients under particular health conditions have an equal chance to receive the TAVR treatment.

In 2019, Wijeysundera et al. (2019) examined the wait time of TAVR across Canada with a Cox proportional hazard model. This research showed that with increasing TAVR capacity, overall, the average wait time increased from 107 days in 2014 to 135 days in 2016. Regional differences were also observed: patients from Newfoundland waited for 71.5 days, while those from Alberta waited for 213 days.

In addition, a few previous studies exploring the association between non-clinical factors and TAVR treatment decisions have been published in the United States (Damluji et al., 2020; Nathan et al., 2021, 2022). These articles focused on the possible disparities of TAVR and SAVR operation rates determined by the distance to medical resources, family income, ethnicity and age. Published in 2020, Damluji et al. (2020) analyzed patients with severe cardiovascular disease living in the State of Florida from 2011 to 2016. They found that older patients in rural areas have lower TAVR operation rates and need longer travel time to receive TAVR treatment. Even among those who received TAVR, rural patients had higher mortality rates.

Nathan et al. (2021) focused on patients who lived in a metropolitan area where a TAVR center existed between 2012 to 2018 in the US. Controlling for age and clinical

comorbidities, they found that patients with higher family income and economic well-being were more likely to start the TAVR program. In addition, the number of medicare beneficiaries was higher for those patients. Inspired by his findings, Nathan et al. (2022) continued to investigate inequalities in AS treatment decisions caused by race and ethnicity differences. In studying patients in 25 metropolitan areas, Nathan et al. (2022) applied a generalized linear model to conclude that zip code areas with more Black and Hispanic patients have lower TAVR operation rates, as do areas with socioeconomic disadvantages.

Studies in the US show inequalities in income, residence, and ethnicity in the distribution of TAVR operations among patients requiring AS surgeries. However, as opposed to the US mixed private and public health care system, Canada offers universal health care to all its residents, which is supposed to ensure an unbiased treatment decision for all medically deserving patients. Therefore, we propose testing for equities in TAVR accessibility among Quebec patients and investigating any association with non-clinical factors. Bergeron (2019) analyzed the epidemiology of aortic stenosis with Quebec provincial administrative databases from 2002 to 2010. He observed that the incidence rates varied across geographical regions. There were areas with low diagnostic rates but high SAVR rates without prior patient-level epidemiological reasons. Based on these data, it was deemed essential to investigate the possible role of geographical and social/economic factors when exploring possible causal factors for a TAVR decision.

Through a logistic regression model, we explore the potential association between the group/population-level factors and aortic stenosis for patients in Quebec using administrative data. Since the administrative dataset we used is not designed to find the critical predictors of treatment decisions, not all predictors we have are related to the outcome. To identify the important predictors, we apply the regularized Horseshoe prior (Piironen and Vehtari, 2017) for variable selection and combine it with spatial models to include all the information we could obtain from the data. For incorporating the spatial information, we either specify the patients' regions as random effects and assign the variance of these regional effects a weight matrix based on their adjacency or calculate the short-

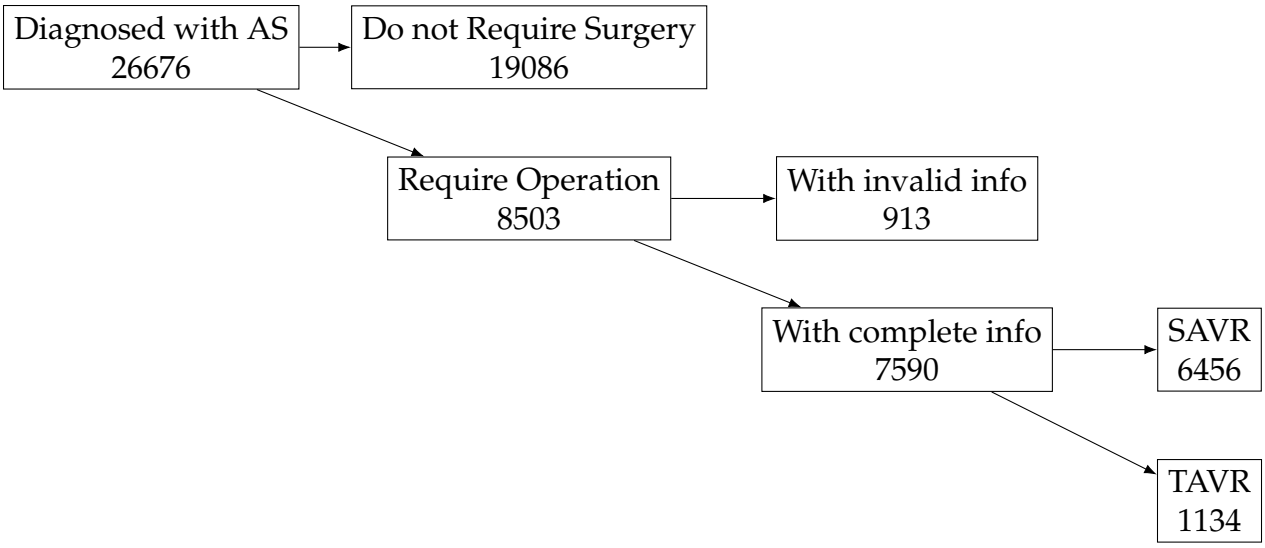

Figure 3.1: The identification process of our desired study cohort: the selection criterion and the corresponding number of Quebec patients from 2011 to 2018

est distance between patients and TAVR centers as an additional predictor. Given the complicated model structure, we estimate the parameters using Bayesian techniques to avoid constructing p-values and calculate the probability of patients receiving the new TAVR procedure with probability priors to reflect our beliefs (Hackenberger, 2019). In this analysis, we adjust the effect of important clinical predictors on treatment decisions. We focus on investigating whether a patient's geographical region and social-economic status determined his/her chance of receiving the more recent TAVR operation.

3.2 Methods

3.2.1 Data

Study cohort

The patients' general information and medical records, including physician billings, diagnosis and hospitalization records, were provided by Institut national de santé publique du Québec (Bergeron, 2019). These files were used to identify Quebec residents involved in

the Quebec Health Insurance Plan aged 45 years or older with a diagnosis of aortic stenosis and undergoing either a SAVR or TAVR procedure between January 1st, 2001 and June 31st, 2018. Patients undergoing AS surgeries were identified using ICD-9-CM (International Classification of Disease, 9th Revision, Clinical Modification) procedure codes and ICD-10-CM (International Classification of Disease, 10th Revision, Clinical Modification) procedure codes (Quan et al., 2005).

We selected patients who had received the treatment since 2011 (8503 in total) due to limited data before this date as the new technology was not yet fully deployed across the province. Seven thousand five hundred ninety patients with complete and valid records for all predictors are selected as our study cohort. For each patient, non-clinical predictors include age, sex, material and social deprivation index in quantile, CLSC, the regional health center most closely associated with the residency area of the individual AS patient, of visit and postal code. The age of the patients is available for each medical record. Compared to the age at AS diagnosis, we believe their age at operation will be more deterministic of the selected AS procedure. Among the available files are all physician visits, hospital visits, and prescription drugs.

Following the Canadian Cardiovascular Society, we consider TAVR programs in hospitals with existing cardiac surgery programs and performed over 10 TAVR procedures over the past calendar year as a TAVR center (Asgar et al., 2019). By the end of 2018, this results in six Quebec TAVR centers, including four centers in Montreal (Sacre Coeur Hospital, McGill University Health Centre, The Centre Hospitalier de l'Université de Montréal, Montreal Heart Institute), one in Quebec City (Quebec Heart and Lung Institute in Quebec) and one in Sherbrooke (Centre Hospitalier Universitaire de Sherbrooke). While other centers have been operating long ago, the Sherbrooke TAVR center has been operating only since March 2014 (Labelle et al., 2020).

Location

Two variables in our dataset contain patient-level spatial information: the postal code with only the first three digits, known as the Foward Sortation Area (FSA), and the CLSC code, which represents the regional health service closet to the patient(CIUSSS and Centres, 2022). Based on the analysis given in previous papers and the variation in CLSC during the treatment years, we choose to locate the patients to their 3-digits postal code.

Because the Quebec Government Health Ministry allocates funding according to Quebec Health regions, we naturally intend to explore the difference in TAVR operation rates among those 17 regions (de la Santé et des Services sociaux, 2018). The rate of each region is presented in Table 3.1. In order to account for regional dependencies, we applied a 0-1 neighborhood structure to account for adjacency, which is explained more in Section.5.2.2.

As the distance from the centroid of one patient's 3-digit postal code address to the closest TAVR center becomes a variable of interest, we also record the latitude and longitude of the six TAVR centers for further measurements.

Social-economic factor

We evaluate the social-economic status of a patient using the social and material deprivation indexes provided in the National Household Survey (Gamache et al., 2019). While social deprivation is defined as the fragility of the social network, from family to community, the NHS describes material deprivation as a lack of access to daily goods and amenities (Pampalon et al., 2012). Both are measured from six perspectives for the population aged 15 years and over:

1. The proportion of the population without a high school diploma or equivalent;
2. The proportion of unemployment;
3. The average income;
4. The proportion of living alone;

5. The proportion of divorced, separated or widowed;

6. The proportion of single-parent families;

The deprivation level is indicated as quantiles from 1 to 5, with 1 the most privileged and 5 the most deprived.

In addition to the score among the six perspectives, both deprivation indexes are computed based on the smallest area units, the dissemination areas, defined from the Canadian censuses (Gamache et al., 2019). Each dissemination area can be linked to specific postal codes and contains residents with relatively homogeneous social-economic conditions, which means that the index of a patient solely depends on his or her address (Pampalon et al., 2012). Besides the two separate deprivation indices included, a combined deprivation index is available with a different suggested way of grouping. In general, metropolitan areas are less deprived of material status but more deprived of social status due to increased social isolation and the varied living conditions of residents.

The measurement and calculation of the deprivation index are performed by INSPQ and transmitted to us for research purposes. In our analysis, we use the quantiles of both deprivation indexes and treat them as factors rather than continuous variables in our logistic regression.

Clinical variables

To measure patients' disease severity, we considered the following parameters: Charlson score, other cardiac diagnoses, number of hospital visits, and number of drugs taken from their first AS diagnosis to TAVR / SAVR surgery. We believe that combining these measurements will give us a good appreciation of the patient's health status at the time of the decision for the aortic valve intervention.

Charlson score, or the Charlson comorbidities, is a measurement tool to evaluate the burden of disease or case-mix administrative data (Quan et al., 2011). It includes 17 co-

morbidities[1] and has been shown to stratify patients according to overall disease severity reliably. To calculate the Charlson score for each patient, we extract the corresponding code of disease from the ICD-9-CM and sum the presence of each comorbidity concerning their coefficient index. Other than the comorbidities mentioned in the Charlson score, the presence of hypertension is also critical for evaluating patients with aortic stenosis.

In addition to the information extracted from the ICD-9-CM records, the number of hospitalization and the number of drugs reflects the general health status of each patient. No matter the length of stay, we count a hospitalization record as one visit and the bill for a new medication as one drug. We do not assign weights to repeated medications in our model.

Invalid values

Invalid values exist in the administrative data transmitted to us. For the CLSC variable and postal code, unidentifiable addresses are found for 54 patients. Though not officially specified, these addresses cluster in military bases or around the border. Invalid quantiles of zero also appear in the social-economic factors. One hundred and one patients have the quantile "zero" in at least one of the deprivation indexes, consisting of 1.2% of the population. According to the explanation provided by INSPQ, the quantile is zero for two possible reasons (Hamel and Gamache, 2020). It is recorded as "zero" as a placeholder because neither the index quantiles were associated with the complete 6-digits postal code nor they were collected from institutions or collective houses. The earlier the operation, the more invalid quantiles we will find in the second situation.

Because we have no available information to impute the invalid values and the proportion of missingness is small, we decide to drop the patients with invalid values and apply the statistical analysis to the complete cases exclusively.

[1]The 17 comorbidities include: myocardial infarction, congestive heart failure, peripheral vascular disease, cerebrovascular disease, dementia, chronic pulmonary disease, rheumatologic disease, peptic ulcer disease, mild liver disease, diabetes without chronic complications, diabetes with chronic complications, hemiplegia or paraplegia, renal disease, any malignancy, moderate or severe liver disease, solid metastatic tumour and AIDS/HIV (Quan et al., 2005).

3.2.2 Statistical analysis

In this section, we introduce three models to estimate the effect of a patient's residence on his/her treatment decision. The first two models use the patient's health care region to represent the spatial effect. Model 1 treats regions as normally distributed random effects, while model 2 further estimates the regional effects of a specific region by adjusting its means and variances based on its neighbouring regions. In the last approach, we calculated the shortest distance from a patient's address to the closest TAVR center and treated it as a continuous predictor.

