A Simulation Study to Evaluate Bayesian LASSO's Performance in Zero-Inflated Poisson (ZIP) Models

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon

By

Yue Dong

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Mathematics and Statistics
Room 142 McLean Hall
106 Wiggins Road
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5E6

Abstract

When modelling count data, it is possible to have excessive zeros in the data in many applications. My thesis concentrates on the variable selection in zero-inflated Poisson (ZIP) models. This thesis work is motivated by Brown et al. (2015), who considered the excessive amount of zero in their data structure and the site-specific random effects, and used Bayesian LASSO method for variable selection in their post-fire tree recruitment study in interior Alaska, USA and north Yukon, Canada. However, the above study has not carried out systematic simulation studies to evaluate Bayesian LASSO's performance under different scenarios. Therefore, my thesis conducts a series of simulation studies to evaluate Bayesian LASSO's performance with respect to different setting of some simulation factors.

My thesis considers three simulation factors: the number of subjects (N), the number of repeated measurements (R) and the true values of regression coefficients in the ZIP models. With different settings of the three factors, the proposed Bayesian LASSO's performance would be evaluated using three indicators: the sensitivity, the specificity and the exact fit rate. For applied practitioners, my thesis would be a useful example demonstrating under what circumstances one can expect Bayesian LASSO to have good performance in ZIP models. After sorting out the simulation results, we can find that Bayesian LASSO's performance is jointly affected by all the three simulation factors, while this method of variable selection is more reliable when the true coefficients are not close to zero.

My thesis also has some limitations. Primarily, with the time limitation of my thesis, it is impossible to consider all the factors that can potentially affect the simulation results, and using other penalty forms other than L_1 penalty is also left for future researchers to work on. Moreover, the current variable selection method is only for fixed effects selection while the variable selection for the mixed effect selection in ZIP models can be a direction for future work.

Acknowledgements

I want to express my sincere gratitude to my supervisor: Prof. Juxin Liu. With her help and mentorship, I not only finished my master thesis, but also learnt to be a more caring and responsible person. I can clearly remember each time we discuss my research in detail, when Prof. Liu always encourages me to think individually but provides me a lot of insightful suggestions. I also want to thank Prof. Liu for all her support when I had a family emergency – without her understanding, I could not be mentally strong enough to go through all the difficulties and concentrate on my study during these two years.

I also want to thank Prof. Longhai Li and Prof. Artur Sowa for being my committee members and providing me all the help during my master study. As a student who did not have enough statistical background when entering this program, I have been very lucky to have the professors who can give me patient guidance and answer my questions that can be quite basic. The past two years, though difficult for me academically, was very fruitful for me with the help and support from all the professors.

I then want to give my appreciation to all of the professors, graduate students, and staff in the Department of Mathematics and Statistics. Thank you so much for your help, advice, support and friendship. It has been a nice and wonderful time to work and cooperate with all of you.

I am also very thankful for all my family members and all my friends for providing me unconditional love everyday. For most of the time, we are separated in different countries but I believe our hearts are very close.

Lastly, thank you God for giving me all the strength to keep proceeding. All I have are the best gifts from you.

Contents

List of Tables

List of Figures

List of Abbreviations

LOF	List of Figures
LOT	List of Tables
LASSO	Least Absoluate Shrinkage and Selection Operator
ZIP	Zero Inflated Poisson
MCMC	Markov Chain Monte Carlo

Chapter 1

Introduction

1.1 Overview

When modelling count data, many applications may encounter an excess amount of zeros. In statistical literature, when an extra proportion of zeros is added to the proportion of zeros from the original discrete distribution, then the dataset is considered to have zero inflation(Van den Broek, 1995). For example, researchers found a considerable proportion of zeros in the distribution of motor vehicle crashes data (Lord et al., 2005); when health scientists examine the utilization of some patient services, they may find the a lot of patients report no utilization at all (Neelon et al., 2010). The dataset I use in my thesis is also an example with zero inflation: when examining the data of post-fire tree seedling recruitment, the responses variable is the count of new juvenile trees in each plot in the study, and it turns out that there is an excess amount of zeros. Therefore, compared with the models neglecting zero inflation, models considering zero-inflation is more suitable in this case for model estimation or variable selection (Lambert, 1992; Greene, 1994; Rose et al., 2006).

For modelling count data with zero inflation, zero-inflated and hurdle models are mostly commonly used. Among all these models, the zero-inflated Poisson (ZIP) model and zero-inflated negative binomial (ZINB) model are commonly applied. For the post-fire recruitment data set, the ZINB model does not over-perform the ZIP model in terms of model fitting (Brown et al., 2015). Therefore, the ZIP model is applied in my thesis for the tree recruitment dataset. In the ZIP model, zeros are modelled in two different processes: the first modelling process of zeros is the Poisson count model, and the second one is the logistic model of the zero inflation (Lambert, 1992).

For the regressions on the Poisson mean and the probability of extra zeros besides Poisson

count model, there are a lot of potential variables in our dataset. Therefore, how to select the important variables to be included in ZIP models is a question that my thesis tackles. There are a lot of existing variable selection approaches which I will discuss in the literature review chapter. Considering the zero inflation in the count data as well as the random effect, there are no other existing methods to accomplish variable selection. In my thesis Bayesian Least Absolute Shrinkage and Selection Operator (LASSO) (Park and Casella, 2008) is used for variable selection.

When conducting the simulations, there are some simulation factors that may potentially affect the performance of the Bayesian LASSO in variable selection. In the longitudinal study, if we change the number of subjects in the dataset, or change the number of repeated measurements within each subjects, the performance of the Bayesian LASSO might be different as well. Therefore, the **objective** of my thesis is to investigate how the proposed Bayesian LASSO method's performance may be affected by different simulation factors. In my thesis, I will consider several simulation factors (including the magnitudes of the true regression coefficients, the number of subjects and the number of repeated measurements within each subject). Based on the simulation results under all scenarios formed by the three factors, I will evaluate the proposed Bayesian LASSO's performance based on sensitivity, specificity and exact fit rate, which are the criteria Buu et al. (2011) used to evaluate LASSO's performance in the ZIP models.

1.2 Motivating Example

In this section, I will introduce the motivating example, which is the post-fire tree recruitment problem that Brown et al. (2015) discussed for the effects of prefire legacies and environmental factors on the trees' regeneration.