Baseline model

We denote the total number of patients as N and let $i = 1, \ldots, N$. For patient i, we construct a logistic regression without spatial information as our baseline model for the treatment decision outcome, y_i, as

$$y_i \sim \text{Bernoulli}(\pi_i),$$

$$\text{logit}(\pi_i) = \alpha + \text{age}_i\beta_1 + \text{sex}_i\beta_2 + \text{social}_i\beta_3 + \text{material}_i\beta_4 + \mathbf{clinical}_i\boldsymbol{\beta}_{\text{clinical}}, \tag{3.1}$$

where π_i is the probability for patient i to receive TAVR operation. We use $\text{sex}_i = 1$ to denote female patients, age_i to denote the age of patient centered at 73, $\text{social}_i, \text{material}_i$ to denote the social and material deprivation index in quantiles, and the vector $\mathbf{clinical}_i$ of length 11 to include all the clinical factors like Charlson score, number of emergency visit, hospitalization, diagnosis of other cardiac disease, and number of drugs taken. The intercept α, which follows a standard Normal distribution, $\alpha \sim \mathcal{N}(0, 1)$, represents the baseline log-odds for a 73 years-old male patient in the 3rd quantile for both deprivation index with a Charlon's score of 3.2.

Random effects model

Because the health care region a patient belongs can affect the specialist accessibility and hospital transfer between regions, we add the regions into the baseline model as a ran-

dom effect, denoted as ϕ_k, where $k = 1, \ldots, 17$ representing the region number patient i registered. The model is expressed as:

$$y_i \sim \text{Bernoulli}(\pi_i),$$

$$\text{logit}(\pi_i) = \alpha + \phi_k + \text{age}_i\beta_1 + \text{sex}_i\beta_2 + \text{social}_i\beta_3 + \text{material}_i\beta_4 + \mathbf{clinical}_i\boldsymbol{\beta}_{\text{clinical}}, \quad (3.2)$$

with all notations same as those in Eq. 3.1, except for the spatial effect ϕ_k. We further assume the random effects of all regions follow the same distribution, assign $\phi_k \sim \mathcal{N}(0, \sigma_\phi^2)$ for $k = 1, \ldots, 17$, and set $\sigma_\phi^2 = 1$ for a non-informative prior distribution.

Contiguous neighbours model

Region	Population	TAVR rate
Bas-Saint-Laurent	262	0.176
Saguenay-Lac-Saint-Jean	230	0.113
Capitale-Nationale	552	0.255
Mauricie	198	0.162
Estrie	674	0.139
Montreal	1231	0.219
Outaouais	33	0.091
Abitibi-Temiscamingue	100	0.100
Cote-Nord	74	0.162
Nord-du-Quebec	14	0.071
Gaspesie-lles-de-la-Madeleine	129	0.147
Chaudiere-Appalaches	300	0.230
Laval	300	0.163
Lanaudiere	382	0.139
Laurentides	438	0.121
Monteregie	1020	0.116
Centre-du-Quebec	209	0.158

Table 3.1: Distribution of patients in 17 health care regions, recorded from 2011 to June, 2018

Based on the random effects model, we intend to study the possible correlation between regions based on their locations. In general, residents in neighbouring regions have similar opportunities to receive TAVR interventions, as shown in Table 3.1. The Outaouais district has an exceptionally negative effect because residents needing TAVR

from Outaouais are not limited to receiving medical care in Quebec. Those patients who received aortic stenosis treatments in TAVR centers in Ottawa, Ontario, are not captured by the Quebec Health Database. Based on this observation, we model the spatial information with a binary adjacency weight matrix (Banerjee et al., 2015).

Based on the FSA, for all patients registered in the health care region k, we assumed they share the same regional ϕ_k, and the logistic regression model is the same as described in Eq. 3.2. We assumed that the difference of regional effects between neighbour regions would be smaller, where region k and k^* were neighbours if any part of the two borders is connected, marked as $\omega_{k,k^*} = 1$. Otherwise, we have $\omega_{k,k^*} = 0$. Therefore, given a region k, the total number of neighbours it has is computed as $d_k = \sum_{k^*=1,k^*\neq k}^{17} \omega_{k,k^*}$.

We assign ϕ_k a prior distribution of

$$\phi_k \sim \mathcal{N}\left(\frac{\sum_{k^*=1,k^*\neq k}^{17} \omega_{k,k^*}\phi_{k^*}}{d_k}, \frac{\sigma^2}{d_k}\right),$$

with some $\sigma^2 > 0$. Here we assign $\sigma^2 = 1$ for a non-informative choice.

Continuous distance model

In addition to modelling the regional difference as random effects, inspired by the research of Damluji et al. (2020), we alternatively interpreted the spatial information by measuring the distance between a patient and TAVR centers. The model can be described as

$$\begin{aligned} y_i &\sim \text{Bernoulli}(\pi_i), \\ \text{logit}(\pi_i) &= \alpha + \text{distance}_i\beta_{dis} + \text{age}_i\beta_1 + \text{sex}_i\beta_2 + \text{social}_i\beta_3 + \text{material}_i\beta_4 + \mathbf{clinical}_i\boldsymbol{\beta}_{\text{clinical}}, \end{aligned} \tag{3.3}$$

with α, β_1 to $\boldsymbol{\beta}_{\text{clinical}}$ the same as described in model 1. Furthermore, instead of measuring the regional effects, we use a continuous predictor distance_i to represent the log of the distance between the centroid of the patient i's address and the nearest TAVR center and let β_{dis} be the effect coefficient corresponding to log-distance. Thus, if patient i lived close to TAVR hospitals, distance_i is small, with the log-odds of 1 unit increase in the log of distance equals β_{dis}.

Shrinkage prior

Because the AS data are administrative, we extract as many predictors as possible and try to identify the relationships between some predictors and the AS treatment decision. Thus, given that not all predictors we extracted from the dataset are helpful, we add the Bayesian shrinkage prior to our proposed spatial models to select the truly important predictors among all available information. In addition, the social and material deprivation indexes relate to a patient's location; the correlation among Charlson score, hospitalization, drug intake, and cardiac disease diagnosis is positive. From the comparisons in Lu and Lou (2021), the shrinkage priors help to distinguish the individual effect of each predictor when there exist correlations among predictors. As discussed and showed through simulations in Piironen and Vehtari (2017), the regularized Horseshoe prior reaches convergence the quickest and generates the most consistent estimation for variable selection purposes compared to the Laplace and the Horseshoe prior. Therefore, we apply the regularized Horseshoe prior to filter out the redundant predictors and to reduce the bias caused by collinearity.

With notations of the distributions and hyperparameters the same as explained in Chapter 2.3.2, the prior we put on all βs in Eq. 3.1, Eq. 3.2 and Eq. 3.3 can be expressed as:

$$\beta_j \sim \mathcal{N}(0, \tau^2 \tilde{\lambda}_j^2)$$
$$\tilde{\lambda}_j^2 = \frac{c^2 \lambda_j^2}{c^2 + c^2 \lambda_j^2}$$

where

$$\lambda_j \sim \mathcal{C}^+(0, 1),$$
$$c^2 \sim \text{IG}(\tfrac{\nu}{2}, \tfrac{\nu}{2} s^2),$$
$$\tau \sim \mathcal{C}^+(0, \tau_0).$$

The idea of the regularized Horseshoe prior is to shrink all $\boldsymbol{\beta}$s to 0 by a global parameter τ, which follows a half-Cauchy distribution, while allowing some of $\boldsymbol{\beta}$ to escape from the shrinkage by adding a local parameter $\tilde{\boldsymbol{\lambda}}$. For parameter j, when $\tilde{\lambda}_j$ small, β_j concentrated

around 0, and when $\tilde{\lambda}_j$ large, β_j is estimated with large values. The additional level of $\boldsymbol{\lambda}$ and c^2 bound the maximum magnitude of the estimated $\boldsymbol{\beta}$ and prevent the large values from going to infinity. According to the recommendations made by Piironen and Vehtari (2017), the values of the hyperparameters are set by $\nu = 20$, $s^2 = 4$ and $\tau_0 = 0.001$.

3.2.3 Model selection

All models are implemented through Hamiltonian Monte Carlo in `RStan` with four chains of 3000 iterations each, and the adaptive delta equals 0.95 as recommended in the user manual (Stan, 2019).

We compare the performance of the models above using the Brier score to evaluate model's prediction performance. Measuring from the prospective of calibration, Brier score calculates the distance between the predictive and the true values (Goldstein-Greenwood, 2021). It can be expressed as

$$Brier = \sqrt{\sum_{i=1}^{N}(\hat{\pi}_i - y_i)^2/N},$$

the square root of the averaged squared difference between the predicted probability of receiving TAVR and the actual operation performed for each patient. The predicted probability of patient i to receive TAVR is denoted as $\hat{\pi}_i$, and the actual operation patient i received is denoted as y_i.

In addition to the prediction performance, we use the widely applicable information criteria (WAIC), which estimates the predictive loss, to compare the goodness of fit for the models. In general, the model with a smaller WAIC should be selected because it preserves more information and thus has better performance. We also present the time that each model takes to complete 3000 iterations for reference.

After checking the convergence of all models through the reported $\hat{R}$ scores and traceplots of all estimated parameters, we summarize the values of these different criteria in Table 3.2. We found that the contiguous neighbour model has smaller Brier score and WAIC while finishing the MCMC with moderate computation cost. Therefore, we con-

clude that the region where a patient lives gives a better representation of this AS dataset's spatial information. Hence, we study the estimates obtained from the contiguous neighbour model and make further analysis.

	Computation Time(s)	Brier Score	WAIC(SE)
Baseline Model	8078	0.0932	4729.5(108.4)
Random Effects Model	9078	0.0917	4710.6(103.3)
Contiguous Neighbours Model	9232	0.0917	4704.0(102.5)
Continuous Distance Model	8114	0.0926	4727.6(102.7)

Table 3.2: Comparison of spatial model performance with the regularized Horseshoe prior

3.3 Results

3.3.1 Descriptive analysis

Characteristics	Mean (SD)	Percentage (N)
Sex	/	33.6% (2250)
Age at Surgery	73.1 (8.8)	/
Wait time (day)	356.0 (831.6)	/
Distance (km)	80.2 (134.9)	/
Charlson Score	3.2 (2.8)	/
Hospitalization	1.7 (1.4)	/
Drugs Intake	10.3 (16.1)	/
TAVR Operation Rate	/	18.1% (1374)

Table 3.3: Baseline characteristics of patients

Among the 7590 patients who received AS operations from January 1st, 2011, to June 30th, 2018, 5042 were male, 2548 were female, and 18.1% patients were selected for TAVR operations. Based on the means and standard deviations of the predictors in Table 3.3, the population largely varies in social and economic status and experiences different levels of disease severity. On average, a patient received an AS operation one year after their first diagnosis and 91.4% of the patients went through hospitalization during their wait

	Quantile	Population	TAVR/k	Wait Time	Age	Charlson	Female(%)
Social	1st	1396	136	379.7 (843.4)	72.6 (8.7)	3.03 (2.76)	0.322
	2nd	1550	145	382.3 (874.4)	72.8 (8.6)	3.20 (2.78)	0.292
	3rd	1616	147	420.3 (918.0)	73.3 (8.8)	3.14 (2.81)	0.302
	4th	1578	161	405.8 (908.2)	73.9 (8.6)	3.30 (2.83)	0.350
	5th	1450	157	399.0 (867.1)	72.8 (9.3)	3.32 (2.82)	0.417
Material	1st	1262	159	346.4 (822.8)	74.3 (8.6)	2.87 (2.67)	0.330
	2nd	1378	158	415.2 (908.8)	73.6 (8.6)	3.13 (2.84)	0.304
	3rd	1607	156	426.1 (924.7)	73.3 (8.4)	3.36 (2.81)	0.330
	4th	1686	135	403.6 (882.4)	72.5 (8.9)	3.21 (2.78)	0.354
	5th	1657	144	389.9 (869.3)	72.3 (9.4)	3.34 (2.85)	0.353

Table 3.4: Patients' characteristics grouped by their social-material deprivation index. Includes summary of the TAVR operation rate followed by their wait time, age, Charlson's score and sex. Continuous variables are presented as mean(sd).

time. Females (18.1%) had a higher proportion of TAVR operations compared to males (13.3%), with shorter waiting times on average (333.9 days for females compared to 367.1 for males). In general, female patients are older (74.1 vs 72.6 years old), more deprived in social (45% in quantile 4 or 5 vs 37% in quantile 4 or 5) and material status (24% in quantile 4 or 5 vs 22% in quantile 4 or 5), and have slightly more severe health conditions with higher Charlson scores (3.52 vs 3.10) and more drugs intake (11.7 vs 9.7).

Without adjusting for other factors, patients who are privileged in material and deprived of social status are more likely to receive TAVR treatment even if they are healthier and younger. As analyzed in Section 5.2.2, these areas with higher TAVR rates are likely to be close to the metropolitan area, which means that these patients are close to the TAVR centers.

3.3.2 Estimates from the model

We then use the contiguous neighbour model described in Section 3.2 to analyze the effect of each predictor on the type of valve replacement operation. With four chains running in parallel through `Rstan`, the model takes 9232 seconds to complete. We checked the $\hat{R}$ values and traceplots to ensure the model converged. The traceplots for two predictors,

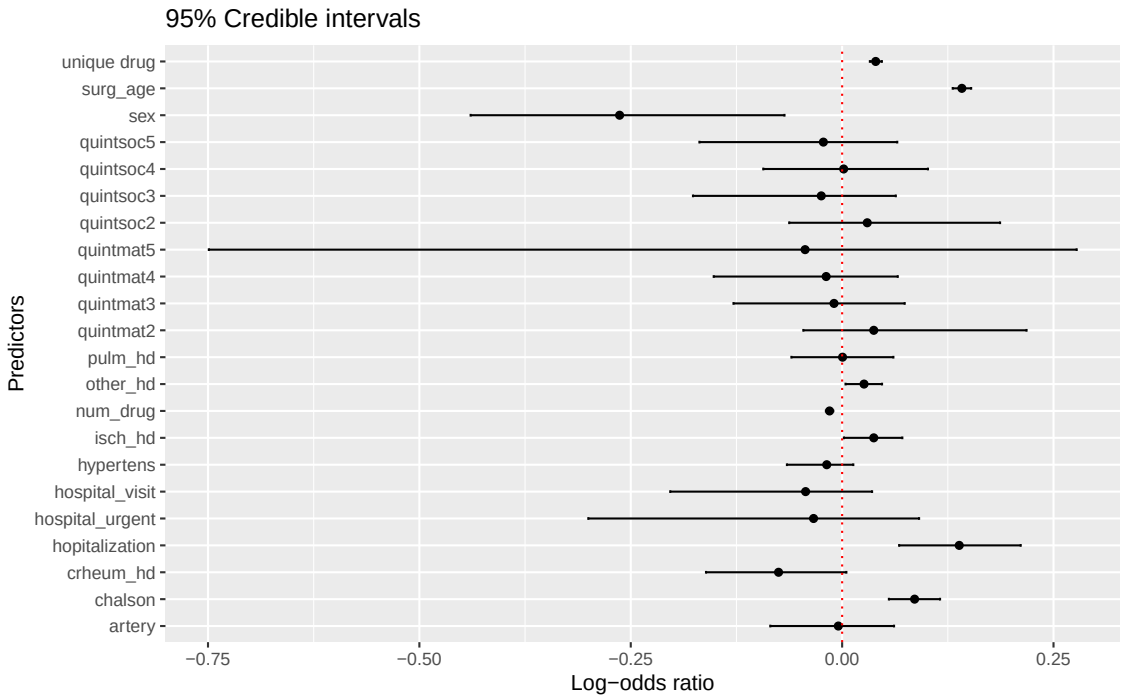

Figure 3.2: Posterior means and 95% credible intervals of the predictors' log-odds ratio sampled from the contiguous neighbors model.

age at surgery and **coronary heart disease** are presented in Figure 3.7 in Appendix 3.6. We obtain the mean and 95% credible intervals base on the posterior samples drawn from the Hamiltonian Monte Carlo, presented in Figure 3.2.

Based on estimation, clinical factors play a significant role in treatment decisions, as we hypothesized. Patients with higher Charlson scores are selected for TAVR surgeries because of their health conditions. For each new drug a patient intakes, the odds of TAVR increase by 3.3%. From the first AS diagnosis to operation, the odds of patients selected for TAVR increase by 7.0% for each hospitalization record. Our estimates on other factors match the observation obtained from the descriptive analysis. Older patients receive TAVR, with an odds of 1.140 times per year younger. Adjusting for age, sex and clinical variables, fewer male patients receive TAVR. The odds for a male are 0.644 times that for a female.

The estimated regional effect is shown in Figure 3.3 for the remote regions in the north and Figure 3.4 for the more concentrated regions in the south of Quebec. The plots clearly

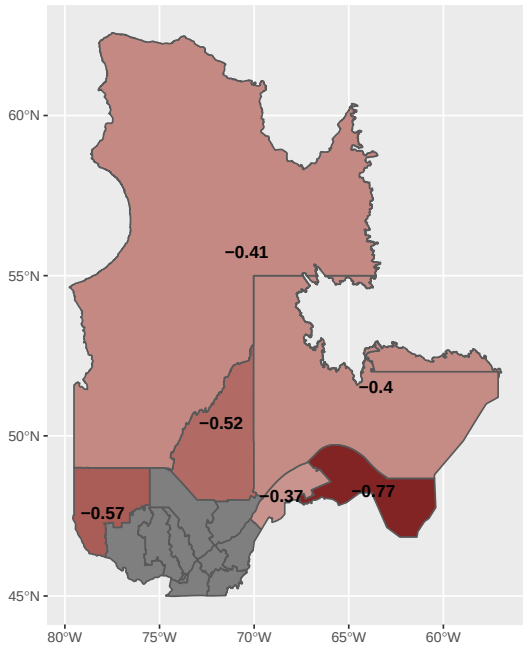

Figure 3.3: Posterior mean of spatial effect for the remote regions. A region in red has a negative log-odds ratio, representing a lower odds and lower probability for patients in this region to receive TAVR. The spatial effects of regions in grey are plotted in Figure 3.4.

show that the patients in the metropolitan area, especially Quebec City and Montreal, are more inclined to be selected for the TAVR operation and less favourable for the remote regions. The result offers an interesting finding: while patients from Outaouais have options to receive TAVR in hospitals in Ontario, compared to those living in Montreal and Quebec City, patients in the surrounding areas are significantly less likely to have TAVR, even compared to the patients living in very remote areas. Distance to the medical centers alone no longer explains this phenomenon. However, these data cannot adequately investigate other possibilities, including unequal distribution of medical resources among healthcare regions assigned by the province. The exact effect sizes and their 95% CI can be found in Table 3.5 in Appendix 3.6.

Compared to the sex and patients' locations, social and material status have minor effects on the treatment decision. The most positive log-odds ratio of these effects is 0.037 for patients in the second material quantile, and the most negative effect is -0.044

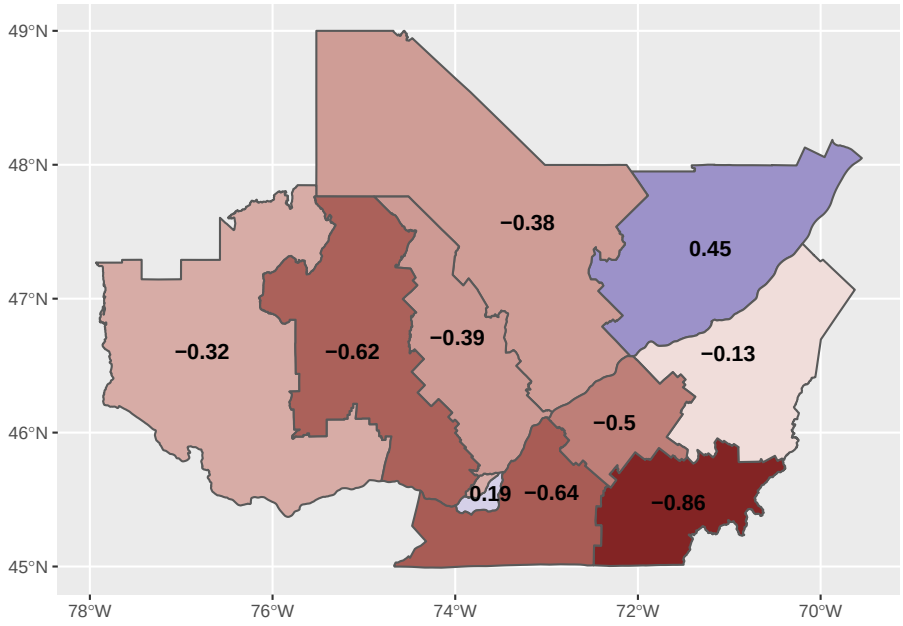

Figure 3.4: Posterior mean of spatial effect for the center regions. A region in red has a negative log-odds ratio, representing a lower odds and lower probability for patients in this region to receive TAVR, and a region in blue has a positive log-odds ratio, representing a higher odds and higher probability to receive TAVR.

(in the log-odds ratio) for patients in the fifth material quantile. Nevertheless, the effect sizes of these deprivation indexes show a clear trend: patients in the second quantile of both indexes have the most priority in receiving TAVR, followed by quantile 1 (the least deprived residents), quantile 3, quantile 4, and the most deprived residents in the fifth quantile. Though not significant, we can still find some inequalities caused by the lack of social-economic access. Worth to notice that the credible interval for material status level 5 (the most deprived) patients is particularly wide. One reason for such phenomenon is the heterogeneity in this population: patients in this category have extreme variety in living conditions, consisting of ones from metropolitan areas and ones from very remote regions of Quebec, as shown in Figure 3.6. The exact effect sizes and credible intervals can be found in Table 3.6 and Table 3.7.

3.4 Discussion

This paper analyzes the effect of clinical and non-clinical factors on AS treatment decisions. We accounted for disease severity from the available clinical variables using the Charlson scores while being aware that potential residual confounding variables are unobserved. Though not the first research to study the equity of medical resource allocation, we are the first to solve problems brought about by multi-collinearity. Given the nature of administrative data, we collect more information than required and thus apply the regularized shrinkage prior for selecting the related predictors while dropping the useless ones. Moreover, to model the spatial information in the dataset, we apply a conditional autoregressive prior with a 0-1 neighbourhood structure on the region-specific effect and measure the distance between patients and the closest TAVR centers. Compared to the models applied in previous research, our model successfully incorporates spatial structure into the regression model and presents an inclusive evaluation of patient's health conditions. Other than the clinical determinants, we identify the effect of patients' residence and interpret the effect of social-economic class concerning distance.

3.4.1 Limitations

Our analysis experienced four types of limitation:

1. Spatial confoundings: As we explore the estimates of covariates' effects on the outcome, the existence of collinearity between spatial information and covariates like social/material status may make the estimates biased. In this analysis, we ignored the possible spatial confoundings in our model. Looking at the estimates generated by the model without spatial effects and the contiguous neighbours model, though most effects are close, for **coronary heart disease**, **social status quantile 2**, and **material status quantile 5**, their corresponding effects are doubled, tripled or diminished to zero. Therefore, we can reduce the bias caused by spatial confoundings with more sophisticated spatial models.

2. Selection bias:

   Because we only selected the patients who received either the SAVR or TAVR operations, not all eligible patients could be assessed. For example, the AS operations performed on male patients since 2011 take 66% of the total, which exceeds the proportion of male patients diagnosed with AS (57%). Thus, comparing the two proportions, we observe that more male patients were selected for AS operations. Without evidence to prove that male AS patients are more severe, our analysis will be more convincing if we could first study the association between sex and operation selections.

3. Model accuracy limited by residual confoundings:

   As mentioned in previous US research, the race and ethnicity of the patients have been shown to be associated with the treatment decision. A limitation of this project is the absence of race, ethnicity and nationality data. As the only correlated predictor, the location of the patients is not fully adequate to explain the difference in TAVR rates observed in our dataset. Moreover, because of the characteristics of the administrative dataset, we do not have the details of the diagnosis, which means it is impossible to know whether patients are selected for surgery based on particular symptoms. We can only generate surrogate measures for severity.

4. Information bias:

   The last limitation of this research is the unrecoverable invalid values in the dataset. With mistakes in collecting patients' postal codes and matching the social/material deprivation index, the quantiles of 102 patients could not be found or imputed with other predictors. Though we assumed the invalid values existed at random, we can never validate this assumption, and thus may cause bias in our estimation by simply removing these samples.

3.4.2 Future work

As described in the limitations, one possible future work direction is incorporating other spatial models designed to solve the biases caused by spatial confoundings. For example, one model we could use is the *Spatial+* model proposed by Dupont et al. (2022), which reduces the collinearity by replacing the covariates with their regressed residuals. To apply the *Spatial+* idea in our Bayesian model, we can refer to the construction of the joint hierarchical model proposed by Michal et al. (2022), which extends the *Spatial+* model to a Bayesian setting.

Another direction for future work is to incorporate more longitudinal predictors. The current model does not consider the variation in patients' address and social-economic quantiles since their first diagnosis. The value of predictors we include in our analysis is based on the latest update at the surgery. Including the time-related information will help us to differentiate patients based on their wait times and investigate effects brought by the change of social/material status or the health regions.

Finally, an improvement could be made if the INSPQ provides information on the relation between hospitals' establishment numbers and their regions and functionality. We want to add a grouping effect based on the hospitals a patient visited, which will help us evaluate the distribution of specific medical resources among hospitals. The group can either be based on the hospital's functionality as a general, research or long-term hospital or based on the location and size of the hospital.