The data on variables in my thesis is the same as that considered by Brown et al. (2015). The data is collected from four regions of interior Alaska (USA) and northern Yukon (Canada) as shown in Figure 1.1. The study areas included were previously upland forested areas but attacked by wildfires in 2004 and 2005. Brown et al. (2015) specifically introduced the strategy for study sites selection. In short, the sites were selected based on the severity

levels of the wildfire, accessibility of the sites from existing road networks, elevation, and the availability of the site drainage before the fire (see Brown and Johnstone (2012) for the detailed site selection procedures; see Johnstone et al. (2009) for detailed study area map in interior Alaska, USA, and Brown and Johnstone (2011) for detailed study area map in northern Yukon, Canada).

In the data set, there are 55 sites, while within each site, there are five blocks of seedling treatments while each block contains four to six $0.50m \times 0.50m$ plots being randomly assigned to one of the seedling treatment (see Brown et al. (2015) for the detailed description of the seedling treatments); meanwhile, one or two plots without any treatments are unseeded control plots. The response variable represents the new juvenile tree seedlings in each plot by the time of data collection. The response variable does not simply follow a Poisson distribution since 62.98% of the observations are all zero when collapsing all the plots in each site, thus we have a notable zero-inflation in the data. Therefore, variable selection should consider the zero-inflation in the real data. Data is also collected on some environmental factors that will be introduced in Chapter 3 in detail.

Given that our dataset has a considerable number of variables, and merely from the biological background knowledge it is not sure which variables should be included in our models to explain both the probability of excessive zeros and the mean of the non-zero part of the model. Brown et al. (2015) proposed the Bayesian LASSO as the variable selection method to accommodate for both zero-inflation and hierarchy structure of the data. As a continued work of Brown's study, my thesis conducts a series of comprehensive simulations that to examine the performance of the proposed Bayesian LASSO under all scenarios formed by the three factors.

1.3 Outline of thesis

The remaining part of my thesis is organized as follows: in Chapter 2, I will give a literature review on the ZIP models and commonly used variable selection methods (in general and also in the ZIP models in particular). In Chapter 3, I will introduce the proposed Bayesian LASSO method for variable selection, and I will give a detailed description on my simulation

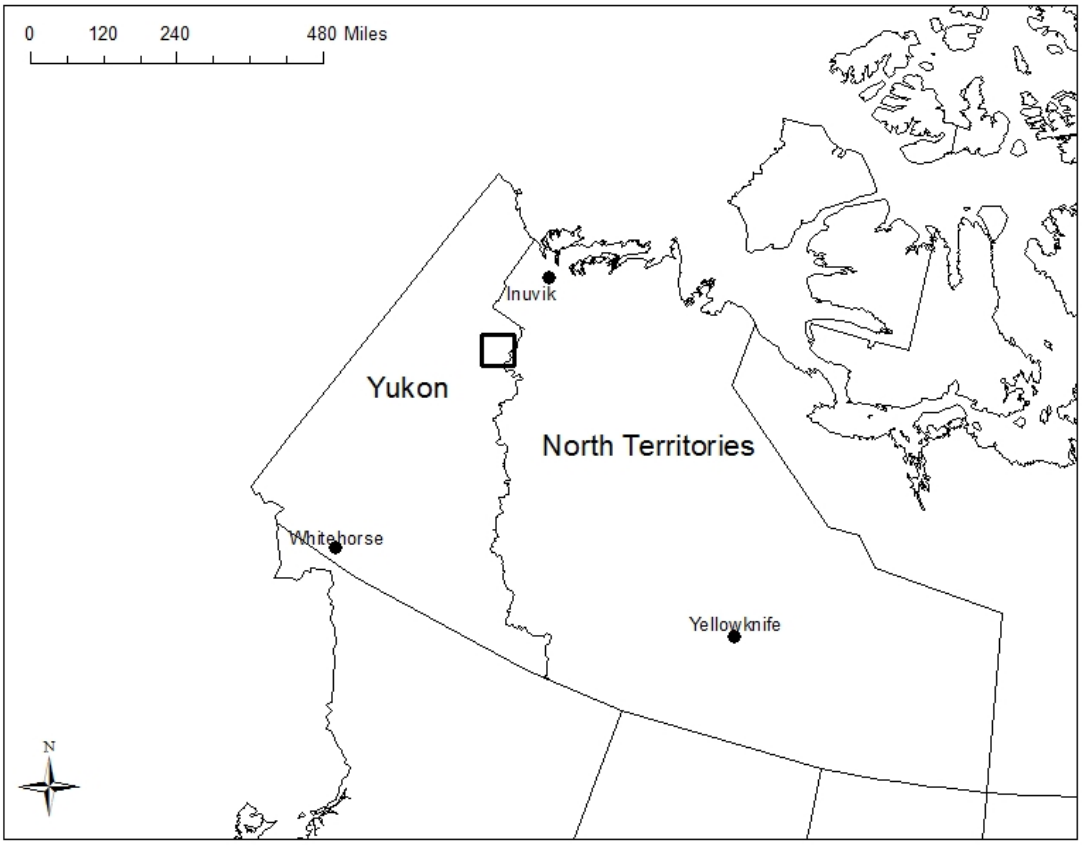


Figure 1.1: Study area in northern Yukon, Canada, with burned area in the rectangle