3.5 Conclusion

This research applies a Bayesian shrinkage model to explore the effect of clinical and non-clinical predictors on the valve replacement operation a patient received, with the regional information as random effects modelled through a binary adjacency variance structure for patients diagnosed with aortic stenosis from 2011 to 2018. Based on the estimates, severe patients with older age, more hospitalization records and ischemic heart

disease at the surgery are considered to encounter more risks for the traditional open heart operation plans. Despite clinical determinants, female patients in Quebec have a significantly higher chance of receiving TAVR treatment. Compared to social and material deprivation status, the healthcare region a patient belongs to has a strong correlation with the treatment decision. These findings emphasize the necessity of extra care for patients residing in remote regions and the importance of establishments for more comprehensive hospitals outside of the metropolitan areas.

3.6 Appendix

Region	Log-odds ratio mean	95% CI
Bas-Saint-Laurent	-0.369	[-0.678,-0.067]
Saguenay-Lac-Saint-Jean	-0.517	[-0.833,-0.209]
Capitale-Nationale	0.0.446	[0.156,0.739]
Mauricie	-0.380	[-0.694,0.075]
Estrie	-0.863	[-1.135,-0.577]
Montreal	0.192	[-0.119,0.485]
Outaouais	-0.320	[-0.704,0.065]
Abitibi-Temiscamingue	-0.569	[-0.954,-0.209]
Cote-Nord	-0.400	[-0.733,-0.078]
Nord-du-Quebec	-0.406	[-0.782,-0.059]
Gaspesie-lles-de-la-Madeleine	-0.774	[-1.064,-0.478]
Chaudiere-Appalaches	-0.133	[-0.439,0.137]
Laval	-0.310	[-0.619,-0.016]
Lanaudiere	-0.388	[-0.690,-0.096]
Laurentides	-0.616	[-0.901,-0.335]
Monteregie	-0.645	[-0.917,-0.369]
Centre-du-Quebec	-0.946	[-0.723,-0.193]

Table 3.5: The posterior mean of spatial effect for 17 health care regions in log-odds ratio.

	Log-odds ratio Mean	95% Credible Intervals
social 2nd	0.030	[-0.063,0.187]
social 3rd	-0.025	[-0.176,0.064]
social 4th	0.002	[-0.093,0.102]
social 5th	-0.022	[-0.169,0.065]
material 2nd	0.037	[-0.046,0.218]
material 3rd	-0.010	[-0.129,0.074]
material 4th	-0.019	[-0.152,0.066]
material 5th	-0.044	[-0.749,0.277]

Table 3.6: Estimated log-odds ratios of social and material deprivation index quantiles on the treatment decision

Predictor	Log-odds ratio mean	95% CI
Sex	-0.263	[-0.439,-0.068]
Age	0.142	[0.131,0.153]
Charlson score	0.086	[0.055,0.116]
Hospitalization	0.139	[0.067,0.211]
Drugs Intake	-0.015	[-0.019,-0.011]
Unique Drugs	0.040	[0.033,0.047]
Hypertension	-0.018	[-0.065,0.013]
Pulmonary HD	0.001	[-0.060,0.061]
Ischemic HD	0.037	[0.002,0.071]
Coronary HD	-0.075	[-0.161,0.005]
Artery HD	-0.005	[-0.085,0.061]
Other HD	0.026	[0.004,0.047]

Table 3.7: Estimated log-odds ratios on patients with cardio diseases on the treatment decision. HD represents ”heart disease”.

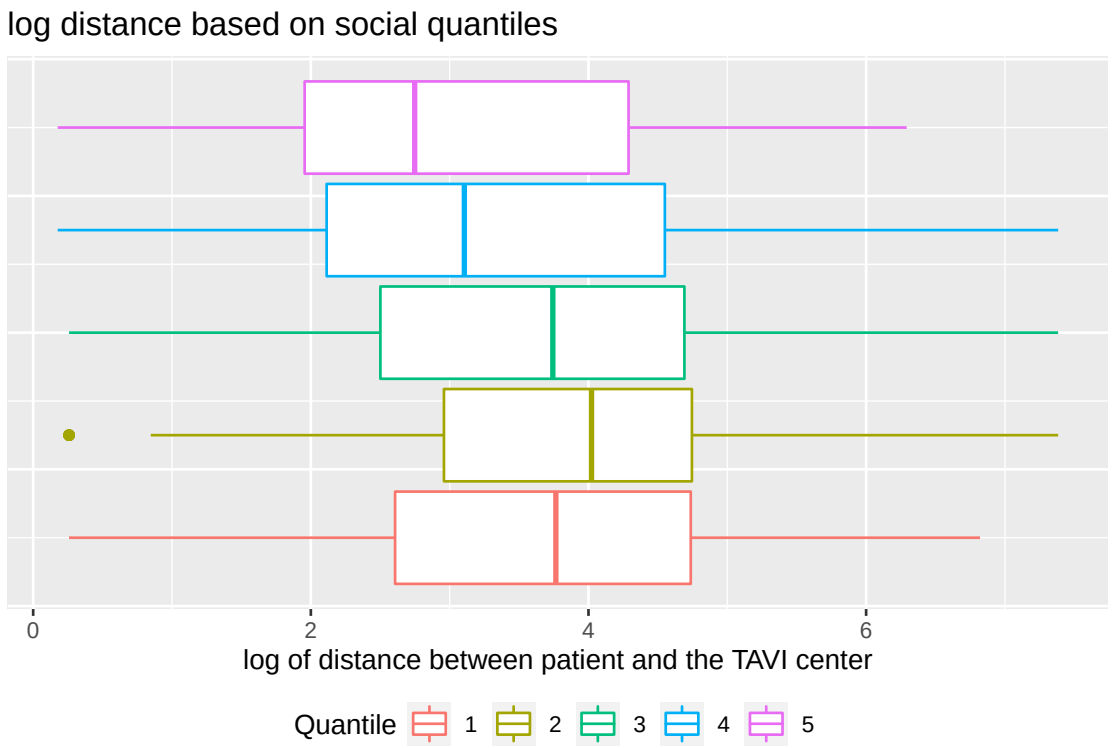

Figure 3.5: Distribution of log(D) with respect to social deprivation index quantiles, where D is the distance between a patient and the closest TAVR center

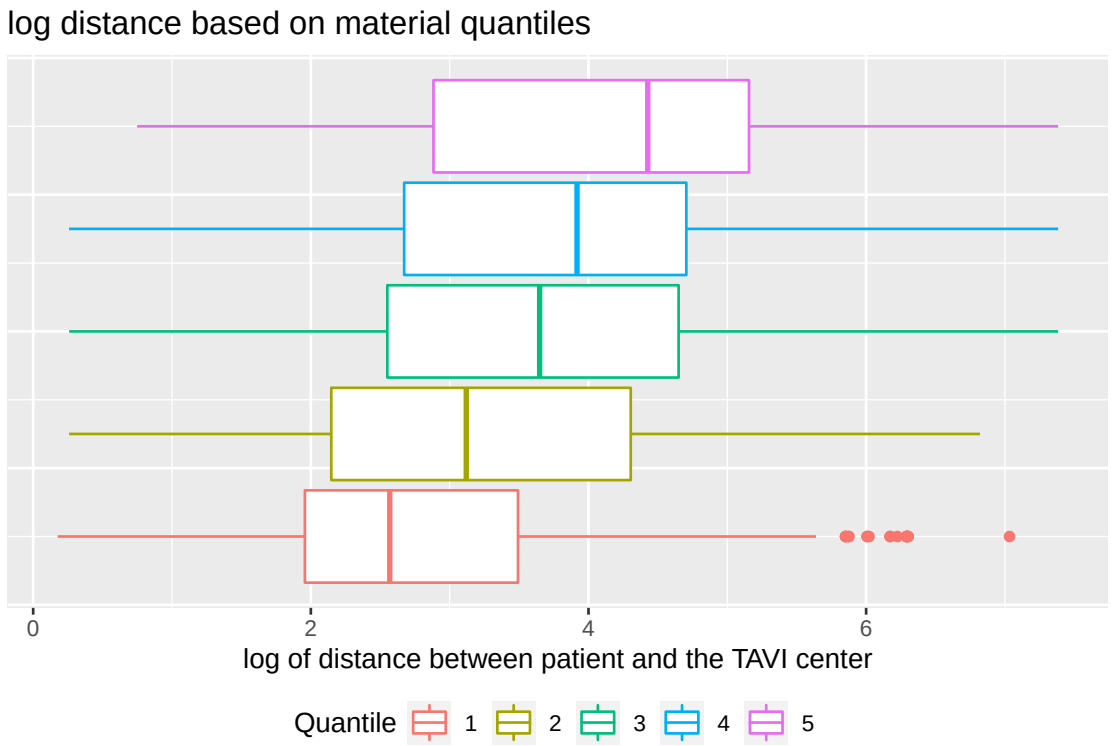

Figure 3.6: Distribution of log(D) with respect to material deprivation index quantiles, where D is the distance between a patient and the closest TAVR center

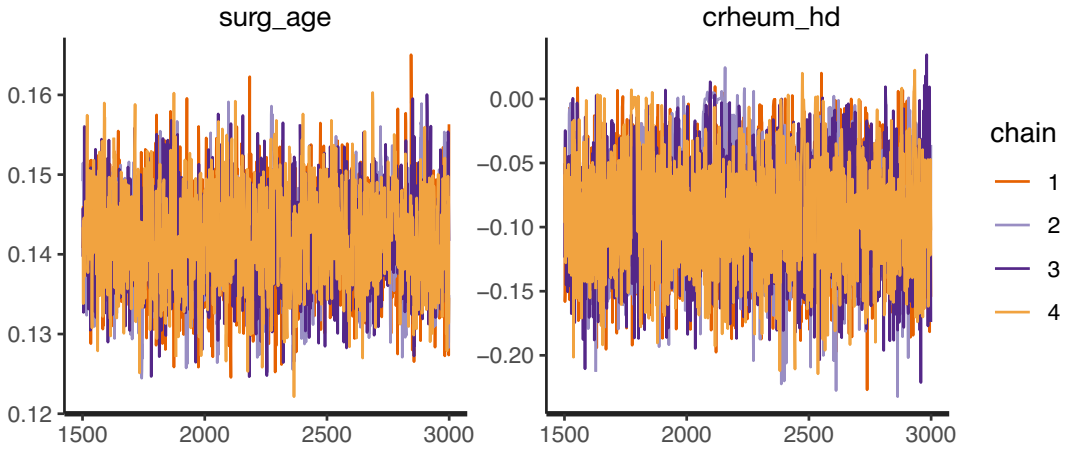

Figure 3.7: Traceplots of the estimates corresponding to **age at surgery** and **coronary heart disease** with four chains run in parallel. The x-axis shows the iteration number and the y-axis records the sample values.

Chapter 4

Exploring factors correlated with HIV infection through variable selection

4.1 Introduction

HIV, the human immunodeficiency virus, attacks the human immune system and cannot be effectively cured. Most people get HIV by sex or through the repeated use of injection equipment, as the virus transmits through body fluid (CDC, 2021), but can only know their diagnosis through testing. HIV testing provides benefits like early diagnosis and treatment initiation, prevents the transmission of HIV between partners, and prolongs the lifespan of infected patients (WHO and UNAIDS, 2017). However, in South Africa, the country with the largest population of HIV-infected patients, the prevalence of in-person HIV testing is impeded by the lack of healthcare facilities and social factors such as discrimination (Strauss et al., 2015). In order to provide individuals who had difficulty coming to clinics opportunities to evaluate their risk of HIV, a self-testing mobile application called *HIVSmart!* was developed (Janssen et al., 2019). This app is designed to help users monitor their health conditions by answering a series of questions, which are then uploaded to an online database.

Since self-testing through a model app was an innovative approach, previous studies tried to show that the self-testing procedure brought benefits in early diagnosis and illustrated its convenience compared to other forms of a health tracker. Regan et al. (2013) performed analysis on the performance of self-testing to show that frequent reporting reduced HIV infections and helped the treatment of HIV. Pai et al. (2021) showed that the digital self-reporting HIV test submitted on the app could monitor an individual's health condition in the same way as the conventional tests. Though proven to provide the same benefits as the traditional tests, the self-testing approach will only be beneficial if we can find the relationship between the sexual behaviour questions, the personal background information and the HIV status. Therefore, by conducting regular self-reporting, epidemiologists can warn people with specific answers about their risk of being infected with the estimated effects.

However, the presence of missing values in the self-testing data complicates the estimation of the HIV infection determinants. Unlike traditional HIV testing, because the self-testing reports could be done without the monitoring of medical professionals, they usually contained missing values due to technical malfunctions. For example, answers on certain questionnaire pages accidentally could not be uploaded to the database. Moreover, it is possible that participants refused or forgot to answer some questions and hence generated missing values. The previous studies about the self-testing data analysis dropped all observations with missing values and only analyzed the complete ones, which may cause the estimates to be biased. Therefore, we intend to improve the estimation by imputing these missing values based on non-informative prior distributions and taking the uncertainty of missingness into account.