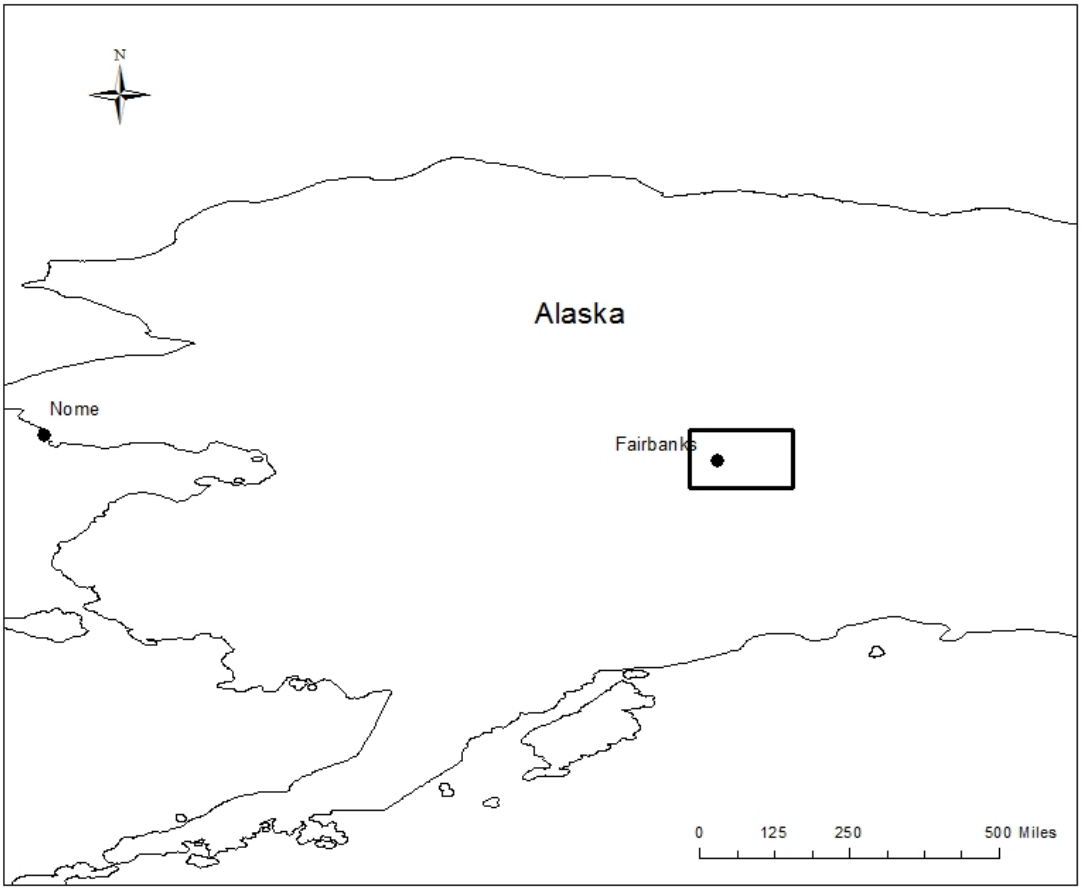


Figure 1.2: Study area in Interior Alaska, USA, with burned area in the rectangle

design. In Chapter 4, I will summarize the results of the simulations by plots and tables, and discuss how the simulation factors would affect Bayesian LASSO's performance. In the last chapter, Chapter 5, I will summarize the main findings of my thesis, together with the discussion of the possible improvement for future work.

Chapter 2

Literature Review

In this chapter, I will review the related literatures in the following aspects: Section 2.1 gives the literature review on the ZIP models; Sections 2.2 discusses the literatures on variable selection methods in general and also the methods in the ZIP models in particular.

2.1 Models for zero inflation

In many real world cases it is possible to have an exceeding amount of zeros when modelling count data. There are different kinds of models that can account for the zero inflation. For example, hurdle model is the kind of model which is composed by two parts: a point mass at zero and a truncated count distribution (e.g. Poisson distribution) at other non-zero points (Mullahy, 1986; Heilbron, 1989). There is another modelling method called zero-inflated count model which is composed by a point mass at zero and an untruncated count distribution (Lambert, 1992; Greene, 1994). When choosing the "best" model between the hurdle Poisson model and the zero inflated Poisson (ZIP) model, the decision should be based on model appropriateness according to the researchers' model assumptions (Rose et al., 2006). For example, unlike the hurdle model where both zero inflation and zero deflation can both be included, the ZIP model only allows for zero inflation (Neelon et al., 2010). As for the result of my reading, there is no universal rule that one of these models can always dominate others as for model fitting or prediction. In my case, I choose the ZIP model since when examining the data we found the proportion of zero is substantially high (thus considering the zero deflation case is not necessary); Moreover, my thesis is a continued work to examine the performance of Bayesian LASSO variable selection method that is proposed by Brown et al. (2015). I hence follow Brown's modelling strategy and focus on the ZIP model in my

thesis.

There are quite a lot studies on the ZIP models with applications in different areas. For example, Lambert (1992) firstly outlined the ZIP model and applied the model to manufacturing defects; Miaou (1994) compared different modelling strategies including Poisson model, ZIP model and negative binomial (NB) model for the road safety dataset; Bohara and Krieg (1996) used ZIP to model the frequency of migration and concluded that the ZIP model has better predicting performance while the model without properly considering zero inflation might lead to underprediction of new migrants; Desouhant et al. (1998) applied the ZIP model into their chestnut weevil dataset, and found that 25 out of 31 datasets could have good data fit; Ridout et al. (1998) provided a comprehensive review on zero-inflated models, and fitted a horticultural dataset to these possible models. Among all the studies regarding the ZIP models, both frequentist and Bayesian approaches have been investigated for model fitting. For example, among the non-Bayesian studies, Yau and Lee (2001) derived a ZIP regression model considering a random effect to fit longitudinal count data with extra zeros. Hall (2000) articulated both zero inflated Poisson model and zero inflated negative binomial model for their whitefly data, and compared the fitness of different models with or without a random effect. Besides these non-Bayesian studies, there are also some studies using Bayesian approaches alternatives to fit the zero inflated models. For example, Neelon et al. (2010) proposed a Bayesian method to fit the repeated measures data and compares competing models as for their performances in fitting the data. However, few papers have been published on variable selection for zero inflated models (Zeng et al., 2014).

2.2 Variable Selection Methods

In this section, the existing studies on the variable selection methods are organized into two categories: firstly, I will give a brief review for the variable selection methods in classical regression models; then the next subsection will review the variable selection methods particularly for the ZIP models.

2.2.1 Variable Selection Methods in Classical Regression Models

Some traditional variable selection methods, including the subset selection (Narendra and Fukunaga, 1977) and the stepwise regression (Hocking, 1976) methods, are easy to use but subject to their limitations. Firstly, when the number of predictors is large, the subset selection would become computationally infeasible (Zou, 2006). Secondly, the subset selection is proved to be lack of stability (Breiman, 1995) while the stepwise regression, as a substitute, starts from an initial model and ends when no single potential variable can improve the fit. However, the stepwise regression might be problematic in several aspects: the number of candidate predictors variables, the degree of correlation between the predictor variables and the order of parameter entry (or deletion) can all affect the variable selection results while increasing the sample size has little meaning in improving this method to select the correct variables (Derksen and Keselman, 1992). Moreover, Fan and Li (2001) pointed out that the stepwise regression and the subset selection also ignore the stochastic errors in the variable selection stage, thus it would be difficult to evaluate the sampling properties of the estimates from these two methods.

On the other hand, penalized likelihood approaches have the strength to overcome the problems mentioned above. Ridge regression (Hoerl and Kennard, 1970) and the least absolute absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) are the members as penalized likelihood approaches. Ridge and LASSO use different forms of the penalty functions (known as L_2 penalty and L_1 penalty respectively). Frank and Friedman (1993) also proposed a generalization of ridge and subset selection called bridge regression. Because of the time limitation of my thesis, only LASSO method and its L_1 penalty are investigated, and the comprehensive simulations for other penalty forms are still left for the future works.