In this analysis, we employ a Bayesian hierarchical model that can impute the missing values, especially for multiple variables, and analyze the effects of the predictors, such as sexual behaviours, based on the HIV data collected through the self-testing app simultaneously. Because of the foreseen level of the sparsity of significant predictors among the available information in the questionnaire, we decide to apply the variable selection

method for analysis. For the three predictors containing missing values, we assume all missing and observed samples of each predictor come from the same distribution and impute the missing values through a Beta-Bernoulli distribution. For the categories with insufficient information due to their small sample sizes or the small number of HIV-infected participants, we either combine these categories when the resulting category is still meaningful or provide explanations for such a phenomenon and propose further research strategies.

We construct the chapter as follows: in Section 4.2, we present a summary of the predictors: questions the self-reporting app asked, the importance the predictors played in HIV infection and their distribution. Then, in Section 4.3, considering the missingness in our dataset, we describe how the fully Bayesian approach imputes the values and list the missing variables in our HIV dataset. The hierarchical model that simultaneously imputes the missing values and estimates the predictors' effects is presented in Section 4.5.2. We assign the regularized Horseshoe prior to all the estimates of predictors with a sparsity assumption. After constructing the model, we present the result with a descriptive analysis of the predictors and their posterior distributions generated by the model. Finally, we discuss the social and epidemiological meaning behind these numerical results and conclude our findings.

4.2 Study cohort

The study was conducted in Capetown, South Africa, from January 2017 to June 2018 and used convenience sampling for recruiting participants. All the participants were 18 years or older, with unknown HIV status at self-reporting. As owners of an Android/Apple smartphone, the participants answered the questions for the self-report testing using the app *HIVSmart!*.

Participants were asked to respond to a list of questions in the app. Janssen et al. (2019) collected the answers and named the predictor of interest in each question. The

details of the predictors, including the question in the report, available answers and their distributions, are presented in Table 4.1. It is worth noting that there are only two answer options, male and female, for the **gender** question. Therefore, the **gender** variable here actually represents the "biological sex". However, to be consistent with the questionnaire, we will refer this variable as "gender" in our statistical analysis.

Followed by the background information and health condition, the questionnaire in *HIVSmart!* asked about the sexual behaviours of the participants. The behaviours were listed in a sequence, and the participants were asked to select all applied descriptions. Therefore, Pai et al. (2021) recorded a checked box as "Yes" and an unchecked box as "No". The questions and answers are shown in Table 4.2.

In this study, we consider the HIV status of the participants as our response variable. The participants were tested for HIV after completing their self-reporting test on the app. We denote HIV positive as 1 and negative as 0. Among the 1535 participants, 137 of them were confirmed with HIV.

Predictor	Explanation / Question	Type	Answer :Distribution
Site	At which clinic site the information is collected?	categorical	1.Gugulethu: 528 (34.40%); 2.Langa: 469 (30.55%); 3.Weltevreden: 538 (35.05%)
age	How old are you?	continuous	Mean: 28.23 (SD: 8.83)
tb	Have you ever been diagnosed with tuberculosis?	binary	Yes: 114 (7.43%)
postsec education	Have you received postsecondary education?	binary	Yes: 312 (20.33%)
dwelling	Do you live in hostel or informal dwelling?	binary	Yes: 677 (44.10%)
comorbidities	Have you ever been diagnosed with: diabetes, other lung disease, or hypertension?	binary	Yes: 90 (5.86%)
employment	Are you full-time employed?	binary	Yes: 544 (35.44%)
gender	biological sex	binary	Female: 994 (64.76%)
monthly income	What is your approximate monthly income (in rand)?	categorical	$\leq$ 3000: 1190 (77.52 %); 3000-6000: 153 (9.97%); 6000-9000: 61 (3.97%); $\geq$ 9000: 68 (4.43%)
township	In which township do you live?	categorical	1. Klipfontein: 391 (25.47%); 2. Mitchells Plain: 486 (31.66%); 3. Western District: 595 (38.76%);
tested	Have you been tested for HIV in the past 6 months?	binary	Yes: 707 (46.06%)
inject drugs	In the past 6 months, have you injected drugs?	categorical	Yes = 96 (6.25%); No = 1350 (87.95%); Abstain = 25 (1.63%)
HIV exposed	In the past 6 months, have you been exposed to HIV at your work place?	categorical	Yes = 161 (10.49%); No = 1261 (82.15%); Abstain = 49 (3.19%)
sexual orientation	What's your sexual orientation?	categorical	Homo: 129(8.4%); Hetero: 994(64.76%); Bi: 52(3.39%); Abstain: 297(19.35%)

Table 4.1: Summary of personal background and living conditions questions based on the actions in the past six months. Rand, or South African Rand, is the official currency in South Africa, with 1 rand = 0.055 USD in Oct. 2022.

Predictor	Question	Type	# of Yes(%)
sex active	You are sexually active.	categorical	1223 (79.67)
current partner	Your current partner is your husband or wife.	binary	429 (27.95)
without condom	I have had sex without a condom	binary	937 (61.04)
with many ppl	I have had sex with multiple partners.	binary	167 (10.88)
with sexworker	I have had sex with a commercial sex worker.	binary	21 (1.37)
with HIV ppl	I have had sex with an HIV-infected partner.	binary	42 (2.74)
with alcohol	I have had sex under the influence of alcohol.	binary	73 (4.76)
with drugs	I have had sex under the influence of drugs.	binary	32 (2.08)
abstain	I do not wish to answer.	binary	58 (3.78)

Table 4.2: Summary of sexual behaviours based on the action in the past six months

4.3 Missingness and missing data

When we observe missing values in our data, prior to handling them, it is crucial to understand why the data are missing. Here we summarize three general reasons for missingness (Rubin, 1976):

Missing completely at random (MCAR)

We say a variable is missing completely at random if the probability of missing is the same for every observation. For example, if an answer in an online questionnaire is missing because of an internet speed problem in the uploading process, we can treat it as MCAR.

Missing at random (MAR)

In a more general setting than MCAR, variables are missing at random if the probability of missing depends only on other observed variables, such as measurements taken by a piece of traditional equipment are more likely to be missing compared to that taken by an innovated equipment. However, measurements taken with the same equipment have the same probability of missing.

Missing not at random (MNAR)

(a) Missingness depends on unobserved variables

Sometimes the missing values of a variable for some observations exist because of other values that have not been recorded or available for imputing or modelling (Gelman and Hill, 2006). For example, participants born in more conservative or religious families are less likely to answer questions about sexual orientation or immoral sexual behaviours.

(b) Missingness depends on the value itself

The probability of missing depends on the value of the missing variables. For example, people might be reluctant to answer questions related to income, especially if they are extremely rich or poor.

4.4 Missing data imputation approach

When one or more predictors contain missing values, we need to remove, directly impute or impute through modelling. Here we present some popular approaches among the various imputation methods available nowadays.

4.4.1 The frequentist approaches

Methods for dealing with missing data

1. Discard the missing data:

   The easiest way is to throw away samples with missing values. Removing the missing variables does not cause any bias when the data are missing completely at random. Furthermore, even with missing at random data, it is reasonable to remove the missing variables as long as we ensure that there is no unmeasured variable that affects the probability of missingness. However, when data are missing not

at random, discarding the missing values will introduce bias in the estimation of predicting variables' effects on the response.

2. Single imputation:

   This approach fills in the values of missing inputs to obtain a complete data set. Some commonly used methods are "mean imputation", which imputes by the mean of the observed values, "last value carried forward" , which imputes by the last observed value for this sample in longitudinal data analysis, "indicator variables for missingness", which impute as an additional category, exclusively for unordered categorical variable, and simple random imputation, which impute each missing value by a randomly sampled observed value (Gelman and Hill, 2006).

3. Model through regression:

   We can use the individual-level information to impute the missing values through a multiple regression model by treating the variable with missingness as the outcome. We fit the model with the complete cases and then use the estimated coefficients to impute the missing observations. With the missing at random assumption, we can model the missing variables through univariate/multivariate regressions depending on the number of variables with missing values. However, for data missing not at random, imputation through regressions will lead to a biased estimation of the predictors' effects on the outcome.

4. Multiple imputation:

   In addition to the approaches mentioned above, multiple imputation (MI) is often applied to missing data problems. Briefly, multiple imputation is a method similar to "model through regression" but generates multiple datasets and analyzes each dataset using a regression model. The results obtained from these analyses are averaged as the imputed values (Van Buuren, 2018). Among all the approaches, the multiple imputation approach has gained popularity in dealing with missing data problems. It accounts for statistical uncertainty compared to single imputation, re-

quires fewer missing data mechanism assumptions compared to the complete-case analysis, and applies to more types of models than the maximum likelihood approach (Azur et al., 2011).

Though generally, multiple imputation is less biased, it is not powerful at all times (Van Buuren, 2018). There are cases where other methods outperform. For example, when the complete-data model is a regression between the outcome $\boldsymbol{y}$ and predictors $\boldsymbol{X}$, and only $\boldsymbol{y}$ contains missing values, the multiple imputation is equivalent to complete-case analysis.

4.4.2 The fully Bayesian approach for ignorable missingness

The fully Bayesian approach treats missing values as parameters for prediction. As suggested by Ma and Chen (2018), it is performed through four steps: 1) proposing the response model (missing data distribution if needed), 2) constructing the prior distribution, 3) calculating the posterior through MCMC and 4) performing a sensitivity analysis due to the inability of knowing the true missing mechanism.

Generally, we denote the predictors with missing values as $\boldsymbol{X} = (\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(0)})$, with (1) the observed and (0) the missing part, the corresponding response as $\boldsymbol{y} = (\boldsymbol{y}_{(1)}, \boldsymbol{y}_{(0)})$, and N_1 the number of observed samples. With the predictors' effects $\boldsymbol{\delta}$ on response and parametrization coefficients Δ on the distribution of $\boldsymbol{X}$, we write their joint posterior distribution as (Ma and Chen, 2018)

$$p(\boldsymbol{\delta}, \Delta | \boldsymbol{X}_{(1)}, \boldsymbol{y}_{(1)}, \boldsymbol{y}_{(0)}, N_{(1)}) \propto \int_{\boldsymbol{X}_{(0)}} f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\delta}) f(\boldsymbol{X}_{(0)}|\boldsymbol{X}_{(1)}, \Delta) d\boldsymbol{X}_{(0)} \pi(\Delta, \boldsymbol{\delta}),$$

where $f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\delta})$ represents the response model, $f(\boldsymbol{X}_{(0)}|\boldsymbol{X}_{(1)}, \Delta)$ the missing data distribution, and $\pi(\Delta, \boldsymbol{\delta})$ is the assigned prior distribution.

Assuming the missing data mechanism is ignorable, which means either MCAR or MAR, we have all values exchangeable, i.e., no matter the values are observed or missing, they follow the same distribution (McElreath, 2016). Therefore, for continuous variable

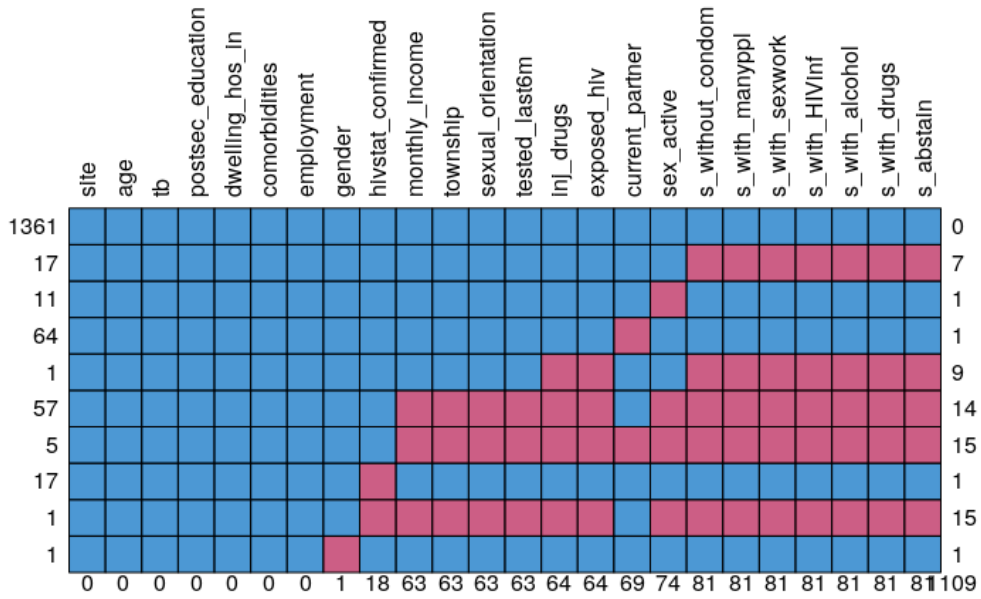

Figure 4.1: Missing data pattern plot: complete cases are painted blue and missing cases are red. The predictors' names are listed on the top, and the number of missing values in each predictor is at the bottom. Each row represents one combination of missing variables: the number on the left denotes how many samples have these variable missing, with the number on the right counts for how many variables are missing.