As mentioned above, to overcome the shortcomings of the existing variable selection methods, aiming to improving prediction accuracy and generating interpretable estimated models, Tibshirani (1996) proposed the least absolute shrinkage and selection operator (LASSO) to simultaneously select the variables and estimate the coefficients for the variables. The LASSO estimator is defined as:

$$\hat{\boldsymbol{\theta}}_{lasso} = argmin(-logL(\boldsymbol{\theta}) + \lambda \sum_{d=1}^{p} |\theta_d|). \tag{2.1}$$

Here $\boldsymbol{\theta}$ is the set of the parameters that we are interested in. Therefore, the LASSO estimates are obtained by minimizing the negative log-likelihood function with a constraint. The term $\lambda \sum_{d=1}^{p} |\theta_d|$ is called the L_1 penalty, and λ is the tuning parameter which is larger than 0. If λ has some large values, more weight will be given to the L_1 penalty, and more coefficients will be shrunk to 0.

Now we can introduce the mathematical form of the ZIP model and see how it can be related to the LASSO method. When we consider ZIP models, the zeros are modelled by two components: we are interested in: the Poisson component and the zero inflation component. Therefore, people are interested in the mean of the Poisson component of the model λ_{ij}, and the probability of the extra zero p_{ij}. For the ith subject and the jth measurement, the models for λ_{ij} and p_{ij} can be defined as:

$$log(\lambda_{ij}) = \boldsymbol{x}_{1ij}^{'}\boldsymbol{\beta} + b_i, \tag{2.2}$$

$$logit(p_{ij}) = \boldsymbol{x}_{2ij}^{'}\boldsymbol{\alpha} + a_i, \tag{2.3}$$

$$y_{ij} \sim \begin{cases} 0 & \text{with probability } p_{ij}; \\ Poisson(\lambda_{ij}) & \text{with probability } 1 - p_{ij}. \end{cases} \tag{2.4}$$

In the above model, the dependent variable y_{ij} represents the response in the jth measurement within the ith subject. $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are the regression parameters, while b_i and a_i are the site-specific random effects. It can be seen from the above models that the variable sets for λ_{ij} and p_{ij} can be different, denoted by $\boldsymbol{x}_{1ij}^{'}$ and $\boldsymbol{x}_{2ij}^{'}$. Then $\boldsymbol{\theta}$ in the LASSO is defined as: $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$. In the above definition of the LASSO estimator, $L(\boldsymbol{\theta})$, as the likelihood function, is defined by:

$$\begin{aligned} L(\boldsymbol{\theta}) = \int \prod_{i=1}^{I} \int \prod_{j=1}^{J} \{ & \frac{u_{ij}}{1 + e^{\boldsymbol{x}_{2ij}^{'}\boldsymbol{\alpha} + a_i}} (e^{\boldsymbol{x}_{2ij}^{'}\boldsymbol{\alpha} + a_i} + exp(-e^{\boldsymbol{x}_{1ij}^{'}\boldsymbol{\beta} + b_i})) \\ & + (1 - u_{ij}) \frac{(e^{\boldsymbol{x}_{1ij}^{'}\boldsymbol{\beta} + b_i})^{y_{ij}} exp(-e^{\boldsymbol{x}_{1ij}^{'}\boldsymbol{\beta} + b_i})}{(1 + e^{\boldsymbol{x}_{2ij}^{'}\boldsymbol{\alpha} + a_i}) y_{ij}!} \} db_i da_i. \end{aligned} \tag{2.5}$$

The above is a brief overview of the LASSO variable selection method and how it can be applied into the background of ZIP models. However, the above LASSO method also has its own shortcomings, and its lack of oracle properties is one of the primary interests

in statistical literature. Fan and Li (2001) pointed out one obvious problem is that the LASSO does not have the oracle properties, which can be referred to the probability of selecting the right variables can converge to one and the estimates of the nonzero coefficients are asymptotically normal with the same means and covariances as if the zero coefficients were known in advance (Fan and Li, 2001; Zou, 2006). To deal with this problem, Fan and Li (2001) proposed smoothly clipped absolute deviation (SCAD) and showed it could enjoy oracle properties as long as the regularization parameters are properly chosen. Zou (2006) also proposed the adaptive LASSO (with an adaptive weight used in the L_1 penalty) and showed that the adaptive LASSO enjoys the oracle properties. The adaptive LASSO estimation, being similar to the LASSO estimation is defined as:

$$\hat{\boldsymbol{\theta}}_{lasso} = argmin(-logL(\boldsymbol{\theta}) + \lambda \sum_{d=1}^{p} \tau \left|\theta_d\right|). \tag{2.6}$$

Here τ is the adaptive weight. In recent studies, OLS estimators (Zou, 2006), LASSO estimators (Lian, 2012), the ratio of standard error of the OLS to the OLS coefficients (Qian and Yang, 2013), maximum likelihood estimators (Zeng et al., 2014), the ratio of the standard error of the maximum likelihood estimator to the ML estimator (Algamal and Lee, 2015) are the examples where different adaptive weights that were used to address LASSO's lack of oracle properties.

Another interest of statistical literature on the LASSO method is how to guarantee its consistency. Zhao and Yu (2006) made the assessment on LASSO's model selection consistency under linear models; they formalized the conditions for LASSO's consistency as strong and week irrepresentable conditions and showed LASSO's ability to select the true model given large or small numbers of potential variables. However, in my thesis, the theoretical properties of LASSO are not investigated, and this can be the focus of future works.

Besides the discussions for the oracle properties and consistency of LASSO, another issue with the ordinary LASSO is that it is difficult to give satisfactory standard errors while the Bayesian version of the LASSO can produce reliable standard errors (Xu et al., 2015). Park and Casella (2008) proposed a Bayesian model, used Gibbs sampler to implement the Bayesian LASSO and also provided the maximum likelihood estimates for the LASSO parameter λ. Some other LASSO's variants also exist to solve some other problems of the

ordinary LASSO (Hsu, 2015; Lim and Hastie, 2013; Yuan and Lin, 2006; Zou and Hastie, 2005). It is beyond this thesis's scope and thus won't be discussed in detail.

In recent years, there are quite a lot new literature that can be found about the LASSO related variable selection methods. Some of them are further methodology developments of existing LASSO methods. For example, LASSO is also integrated into geographically weighted regression (GWR) to make geographically weighted LASSO (GWL) because of the need to address the collinearity issue with GWR (Wheeler, 2009). Czarnota et al. (2015) compared the performance of GWR and GWL in a scenario with independent predictors and another scenario with correlated predictors. They found that when the predictors are correlated, compared with GWL, GWR might suffer more from regression coefficient sign reversal (i.e., reversal paradox). Another example is related to Bayesian variable selection approaches with spike and slab priors–mixture distributions of a point mass at 0 and a continuous distribution (Mitchell and Beauchamp, 1988). Xu et al. (2015) proposed a Bayesian group LASSO model with spike and slab priors for problems that only require variable selection at the group level.

There are also some recent studies on the application of the proposed LASSO methodologies in different areas. One example can be that Zeng et al. (2014) proposed to use adaptive LASSO for zero-inflated count data, and the authors found that the adaptive lasso worked well to identify the important variables. Mortier et al. (2015) also propose an application of adaptive lasso. Since the authors have specified several species groups with measurements taken at different times, there is a sum within the log-likelihood function in the penalty function, and thus the penalized log-likelihood function cannot be maximized analytically. Therefore, the authors use the Expectation-Maximization (EM) algorithm to do the optimization problem numerically.

Besides the LASSO method and its variants, there are also some new studies on other variable selection methods. There are the sequential method by Costa et al. (2015), the weighted quantile sum (WQS) by Carrico et al. (2015) and its application by Czarnota et al. (2015) and so forth. These novel variable selection methods might not be necessarily related to my thesis since their research backgrounds do not involve count data with excess amount of zeros. Therefore I won't go over each of the recent studies in details.

2.2.2 Variable Selection Methods for ZIP Models

In this section, I will briefly review variable selection methods for ZIP models. Unlike generalized linear models, zero-inflated count models are more complex in variable selection since different components of the model may have different explanatory variables (Zeng et al., 2014). However, there are also some studies applying the LASSO-type methods into ZIP models: Buu et al. (2011) applied LASSO into the area of substance abuse field, and respectively compare the variable selection results of Poisson regression with LASSO, Poisson regression with SCAD, ZIP with LASSO and ZIP with SCAD. Zeng et al. (2014) applied adaptive LASSO for both zero inflated Poisson and zero inflated Binomial models in the doctor visit dataset. Tang et al. (2014) combined Expectation-Maximization (EM) algorithm and adaptive LASSO penalty in selecting risk factors for their insurance modelling, and showed the proposed method has oracle properties. Moreover, they also mentioned that their variable selection result is not very good for the zero-inflation part. Wang et al. (2015) focused on the variable selection problem on zero-inflated negative binomial (ZINB) model, and propose an EM algorithm for the different penalties LASSO and SCAD respectively. According to my reading, there is no literature on existing variable selection method for ZIP models with longitudinal count data which involves random effects.

By doing this literature review, we can see that even though there is a systematic development of the variable selection methods and miscellaneous studies on the zero inflated models, literature on variable selection for zero inflated models, especially for the cases with longitudinal count data is still scant. Brown et al. (2015)proposed a Bayesian LASSO variable selection method for longitudinal count data with random effects. Therefore, my thesis's goal is to conduct a series of simulation studies to test the performance of the proposed Bayesian LASSO variable selection method.

Chapter 3

Proposed Method for Variable Selection in ZIP Models

This chapter introduces the models and the Bayesian LASSO variable selection method.

3.1 Notation and Model

As it has been mentioned in Chapter 2, in ZIP models zeros are modelled in two components: the first one is the Poisson component and the other is the zero inflation component. Therefore, when we consider ZIP models, we are interested in: the mean of the Poisson component of the model λ_{ij}, and the probability of the extra zero p_{ij}. In Section 2.2, the mathematical form of ZIP models has already been introduced. In this chapter, I will elaborate the model based on the concrete data example of my thesis.

Section 1.2 has already introduced my thesis's research background: there are 55 sites; the data is collected for site 1 to site 39 with 15 plots for each of site, while site 40 to site 55 have attained 20 plots for each of them. Therefore, if I define i as the site index and j as the plot index, the general ZIP model defined from Equation 2.2 to 2.4 can be specified as:

$$\begin{aligned} log(\lambda_{ij}) = \beta_1 + \beta_2 Moist_i + \beta_3 Latitude_i + \beta_4 Elevation_i + \beta_5 BS.Sown_{ij} \\ +\beta_6 BAstdg_i + \beta_7 TSLF_i + \beta_8 Resid.org_i + \beta_9 BS.nstand_i + b_i, \end{aligned} \tag{3.1}$$

$$\begin{aligned} logit(p_{ij}) = \alpha_1 + \alpha_2 Moist_i + \alpha_3 Latitude_i + \alpha_4 Elevation_i + \alpha_5 BS.Sown_{ij} \\ +\alpha_6 BAstdg_i + \alpha_7 TSLF_i + \alpha_8 Resid.org_i + \alpha_9 BS.nstand_i + a_i, \end{aligned} \tag{3.2}$$

$$y_{ij} \sim \begin{cases} 0 & \text{with probability } p_{ij}; \\ Poisson(\lambda_{ij}) & \text{with probability } 1 - p_{ij}. \end{cases} \tag{3.3}$$

In the above model, the dependent variable y_{ij} represents the count of new juvenile tree seedlings in the jth plot of the ith site. Table 3.1 gives a description of the variables. Please note that besides the plot-level variable BS.Sown, all the other variables are measured at site level. β_i and α_i $(i = 1, 2, 3 \ldots, 8)$ are the regression parameters, while b_i and a_i are the site-specific random effects. In Chapter 4, a systematic simulation study is conducted to evaluate the performance of the proposed Bayesian LASSO method.

Table 3.1: Variable Descriptions

Variable	Variable Description
Moist	Ranking of the site moisture potential
Latitude	Latitude of the plot in degree based on GPS readings
Elevation	Elevation of the plot in meter based on GPS readings
BS.Sown	Binary variable indicating whether a plot was sown with black spruce seed
BAstdg	The area occupied by trees in each site; measured by square meter per hectare
TSLF	Estimated time since last fire
Resid.org	Mean residual organic layer depth of each site
BS.nstand	Ranking of the distance to the nearest stand of black spruce (from 1 to 8)

It is worth mentioning that even though the above ZIP model is specified based on the concrete example of my thesis, the model can still be generalized for longitudinal data's multi-level data structure, which has a certain number of subjects containing repeated measurements.