$\boldsymbol{X} = (\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(0)})$, we assign its distribution as

$$\begin{aligned} \boldsymbol{X}_{(1)} &\sim \mathcal{N}(f(\cdot), \sigma^2), \\ \boldsymbol{X}_{(0)} &\sim \mathcal{N}(f(\cdot), \sigma^2), \end{aligned} \tag{4.1}$$

where we denote the mean of $\boldsymbol{X}$ as $f(\cdot)$, contributed by some corresponding related covariates and the coefficient Δ, and set $\sigma^2 > 0$. For binary or count missing values, we change the distribution in (4.1) to Bernoulli or Multinomial distributions.

4.4.3 Missingness in HIV data

For this HIV self-reporting dataset, we found 81 participants had all sexual behaviour answers missing from the missing pattern plot (Figure 4.1). Based on the analysis provided by Janssen et al. (2019), the phenomenon that all sexual-behaviour answers are missing

at the same time is very likely due to the malfunction of the self-testing app. With almost one-third of the predictors missing for a single sample, imputing all missing values, either through regression models or distributions inferred from the observed values, will induce estimation bias and variance in the imputation process. Moreover, without any sexual behaviour answers available, which played a critical role in the infection of HIV, it is impossible to recover all behaviours for one participant and the relationship between these behaviours and HIV prevalence. Therefore, we assume that the app malfunction happened completely at random and excluded participants with all sexual behaviour answers missing.

After removing the 81 samples with all sexual behaviour response missing, we only have **gender**, **current partner** and **sex active** containing missing values. Due to its nature as a self-identifying report, a third option, "Abstain", is available as an answer other than "Yes" and "No" for participants to select for some questions. To respect all participants' will, we do not consider these "Abstain" answers as missing. For **gender**, only 1 out of 1535 samples is missing while 64 responses of **current partner** and 11 responses of **sex active** are missing. For each variable, we assume all 1437 values, including both the observed and the missing values, are samples of the same distribution. For example, for **gender**, we have the following hierarchical structure:

$$\begin{aligned} \text{gender}_i &\sim \text{Bernoulli}(p_{\text{gender}}), \\ p_{\text{gender}} &\sim \text{Beta}(a_{0g}, b_{0g}), \end{aligned}$$

for $i = 1, \ldots, n$ and n denotes the sample size. We set $a_{0g} = 1, b_{0g} = 1$ for a non-informative prior. We assign the same distribution to **current partner** and **sex active** with $\text{partner}_i \sim \text{Bernoulli}(p_{\text{partner}})$ and $\text{active}_i \sim \text{Bernoulli}(p_{\text{active}})$. Each of the probability p_{partner} and p_{active} follows a Beta distribution with $a_{0p} = b_{0p} = a_{0a} = b_{0a} = 1$. When constructing the model, we estimate these probabilities while integrating out the missing ones in the outcome regression model.

4.5 Method

As discussed above, the existence of missing values complicates our analysis. Furthermore, assuming that not all predictors we extracted from the report are helpful, we need to use variable selection techniques to recover the true effects of significant predictors while dropping the others. Therefore, to better analyze the HIV data, we need a model that considers the uncertainty brought by missing values and efficiently performs variable selection in the model fitting process.

In this section, we employ a Bayesian hierarchical model that simultaneously estimates the predictors' coefficients $\boldsymbol{\beta}$ and imputes missing values to fulfill all requirements.

4.5.1 Notations

We denote the HIV test result of all participants as $\boldsymbol{y}$ with size n. Let $y_i = 1$ if the HIV status of participant i is positive and $y_i = 0$ otherwise. The categorical predictors in our HIV data are first transformed into dummy variables and then combined with other binary or continuous predictors. The categorical data we transformed in the HIV self-report testing analysis includes: **level of monthly income**, **clinical sites**, **residential area**, and all questions with "Abstain" option. Let J denotes the number of predictors, including the transferred dummy variables, in the HIV dataset. The matrix of all predictors is denoted as $\boldsymbol{X} = (\boldsymbol{X}_{com}, \boldsymbol{X}_{\text{gender}}, \boldsymbol{X}_{\text{active}}, \boldsymbol{X}_{\text{partner}})$, with $\boldsymbol{X}_{com}$ a $n \times (J-3)$ matrix of complete variables, and $\boldsymbol{X}_{\text{gender}}$, $\boldsymbol{X}_{\text{active}}$, $\boldsymbol{X}_{\text{partner}}$ the columns of predictors with missing values. The distributions of all missing predictors are described in Eq. 4.4.3.

4.5.2 Model

To model the effect of predictors on the dichotomous response $\boldsymbol{y}$ (HIV status), we establish a logistic regression. For participant i, we build our model as:

$$y_i \sim Bernoulli(\pi_i),$$
$$\text{logit}(\pi_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \alpha,$$

where the coefficients $\boldsymbol{\beta} = (\boldsymbol{\beta_{obs}}, \beta_{gender}, \ldots, \beta_{partner})$ correspond to the effect of the predictors on the HIV status. Specifically, we use $\boldsymbol{\beta_{obs}}$, a vector of length $J-3$, to represent the coefficients of the complete predictors. Furthermore, after centering the continuous variables on their mean values, we use α to denote the baseline odds for male patient at the age of 28.23, who lived in Klipfontein, answered "No" to all the questions, filled in the report in the Gugulethu clinic and had a monthly income below 3000 in the prediction model. We run the model with 4 chains in parallel through `Rstan`. Each chain contains 3000 iterations, with the first 1500 set as burn-in and target average acceptance probability set to 0.95.

We draw inference from the joint posterior of all parameters' distributions, $\boldsymbol{\beta}, \alpha$, in the model, and the structure can be expressed as

$$\pi(\boldsymbol{\beta}, \alpha | \boldsymbol{y}, \boldsymbol{X}) \propto \pi(\boldsymbol{y} | \boldsymbol{\beta}, \boldsymbol{X}, \alpha) \pi(\boldsymbol{\beta}, \alpha)$$

with the right hand side equals to the likelihood times the prior. Furthermore, we construct the prior as the product of prior distributions of the parameters in the model (Golchi et al., 2022):

$$\pi(\boldsymbol{\beta}, \alpha) = \pi(\boldsymbol{\beta}) \pi(\alpha),$$

with α assumed to follow a standard Normal distribution, i.e. $\alpha \sim \mathcal{N}(0, 1)$,

Because we believe the distribution of truly influential predictors is sparse and the correlation among them could be high, we decided to assign a continuous shrinkage prior to the coefficients corresponding to all the predictors. As illustrated by Piironen and Vehtari (2017), compared to the Laplace prior and Horseshoe prior, the regularized Horseshoe reaches convergence more quickly in more conditions and shrinks the large value in the regression. Therefore, we apply the regularized Horseshoe prior to each β_j in $\boldsymbol{\beta}$ for variable selection. The structure of prior is described in Section 2.3.2 with the hyperparameters set at the default values: $\nu = 20$, $s^2 = 4$ and $\tau_0 = 0.001$.

4.6 Result

Among the 1437 participants used in the analysis, 7.9% of them tested HIV positive. The background information and answers to economic questions showed the potential risks of delayed HIV diagnosis, and the optimistic employment, income and housing conditions of the participants. With only 35% employed, 44% of our sample lived in a hostel or informal dwelling, and 77% of them had a monthly income of 3000 rand or less. These participants, who struggled to survive, lacked the ability to monitor their health conditions. Moreover, most participants did not show awareness of protective actions in sexual behaviours. Among the 1223 sexually active participants, 937 of them did not use condoms, and 167 had sex with multiple partners. These discouraged behaviours further increased their exposure to HIV and accelerated the prevalence of HIV.

4.6.1 Inference from model

To ensure the convergence of our model, we checked the $\hat{R}$ and the traceplots for the posteriors. With all $\hat{R} < 1.05$ and traceplots indicating that all chains mixed well, we summarized the mean and 95% credible intervals of the estimated β and plotted in Figure 4.2. The posterior summaries can be found in Table 4.3 in 4.8. We also present the traceplot of two predictors with wide credible intervals, **bisexual orientation** and **comorbidities**, to show that the widths of these intervals are not due to lack of convergence in Figure4.4 as an example.

Estimated from the hierarchical model, we found a few predictors strongly associated with the HIV status of the participants. Though not statistically significant based on their 95% credible intervals, 45.2% more participants who finished the report in the Langa clinic tested HIV positive. Participants who had sex with one or more HIV-infected person is 1.14 times more likely to have HIV. On the other hand, those who checked their HIV status regularly had lower infection risk: participants who had a negative test result within the past six months were 47.3% less likely to be HIV positive at the time of the report,

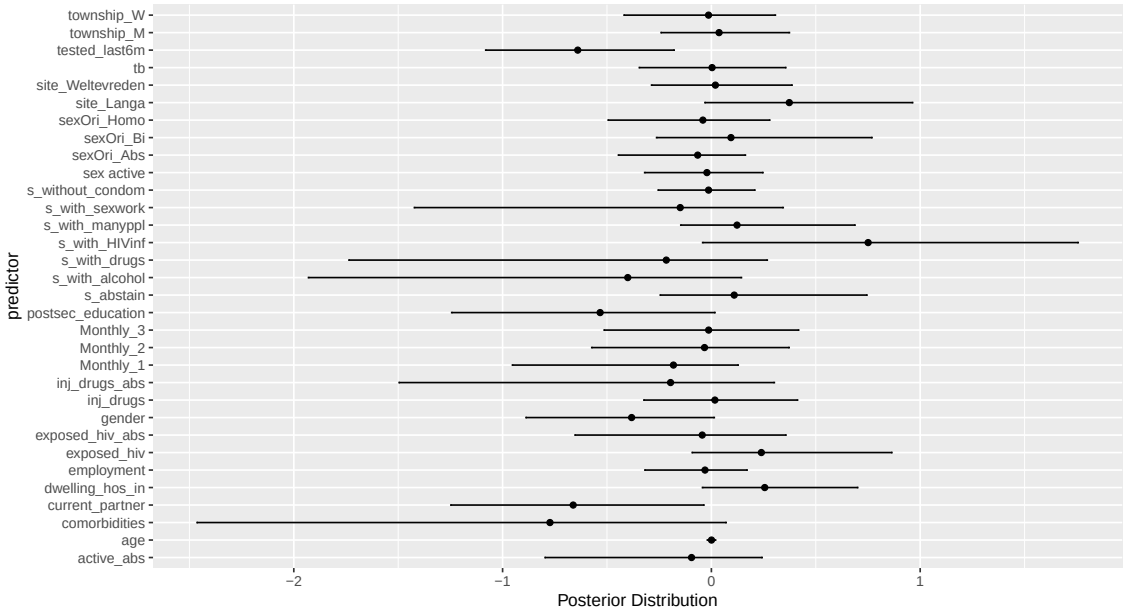

Figure 4.2: Posterior summaries of the estimated effects of predictors on HIV status in the form of log-odds ratios. Solid circles denote the posterior means and bars denote the 95% credible intervals of the posteriors' distributions.

and 41.3% and 53.8% fewer people with a post-secondary degree or comorbidities, such as diabetes and hypertension, were diagnosed with HIV. We find a strong association between one's gender and HIV status. Compared with males, 31.7% fewer females were HIV-infected. Finally, from the estimates, the correlation between age or monthly income and HIV infection is negligible.

4.6.2 Problem in estimated credible intervals

As shown from descriptive analysis, the proportion of HIV-infected participants is small. Therefore, for some categories with only a few participants involved, it is possible to have a wide credible interval due to insufficient information.

Among the variables with wide 95% credible intervals, the "Abstain" category for **drug injection**, **exposed to HIV infection**, and **sexual active** questions obtain large credible intervals because of the limited number of participants we have. Because we cannot determine the condition of these participants or make any assumptions about their

preferred answers, we leave them as a separate category and stay cautious about these estimates. An ideal solution to prevent these low-observed categories is to remove the "Abstain" option from these anonymized reports.

For predictor **comorbidities**, **postsecondary education** and **clinic site Lange**, we observed wide credible intervals together with strong negative or positive effects. We assume these intervals are caused by the variability among participants who answered "Yes" to the above questions: these participants belong to different categories in both observed predictors, such as income and employment, and possibly unobserved predictors, like health conditions and social classes. However, with the strong effects and almost significant credible intervals, we were confident in the direction of their correlation to HIV infection.

Besides, most sexual behaviour questions have a small proportion of "Yes" responses from participants with few positive HIV test results. Therefore, the samples for these minority behaviours are not representative enough, and we combined the answers to these unprotected sexual activity questions into one variable named **unsafe sex**. Hence, if one participant, denoted as participant i, has had sex **with many people** (10.88%) or **with sex-worker** (1.37%) or **with HIV-infected person** (2.74%) or **under the influence of drug or alcohol** (4.76% / 2.08%), we record $\text{unsafe}_i = 1$ and 0 otherwise.