3.2 Bayesian LASSO

As it has been introduced in Chapter 2, considering that $\boldsymbol{\theta}$ is the set of the regression coefficients in the aforementioned ZIP models, and $\boldsymbol{\theta}$ is defined as: $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$. The LASSO of

Tibshirani (1996) achieves the LASSO estimators by:

$$\hat{\boldsymbol{\theta}}_{lasso} = argmin(-logL(\boldsymbol{\theta}) + \lambda \sum_{d=1}^{p} |\theta_d|). \tag{3.4}$$

where $\lambda \geq 0$ determines the shrinkage amount: when λ is 0, the LASSO estimator $\hat{\boldsymbol{\theta}}_{lasso}$ is identical with $\hat{\boldsymbol{\theta}}_{MLE}$; while when λ is sufficiently large, $\hat{\boldsymbol{\theta}}_{lasso}$ shrinks to zero. Under the assumptions of the independence of the subjects and the conditional independence of the repeated measures, $L(\boldsymbol{\theta})$, as the likelihood function, is defined by:

$$\begin{aligned} L(\boldsymbol{\theta}) = \int \prod_{i=1}^{I} \int \prod_{j=1}^{J} \{ & \frac{u_{ij}}{1 + e^{\boldsymbol{x}_{2ij}^{'}\boldsymbol{\alpha} + a_i}} (e^{\boldsymbol{x}_{2ij}^{'}\boldsymbol{\alpha} + a_i} + exp(-e^{\boldsymbol{x}_{1ij}^{'}\boldsymbol{\beta} + b_i})) \\ & + (1 - u_{ij}) \frac{(e^{\boldsymbol{x}_{1ij}^{'}\boldsymbol{\beta} + b_i})^{y_{ij}} exp(-e^{\boldsymbol{x}_{1ij}^{'}\boldsymbol{\beta} + b_i})}{(1 + e^{\boldsymbol{x}_{2ij}^{'}\boldsymbol{\alpha} + a_i}) y_{ij}!} \} db_i da_i. \end{aligned} \tag{3.5}$$

Tibshirani (1996) suggested that LASSO estimates can be viewed as posterior mode estimates when the regression parameters have independent and identical Laplace (i.e., double-exponential) priors, and there are several studies subsequently proposed using Laplace-like priors (Figueiredo, 2003; Bae and Mallick, 2004; Yuan and Lin, 2006). In my thesis, the distributions of site-specific random effects b_i and a_i respectively follow $N(0, \sigma_a)$ and $N(0, \sigma_b)$, while $\sigma_a, \sigma_b \sim Unif(0.001, 10)$. Moreover, $\lambda \sim Unif(0, 100)$. I follow Park and Casella (2008) and consider the unconditional prior for $\boldsymbol{\beta}$ as a Laplace distribution with the scale parameter λ:

$$\pi(\boldsymbol{\beta}) = \prod_{d=1}^{p} \frac{\lambda}{2} e^{-\lambda|\beta_d|}. \tag{3.6}$$

Similarly, the conditional Laplace prior for $\boldsymbol{\alpha}$ is:

$$\pi(\boldsymbol{\alpha}) = \prod_{d=1}^{p} \frac{\lambda}{2} e^{-\lambda|\alpha_d|}. \tag{3.7}$$

I used Rjags (Plummer, 2013) to conduct the Bayesian Markov Chain Monte Carlo (MCMC) samples from the posterior distributions for both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. The MCMC samples enable us to obtain the 2.5% and 97.5% posterior quantiles of the samples, thus we can also get 95% credible intervals for both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. For a single coefficient β_j or α_j, if its 95% credible interval contains zero, it means that the variable associated with this coefficient is not selected implied by Bayesian LASSO. On the contrary, if the 95% credible interval does

not contain zero, it means that the associated variable is selected. I used R (3.1.3) (Team, 2014) for all the statistical summary.

Chapter 4

Simulation Results

In this chapter, I will introduce the simulation design and the main findings from my simulation results.

4.1 Simulation Description

In this section, for the purpose of simplification, I use n to denote the number of subjects in the original dataset, while use r to denote the number of repeated measurements in the original dataset.

4.1.1 Simulation Setting

To generate different simulation scenarios, I consider the simulation factors that are defined as follows:

(1) Number of the subjects: $N = n, 2n, 5n$

As mentioned in section 1.2, in the original dataset, there are 55 subjects named "sites", and a natural question is whether increasing the number of subjects can affect the performance of the variable selection method. To increase the number of subjects, I expand the number of subjects (originally 55) by twice and five times.

To be more specific, when expanding the total number of subjects by twice, the original dataset is duplicated and a new subject (site) index from 1 to 110 is created. The dataset is duplicated twice, thus the new number of observations in the dataset is 1810. Similarly, when expanding the total number of subjects by five times, the original dataset is duplicated for five times (except the response variable). For the simplicity of notation, I use n to denote

the original number of subjects from the dataset, and use $2n$ and $5n$ to represent the cases when the site index is expanded by twice and five times.

(2) Number of repeated measurements: $R = \frac{1}{2}r, r, 2r$

The number of repeated measurements nested within each subject is often of interest and thus is considered as another simulation factor. In the real dataset, the data is collected for site 1 to site 39 with 15 repeated measurements for each of site, while site 40 to site 55 have attained 20 repeated measurements for each of them.

For a choice of fewer number of repeated measurements, I decrease the measurement numbers from 15 to 7 for subject 1 to subject 39; while the subject 40 to subject 55 have decreased in their repeated measurement from 20 to 10. It is safe to decrease number of repeated measurements in this way since all the variables except for BS.sown are site-level, which means they are all identical values within the same site. On the contrary, for the scenarios with larger number of repeated measurement, I double the repeated measurement for each subject. For the simplicity of the notation, the number of original repeated measurements is denoted by r, while the lower level of repeated measurements is denoted by $r/2$ (even though it is not exactly half of the original repeated measurements), and the higher level is denoted by 2r.

(3) True regression coefficients:

I also vary the true regression coefficients' absolute values to check if the variable selection method will perform differently while the effect size of variables changes. The larger the regression coefficient's distance to zero is, the larger effect that the corresponding variable has. By varying the absolute values of the regression coefficients, we can conclude when the Bayesian LASSO would have the better performance, and how the other two simulation factors will work under different true coefficients. In my thesis, I have the below settings that the true values' distances to zero varying from 0.1 to 1. It is worth mentioning that the true coefficients are set to have the same absolute values under each case for the purpose of simplicity.

Therefore, by controlling the subject number, the number of repeated measurements, and the variable effect size, there are 27 simulation scenarios in total. I will elaborate the findings in Subsection 4.2.