With such a combination, we have 20.4% participants who experienced at least one unsecured sexual activity, and 33 were HIV positive, which is more representative of the population. We reran the model with the combined categories; its estimation is shown in Figure 4.3. We found that people with risky sexual experiences had a 7.8% higher chance to be HIV-infected, showing the harm of such behaviours.

4.7 Discussion

In this research, we investigated the relationship among personal background, living conditions, sexual behaviours and HIV infection with self-testing reports submitted by

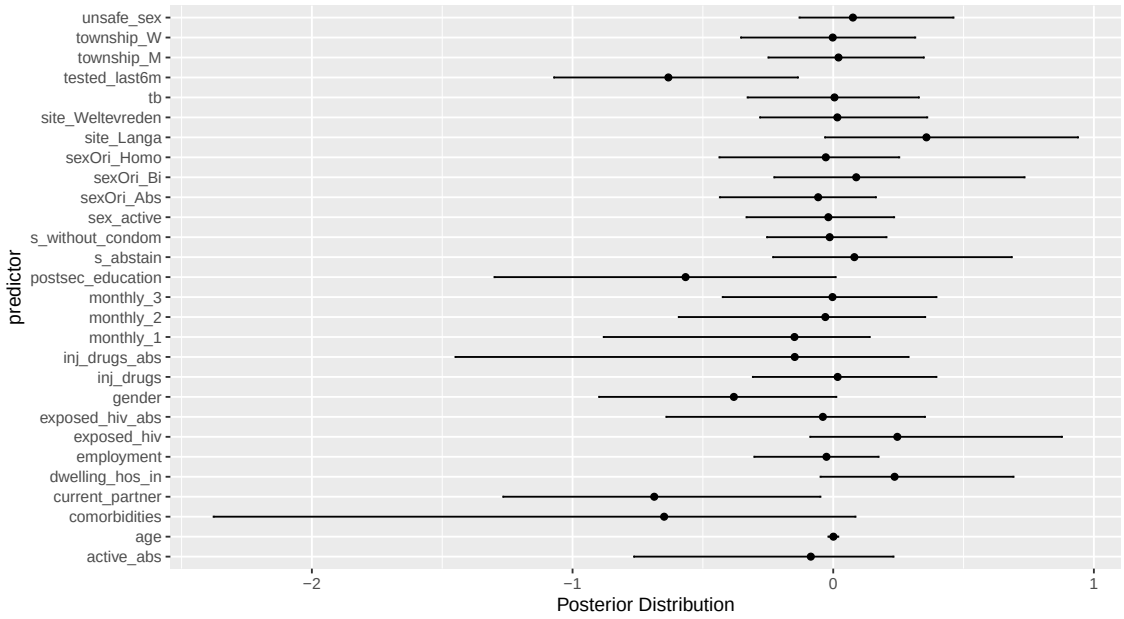

Figure 4.3: Posterior summaries of the estimated effects of predictors on HIV status in the form of log-odds ratios, with risky sexual behaviours combined into the "unsafe sex" category. Solid circles denote the posterior means and bars denote the 95% credible intervals of the posteriors' distributions.

participants through the mobile app *HIVSmart!* in South Africa. The data we collected showed that the participants were generally young, poor, uneducated and without enough caution regarding the protections in sexual behaviours. Impeded by the doubtful living conditions, more participants without regular health checking are HIV positive, and so do those with unsafe sexual experience, especially if they ever had sex with sex workers or HIV-infected persons.

Followed by the descriptive results, we dealt with the missing value problems in this research. Because of the nature of the self-testing report, we do not consider "Abstain" answers as missing but keep all of them. We observed a sequence of missingness overall sexual behaviour questions for 81 participants caused possibly by the app malfunction. Without any related information, imputation for these missing variables can be biased and largely varied. Therefore, we assumed the malfunction happened completely at ran-

dom and removed these samples. For the three missing variables left, we imputed each variable with missing values in a fully Bayesian way.

In order to model the outcome variable, we applied a Bayesian hierarchical model to impute the missing values simultaneously and to estimate the effect of predictors on the HIV result. Through summarizing the posterior means and credible intervals, our research identifies factors related to the infection of HIV. In general, participants who had earned a post-secondary degree, had comorbidities, or had sex with their wives or husbands were less exposed to HIV infection, and so were those who had another HIV test within six months. Conversely, more participants who had sex without condoms or other unsafe sexual behaviour, especially among HIV-infected people, were HIV positive. Surprisingly, the information about sexual orientation and living regions did not reflect on the risk of HIV infection. It is worth noting that many more participants who filled in the report in the clinic at Langa clinic were HIV positive, which requires a further investigation of the social and economic environment around that clinic to find a reasonable explanation.

The limitation of our model is the unidentifiable missing pattern for the sexual behaviour questions. Though we assumed the missingness of the series of sexual behaviour answers was missing at random due to a malfunction of the app, we did not know and could not know the true missing mechanism. Because it is possible that some people were unwilling to answer questions related to their sexual activities and left these answers blank, discarding these values may bring bias to our estimation. Furthermore, imputation would introduce variance because of the large proportion of missingness on one sample. There is no perfect solution to balance the bias and the variance with currently available information.

4.8 Appendix

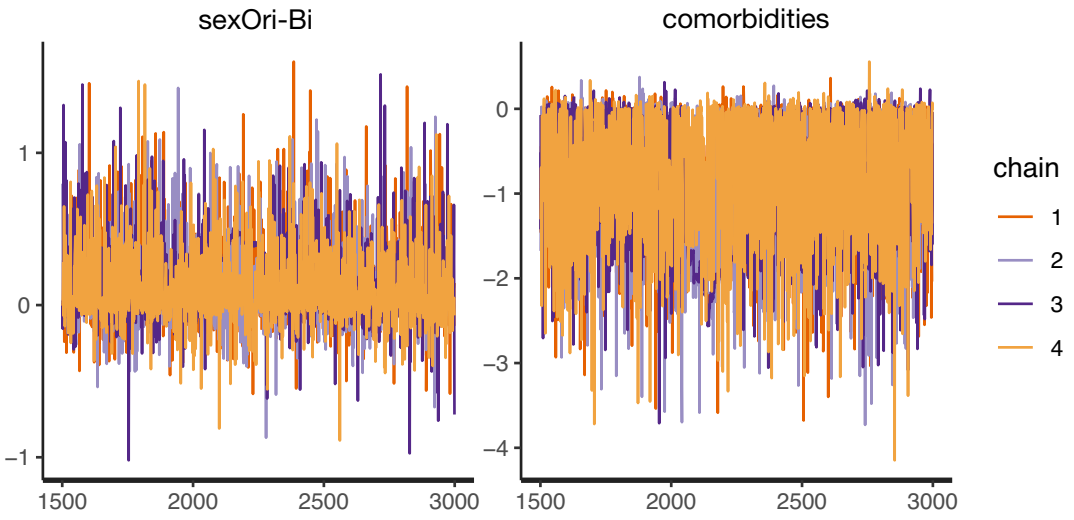

Figure 4.4: Traceplots of the estimates corresponding to **bisexual** and **comorbidities** with four chains run in parallel. The x-axis shows the iteration number and the y-axis records the sample values.

Predictor	Log-odds ratio Mean	95% Credible Intervals
Weltevreden site	0.019	[-0.286,0.386]
Langa site	0.373	[-0.031,0.964]
age	0.001	[-0.019,0.020]
TB	0.003	[-0.345,0.357]
Postsec education	-0.533	[-1.243,0.017]
Dwelling	0.256	[-0.043,0.701]
Comorbidities	-0.773	[-2.463,0.071]
Employment	-0.031	[-0.318,0.171]
Gender	-0.382	[-0.888,0.014]
monthly income 2	-0.033	[-0.573,0.372]
monthly income 3	-0.013	[-0.513,0.418]
Mitchells plain	0.037	[-0.241,0.374]
Western District	-0.014	[-0.418.0.306]
HIV tested	-0.640	[-1.081,-0.178]
Inject drugs	0.017	[-0.323,0.413]
Inject drugs abstain	-0.196	[-1.495,0.302]
HIV exposed	0.239	[-0.092,0.865]
HIV exposed abstain	-0.044	[-0.653,0.357]
Bisexual	0.094	[-0.262,0.769]
Homosexual	-0.041	[-0.494,0.279]
Abstain sexual orientation	-0.066	[-0.445,0.163]
Sex active	-0.021	[-0.319,0.247]
Current partner	-0.662	[-1.248,-0.036]
Sex without condom	-0.013	[-0.255,0.208]
With many people	0.123	[-0.146,0.688]
With sexworker	-0.149	[-1.422,0.344]
With HIV-infected ppl	0.751	[-0.042,1.757]
With alcohol	-0.400	[-1.930,0.144]
With drugs	-0.216	[-1.736,0.268]
Abstain sex question	0.109	[-0.245,0.745]

Table 4.3: Estimated log-odds ratios of predictors in HIV data

Chapter 5

Conclusion

In this thesis, we summarize the current variable selection methods from model selection to lasso penalization, followed by the Bayesian methods, such as the spike-and-slab priors, the shrinkage priors, and the discrete-continuous mixture prior. After describing the benefits of the shrinkage priors in convergence and uncertainty measurement discussed in comparisons from previous literature, we apply the Bayesian shrinkage priors, especially the regularized Horseshoe prior, for variable selection on two epidemiological datasets. Because the medical resource allocation was historically related to the health region a patient belonged to, for the aortic stenosis data, we use a spatial model that estimated the regional differences, together with the regularized shrinkage prior applied to parameters' coefficients in the logistic regression model. Then, for the self-reporting HIV data, we introduce the fully Bayesian data imputation method into the model fitting to deal with the missing values by constructing a hierarchical model that simultaneously performs the imputation and estimation.

Performance of the Bayesian variable selection methods

Because the model selection methods in neither the frequentist nor Bayesian approach are feasible for the high-dimensional dataset, we mainly summarize the working mechanism of the Lasso family, the spike-and-slab priors and the shrinkage priors. According to the

comparison performed by Lu and Lou (2021) and Celeux et al. (2012), the spike-and-slab priors have higher specificity and RMSE than the Lasso methods. The shrinkage priors further avoid model overfitting and are more robust to the change of correlation structure and sparsity levels than the spike-and-slab priors. Moreover, the shrinkage priors are relatively computationally efficient among the high-dimensional Bayesian variable selection methods. As illustrated in Piironen and Vehtari (2017), among the shrinkage priors, the regularized Horseshoe prior outperforms in recovering the parameters with large values. Therefore, we choose to apply it for the variable selection purpose.

Aortic stenosis treatment data

In the first data analysis, we apply the regularized Horseshoe prior to the aortic stenosis dataset to investigate the potential inequalities caused by differences in health regions and social-material status on patients' treatment decisions. The regional information is modelled as random effects through a binary adjacency variance structure and added to the logistic regression that predicted the treatment. Estimates show that older patients with severe diseases were assigned to the new TAVR treatment, and so were patients who were female or lived in metropolitan areas. The results highlight us the necessity of resource allocation to remote health regions and patients deprived of material and social perspectives.

HIV self-reporting data

Next, we use the variable selection method to find the association between participants' living background, sexual behaviours and HIV infection with data collected from self-reporting HIV testing in South Africa. With the assumption that the missing mechanisms for all variables are ignorable, we assign a Beta-Bernoulli distribution to each binary missing variable. We integrate them out within the Bayesian hierarchical structure that fits the outcome logistic regression model. We identify a positive correlation between participants who finished the report in the Langa clinic, had sex with HIV-infected partners and

HIV status. Conversely, regular health check, obtaining a post-secondary degree and being female is related to a lower probability of HIV infection. Due to the limited number of HIV participants in some answer categories, though converged, the credible intervals of some predictors are wide. We solve this issue by combining categories or providing a reasonable interpretation of this phenomenon.

5.1 Future direction

Despite the advantage in interpretation and inference, the Bayesian methods suffer from extensive computation costs, especially in high-dimensional applications. Researchers have been working on the reduction of computation time. For example, INLA, the integrated nested Laplace approximation proposed by Rustand et al. (2021), is an alternative to other computation methods like the MCMC, with an advantage in speed and easy to apply in R. In the future, we will try to replace the MCMC with INLA to generate posterior samples to infer predictors on higher-dimension datasets.

Moreover, in the aortic stenosis data analysis, we focus on identifying the possible correlation between non-clinical factors and the treatment decision among all information provided by the administrative data. Aware of the existence of spatial information in the analysis, we add basic spatial models into the variable selection procedure. Nevertheless, the spatial effects can be modified to solve the problem caused by spatial confoundings. As mentioned in Section 3.4.2, in the future, we will combine the *Spatial+* model proposed by Dupont et al. (2022) with the shrinkage priors to reduce bias caused by the collinearity between predictors and spatial factors.