Table 4.1: True Coefficients

variable Effect Size	$\boldsymbol{\beta}$	$\boldsymbol{\alpha}$
Small	$(0.1, -0.1, 0.1, 0.1, 0, 0, 0, 0)$	$(0, 0, 0, 0, 0.1, -0.1, 0.1, 0.1)$
Medium	$(0.5, -0.5, 0.5, 0.5, 0, 0, 0, 0)$	$(0, 0, 0, 0, 0.5, -0.5, 0.5, 0.5)$
Large	$(1, -1, 1, 1, 0, 0, 0, 0)$	$(0, 0, 0, 0, 1, -1, 1, 1)$

4.1.2 Simulation Design

To evaluate the performance of the proposed Bayesian LASSO variable selection method, I conduct a series of comprehensive simulation studies. The simulation studies are carried out in the following steps:

(1) Set the true values for the model coefficients ($\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$) in the simulation studies. As it has been shown in Table 1.1, there are eight variables in my dataset. Therefore, in the first step, some coefficients are set to be non-zero, meaning that the corresponding variables are included to generate the response variable; the other variables are set to have zero coefficients, which means that these variables are absent from the model. Besides, the intercept terms β_1 and α_1 are respectively set to be -0.5 and 0.2, while the standard deviations for b_i and a_i, denoted by σ_b and σ_a are respectively 0.2 and 0.1.

(2) Generate the count response variable based on the true values of the coefficients, intercepts and the standard deviations of the random effects in step (1).

(3) Use the generated count data together with the variable data, run a Bayesian MCMC for the zero inflation component and Poisson component of the ZIP model. The interval estimates based on the posterior samples of the parameters are used to select the variables; that is, if a credible interval does not cover zero, then the corresponding variable is selected.

(4) Repeat steps (2) and (3) for 200 times and evaluate the performance of Bayesian LASSO method in terms of sensitivity, specificity and exact fit rate (see Section 4.1.3 for more detail).

In step 2, the response variable is generated using the equations 3.1 and 3.2. More detail for the notations for the variables can be referred to section 3.1. The random effects are generated respectively as $b_i \sim N(0, \sigma_b^2)$ and $a_i \sim N(0, \sigma_a^2)$. It is worth mentioning that the

random effects are only specified at the site level. The site-specific random effects can account for the heterogeneity among different sites and the dependence among the plots within each site (Brown et al., 2015). In step 3, as it has been mentioned in Section 3.2, to obtain the Bayesian LASSO estimates, we impose independent double-exponential priors for all β_j and α_j. The overall procedure of the simulation study can be summarized by Figure 4.1.

4.1.3 Evaluation Criteria

For each simulation scenario, the performance of the Bayesian LASSO method is evaluated based on several criteria. I follow Buu et al. (2011) and use the following criteria:

(1) Specificity: specificity is defined as the proportion of zero coefficients that have credible intervals covering zero (the corresponding variables are not selected).

(2) Sensitivity: sensitivity is defined as the proportion of nonzero coefficients that have credible intervals not covering zero (the corresponding variables are selected).

(3) Exact fit: exact fit is defined as the probability of a replication selecting the exact sub-model among the 200 replications at one simulation scenario. For example, at a single simulation scenario, if n of the 200 replications are found to select exactly the variables that are chosen in the step (1) of the simulation procedure, the exact fit value is $\frac{n}{200}$.

Based on the above criteria, the simulation results under each simulation scenario will be evaluated, and I will summarize how the simulation factors can influence the performance of the Bayesian LASSO. Since my thesis concentrates on the variable selection for ZIP models, the simulation results will be reported respectively for the Poisson component and the zero component.

It is worth mentioning that I expect the exact fit rate cannot be quantitatively larger than either of the sensitivity and the specificity. The reason is quite straightforward if we look at the mathematical forms of these criteria. To be more specific, if we only consider the Poisson component (considering the zero inflation component will give the same conclusion as well), the sensitivity, the proportion of non-zero coefficients having credible intervals not covering zero, is mathematically defined as:

$$Sensitivity = P(\hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5); \tag{4.1}$$

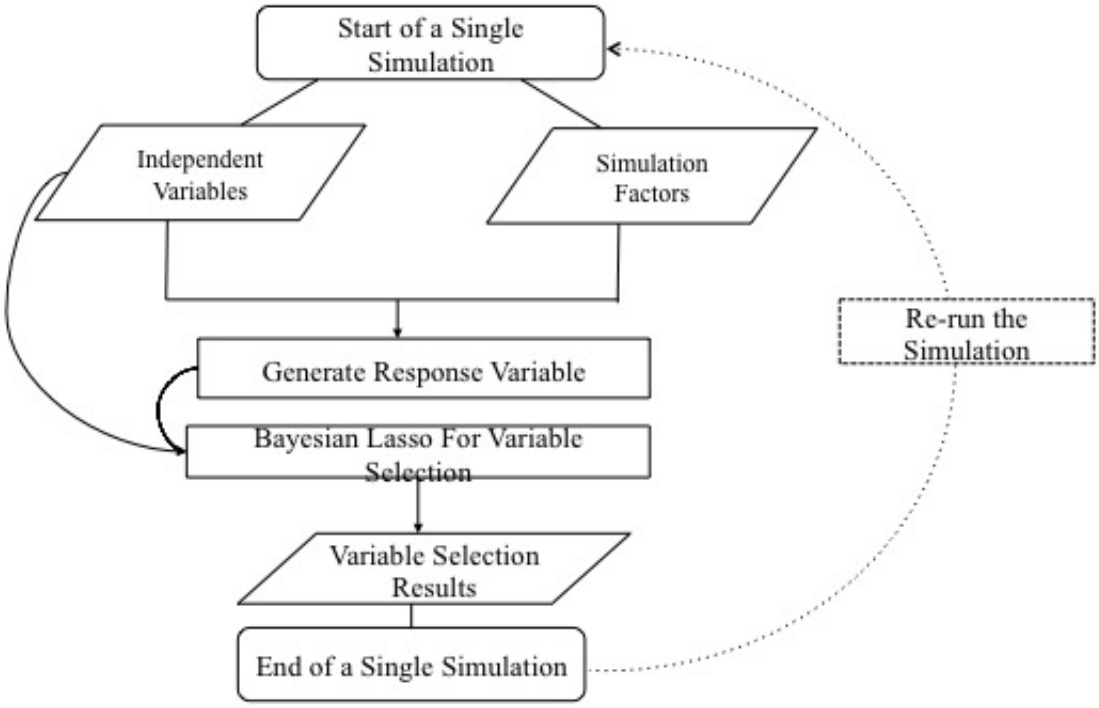


Figure 4.1: The Procedure for Each Simulation Scenario (the data set is repeated generated for 200 times under each simulation scenario)

In the above definition, $\hat{\beta}_2$ to $\hat{\beta}_5$ are respectively the posterior estimates for β_2 to β_5. Similarly, the specificity, which is the proportion of zero coefficients having credible intervals covering zero, can be defined as:

$$Specificity = P(\hat{\beta}_6, \hat{\beta}_7, \hat{\beta}_8, \hat{\beta}_9); \tag{4.2}$$

Again $\hat{\beta}_6$ to $\hat{\beta}_9$ are the posterior estimates for β_6 to β_9.