Finally, the results presented between the predictors and the outcome in this thesis are correlations, not causation. Our model can have more impact in clinical settings if we can establish causal relationships between variables. Different from the residuals in regression, the background factor, though obtained from the same equation, will better reflect reality if we have specific knowledge based on the data (Pearl, 2010). Besides, to make

causal assumptions, we will note their causal relationships, use a graphical model, and show causality through arrows with direction. Nevertheless, difficulty existed because we only have observations rather than experiments. Therefore, performing the counterfactual analysis, especially for one predictor, we will incorporate additional methods like the propensity score model.

Code Availability

The code for the two real data applications can be found at https://github.com/JanetteFu/Thesis_data_application. Each file contains the code used to construct model and a small artificial data for illustration. Follow the **Read me** file for instructions on the purpose and use of each file.

Bibliography

Zihang Lu and Wendy Lou. Bayesian approaches to variable selection: A comparative study from practical perspectives. *The International Journal of Biostatistics*, 17, 2021.

Zahir Noorie and Fatemeh Afsari. Sparse feature selection: Relevance, redundancy and locality structure preserving guided by pairwise constraints. *Applied Soft Computing*, 87, 2020.

Gilles Celeux, Mohammed El Anbariy, Jean-Michel Marinz, and Christian P. Robert. Regularization in regression: Comparing Bayesian and frequentist methods in a poorly informative situation. *Bayesian Analysis*, 7(2):477–502, 2012.

Juho Piironen and Aki Vehtari. Sparsity information and regularization in the Horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11:5018–5051, 2017.

Georg Heinze, Christine Wallisch, and Daniela Dunkler. Variable selection – a review and recommendations for the practicing statistician. *Biometrical Journal*, 60:431–449, 2018.

Hui Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68:49–67, 2006.

Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of American Statistical Association*, 90(430):773–795, 1995.

Andrew Gelman. Politics and COVID: Interesting example of aggregation and correlation. https://statmodeling.stat.columbia.edu/category/miscellaneous-statistics/, 2022. [Online; accessed 21-Apr-2022].

NCSS. Ridge regression. https://www.ncss.com/software/ncss/regression-analysis-in-ncss/#Ridge, 2022. [Online; accessed 03-March-2022].

PSU. Ridge regression. https://online.stat.psu.edu/stat508/lesson/5/5.1, 2022. [Online; accessed 12-Apr-2022].

Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B*, 67(2):301–320, 2005.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

Karthik Sridharan. Minimax rates, statistical learning and uniform convergence. http://www.cs.cornell.edu/courses/cs6783/2018sp/lec3.pdf, 2018. [Online; accessed 11-May-2022].

Anirban Bhattacharya, Debdeep Pati, Natesh S. Pillai, and David B. Dunson. Dirichlet-Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110 (512):1479–1490, 2015.

Anjali Agarwal. Bayesian variable selection with spike-and-slab priors. The Ohio State University, 2016.

Wessel Bruinsma. Spike and slab priors. https://Fwesselb.github.io/assets/write-up/, 2019. [Online; accessed 11-March-2022].

Veronika Ročková and Edward I. George. The spike-and-slab Lasso. *Journal of the American Statistical Association*, 113:431–444, 2015.

Gertraud Malsiner-Walli and Helga Wagner. Comparing spike and slab priors for Bayesian variable selection. *Austrian Journal of Statistics*, 40(4):241–264, 2011.

Hemant Ishwaran and J. Sunil Rao. Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005.

Veronika Ročková and Edward I. George. EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846, 2014.

George Park, Trevor Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. The Horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

Stéphanie van der Pas, Bas J. K. Kleijn, and Adrianus W. van der Vaart. The Horseshoe estimator: Posterior concentration around nearly black vectors. *Electron. J. Statist.*, 8(2): 2585–2618, 2014.

Peter J. Green, Krzysztof Łatuszyński, Marcelo Pereyra, and Christian P. Robert. Bayesian computation: A summary of the current state, and samples backwards and forwards. *Statistics and Computing*, 25:835–862, 2015.

Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Micheal I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50:5–43, 2003.

Development Team Stan. Stan reference manual. Version 2.29. https://mc-stan.org/docs/2_29/reference-manual/index.html, 2019. [Online; accessed 08-Sep-2021].

Radford M. Neal. MCMC using Hamiltonian dynamics. *ArXiv e-prints*, June 2012.

Matthew D. Hoffman and Andrew Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1593–1623, 2014.

John Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. 1991.

Vivekananda Roy. Convergence diagnostics for Markov chain Monte Carlo. Department of statistics, Iowa state University, 2019.

Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–511, 1992.

Stephen P. Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, 1998.

Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006.

Sumio Watanabe. Asymptotic equivalence of Bayes Cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571, 2010.

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

Sumio Watanabe. Equations of states in singular statistical estimation. Precision and Intelligence Laboratory, Tokyo Institute of Technology, 2018.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher,

Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jason Brownlee. A gentle introduction to k-fold Cross-validation. https://machinelearningmastery.com/k-fold-cross-validation/, 2018. [Online; accessed 11-March-2022].

Cosma Shalizi. Lecture 26: Variable selection. https://www.stat.cmu.edu/~cshalizi/mreg/15/, 2015. [Online; accessed 11-March-2022].

Malgorzata Roos, Thiago G. Martins, Leonhard Held, and Havard Rue. Sensitivity analysis for Bayesian hierarchical models. *Bayesian Analysis*, 10(2):321–349, 2015.

Paul Gustafson and Larry Wasserman. Local sensitivity diagnostics for Bayesian inference. *The Annals of Statistics*, 23(6):2135–2167, 1995.

Rob McCulloch. Local model influence. *Journal of the American Statistical Association*, 84 (406):473–478, 1989.

Mayo Clinic. Aortic valve stenosis. https://www.mayoclinic.org/diseases-conditions/aortic-stenosis/symptoms-causes/syc-20353139, 2021a. [Online; accessed 01.26.2022].

Medicine Johns Hopkins. Aortic valve replacement: open. https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/aortic-valve-replacement-open, 2022. [Online; accessed 02.14.2022].

Alain Cribier. Development of transcatheter aortic valve implantation (TAVI): A 20-year odyssey. *Archives of Cardiovascular Diseases*, 105(3):146–152, 2012.

Mayo Clinic. Transcatheter aortic valve replacement. https://www.mayoclinic.org/tests-procedures/transcatheter-aortic-valve-replacement/about/pac-20384698, 2021b. [Online; accessed 10.13.2021].

Anita Asgar, Pual Dorian, et al. National quality report: transcatheter aortic valve implantation. https://ccs.ca/publications/, 2019. [Online; accessed 16-Nov-2021].

Harindra C. Wijeysundera, Kayley A. Henning, Feng Qiu, et al. Inequity in access to transcatheter aortic valve replacement: A Pan-Canadian evaluation of wait-times. *Canadian Journal of Cardiology*, 36:844–851, 2019.

Abdulla A. Damluji, Michael Fabbro, et al. Transcatheter aortic valve replacement in low-population density areas. *Circulation: Cardiovascular Quality and Outcomes*, 13(8), 2020.

Ashwin S. Nathan, Lin Yang, Nancy Yang, Sameed Ahmed M. Khatana, et al. Socioeconomic and geographic characteristics of hospitals establishing transcatheter aortic valve replacement programs, 2012–2018. *Circulation: Cardiovascular Quality and Outcomes*, 14(11), 2021.

Ashwin S. Nathan, Lin Yang, Nancy Yang, et al. Racial, ethnic, and socioeconomic disparities in access to transcatheter aortic valve replacement within major metropolitan areas. *JAMA cardiology*, 7(2):150–157, 2022.

Carolle Bergeron. Évaluation de la cardiologie tertiaire au Québec. https://publications.msss.gouv.qc.ca/msss/fichiers/2008/08-906-01.pdf, 2019. [Online; accessed 12.01.2021].

Branimir K. Hackenberger. Bayes or not Bayes, is this the question? *Croatian medical journal*, 60(1):50–52, 2019.

Hude Quan, Vijaya Sundararajan, Patricia Halfon, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical Care*, 43(11), 2005.

Patricia Labelle, Denis Santerre, and Hélène St-Hilaire. Avis, Guides et normes, État de connaissances, État de pratique. https://www.inspq.qc.ca/sites/default/files/publications/2351_communaute_pratique_outil_pertinent_resume_connaissance.pdf, 2020.

Integrated University Health CIUSSS and Social Services Centres. Local community services centres (CLSCs). https://santemontreal.qc.ca/en/public/montreals-institutions-at-a-glance/clscs/, 2022. [Online; accessed 01.18.2022].

Ministère de la Santé et des Services sociaux. Québec health regions. https://www.msss.gouv.qc.ca/en/reseau/regions-sociosanitaires-du-quebec/, 2018. [Online; accessed 8-March-2022].

Philippe Gamache, Denis Hamel, and Christine Blaser. Material and social deprivation index. https://www.inspq.qc.ca/en/deprivation/material-and-social-deprivation-index, 2019. [Online; accessed 11.15.2021].

Robert Pampalon, Denis Hamel, Philippe Gamache, Mathieu D. Philibert, Guy Raymond, and André Simpson. An area-based material and social deprivation index for public health in Québec and Canada. *Canadian Journal of Public Health*, 103(Suppl.2):S17–S22, 2012.

Hude Quan, Bing Li, Chantal M. Couris, et al. Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *American Journal of Epidemiology*, 137(6), 2011.

Denis Hamel and Philippe Gamache. Assignment program user guide for the 2016 Canadian deprivation index. https://www.inspq.qc.ca/sites/default/files/publications/2639_material_social_deprivation_index.pdf, 2020. [Online; accessed 01.07.2022].

Sudipto Banerjee, Bradley P. Carlin, and Alan E. Gelfand. *Hierarchical modeling and analysis for spatial data (2nd ed.)*. Boca Raton, FL: CRC Press, 2015.

Jacob Goldstein-Greenwood. A brief on Brier scores. https://data.library.virginia.edu/a-brief-on-brier-scores/, 2021. [Online; accessed 23-Aug-2022].

Emiko Dupont, Simon N. Wood, and Nicole H. Augustin. Spatial+: A novel approach to spatial confounding. *Biometrics*, pages 1–12, 2022.

Victoire Michal, Leo Vanciu, and Alexandra M. Schmidt. A joint hierarchical model for the number of cases and deaths due to COVID-19 across the boroughs of montreal. *Spatial and Spatio-temporal Epidemiology*, (42):100518, 2022.

U.S. Department of Health & Human Services CDC. HIV basics. https://www.cdc.gov/hiv/basics/index.html, 2021. [Online; accessed 6-April-2022].

World Health Organization WHO and Joint United Nations Programme on HIV/AIDS UNAIDS. WHO, UNAIDS statement on HIV testing services: new opportunities and ongoing challenges. https://www.unaids.org/en/resources/presscentre/featurestories/2017/august/20170829_HIV-testing-services, 2017. [Online; accessed 14-March-2022].

Michael Strauss, Bruce Rhodes, and Gavin George. A qualitative analysis of the barriers and facilitators of HIV counselling and testing perceived by adolescents in South Africa. *BMC Health Serv Res*, 15(250), 2015.

Ricky Janssen, Nora Engel, Aliasgar Esmail, Suzette Oelofse, Anja Krumeich, Keertan Dheda, and Nitika Pai. Alone but supported: A qualitative study of an HIV self-testing app in an observational cohort study in South Africa. *AIDS and Behavior*, 24:467–474, 2019.

Susan Regan, Elena Losina, Senica Chetty, Janet Giddy, Rochelle P. Walensky, and et al. Factors associated with self-reported repeat HIV testing after a negative result in Durban, South Africa. *PLoS ONE*, 8(4), 2013.

Nitika Pai, Aliasgar Esmail, Paramita S. Chaudhuri, et al. Impact of a personalised, digital, HIV self-testing app-based program on linkages and new infections in the township populations of South Africa. *BMJ Global Health*, 6, 2021.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006.

Stef Van Buuren. *Flexible Imputation of Missing Data*. CRC Press, 2018.

Melissa J. Azur, Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf. Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1):40–49, 2011.

Zhihua Ma and Guanghui Chen. Bayesian methods for dealing with missing data problems. *Journal of the Korean Statistical Society*, 47:297–313, 2018.

Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan.*, volume 122. CRC Press, 2016.

Shirin Golchi, Jingyan Fu, Xiaoyang Liu, Eugene Yu, Reza Forghani, and Sahir Bhatnagar. Sparse Bayesian predictive modelling of tumour response using radiomic features. *Stats*, 11(1), 2022.

Denis Rustand, Janet van Niekerk, Havard Rue, Christophe Tournigand, Virginie Rondeau, and Laurent Briollais. Bayesian estimation of two-part joint models for a longitudinal semicontinuous biomarker and a rerminal event with R-INLA: Interests for cancer clinical trial evaluation, 2021. arXiv:2010.13704v2.

Judea Pearl. An introduction to causal inference. *The International Journal of Biostatistics*, 6 (2), 2010.