On the other hand, the exact fit rate, which is defined as the probability of choosing the exact sub-model in the 200 runs, can be mathematically defined as:

$$ExactFitRate = P(\hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6, \hat{\beta}_7, \hat{\beta}_8, \hat{\beta}_9); \tag{4.3}$$

In an ideal case where $(\hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5)$ and $(\hat{\beta}_6, \hat{\beta}_7, \hat{\beta}_8, \hat{\beta}_9)$ are independent, the exact fit rate, which can be represented by Equation 4.3, is the product of the sensitivity and the specificity, which are defined in Equation 4.1 and Equation 4.2. Given that both the sensitivity and the specificity are quantitatively between 0 and 1, the exact fit rate cannot be larger than any of the specificity and the sensitivity.

However, there is no guarantee for the independence of $(\hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5)$ and $(\hat{\beta}_6, \hat{\beta}_7, \hat{\beta}_8, \hat{\beta}_9)$. In this sense, the exact fit rate cannot quantitatively exceed either of the sensitivity and the specificity.

4.2 Simulation Results

In this subsection, I will summarize the simulation results and point out the findings that may be helpful for designing the experiments for future data collection. As it has been mentioned in the Section 4.1.1, there are three simulation factors I have taken into consideration: the number of subjects, the number of repeated measurements, and the magnitude of the regression coefficients. In this subsection, I will firstly summarize if any of the above three factors alone, or some of these factors jointly, can have influence on the performance of the Bayesian LASSO variable selection approach; if the factors' influence can be found, I will also elaborate how the influence would be.

Table 4.2 and Table 4.3 numerically summarize the simulation results respectively for the Poisson component and the zero-inflation component. In these two tables, each case can

be labeled by three simulation factors in the following order: the true regression coefficients ranging from 0.1 to 1, the number of repeated measurements: r/2, r, and 2r, and the number of subjects n, 2n, and 5n.

By looking at Table 4.2 for the Poisson component, we can have a general understanding of how the three simulation factors can affect the performance of the Bayesian LASSO for the Poisson component. First of all, the exact fit rate and the sensitivity all indicate better performances of the Bayesian LASSO when the coefficients are larger; thus we can say it is easier to select the "important" variables who have larger magnitudes of the true regression coefficients. To be more specific, we can fix the other two simulation factors to be at any level while increase the regression coefficients from 0.1 to 1: the sensitivity will be increased or at least unchanged. However, increasing the regression coefficients does not always increase the exact fit rate, especially when the exact fit rate is already around or larger than 0.9 before increasing the regression coefficients. For example, fixing the number of subject to be $2n$ and the number of repeated measurements to be $2r$, the highest exact fit rate happens when the magnitude of the regression coefficients are 0.5, instead of 1 (the largest exact fit rate does not happen at the largest magnitude of regression coefficients in this case). The specificity is always attained at a quite high level (larger than 0.98). Therefore, larger magnitudes of the true coefficients can promote Bayesian LASSO's performance in terms of the sensitivity; while the specificity remains high and the influence resulted from changing the true coefficients is not obvious. Moreover, increasing the magnitudes of the true coefficients can increase the exact fit rate while there can be some exceptions when the exact fit rate is already close to or larger than 0.9.

Secondly, increasing the number of repeated measurements has its effect on improving the variable selection performance, but its effect is highly dependent on other two simulation factors. The positive effect of the number of repeated measurement only present when evaluated by sensitivity while not exact fit rate. For example, when we fix the true coefficients to be 0.5 and the number of subjects to be $5n$, the highest exact fit rate is not achieved at the highest level of number of repeated measurements $2r$; Instead, the highest exact fit rate is achieved when the number of repeated measurement is $r/2$ (while the sensitivity remains unchanged to be 1). There are also some other examples that can be seen from Table 4.2

Table 4.2: Simulation results for the Poisson Component (with respective levels of true coefficients, R, and N in the brackets)

	Exact Fit	Specificity	Sensitivity
Case (0.1,r,n)	0	1	0.003
Case (0.5,r,n)	0.460	0.986	0.786
Case (1,r,n)	0.955	0.988	1
Case (0.1,r,2n)	0	0.998	0.009
Case (0.5,r,2n)	0.875	0.975	0.989
Case (1,r,2n)	0.910	0.971	1
Case (0.1,r,5n)	0	0.996	0.096
Case (0.5,r,5n)	0.910	0.976	1
Case (1,r,5n)	0.885	0.966	1
Case (0.1,r/2,n)	0	1	0.001
Case (0.5,r/2,n)	0.045	0.968	0.383
Case (1,r/2,n)	0.895	0.974	0.999
Case (0.1,r/2,2n)	0	1	0
Case (0.5,r/2,2n)	0.500	0.978	0.814
Case (1,r/2,2n)	0.910	0.975	1
Case (0.1,r/2,5n)	0	1	0.015
Case (0.5,r/2,5n)	0.920	0.978	1
Case (1,r/2,5n)	0.855	0.958	1
Case (0.1,2r,n)	0	0.998	0.006
Case (0.5,2r,n)	0.855	0.985	0.975
Case (1,2r,n)	0.855	0.963	1
Case (0.1,2r,2n)	0	0.995	0.074
Case (0.5,2r,2n)	0.890	0.971	1
Case (1,2r,2n)	0.855	0.960	1
Case (0.1,2r,5n)	0.030	0.995	0.296
Case (0.5,2r,5n)	0.890	0.968	1
Case (1,2r,5n)	0.890	0.970	1

showing that higher levels of repeated measurements not necessarily lead to higher exact fit rate, while the sensitivity can indeed increase or at least stay unchanged with more repeated measurements. Moreover, it can be clearly indicated by Table 4.2 that setting the repeated measurement to be at its lowest level $r/2$ can still yield the sensitivity to be close to 1 when the true coefficients and the number of subjects are at their higher levels. However, if we set the number of repeated measurements to be at its lowest level r/2 while the true coefficients to be 0.1, the exact fit rate and the sensitivity will be very close to zero. In this case, increasing the number of subjects from n to $5n$ can increase the sensitivity from 0.001 to 0.015 while the exact fit rate is always 0.

Thirdly, changing the number of subjects also has the effect to improve both the exact fit rate and the sensitivity when fixing the other two simulation factors (For example, fixing R to be r and fixing the true regression coefficients to be 0.5; fixing R to be $r/2$ and the true regression coefficients to be 0.5; fixing R to be $2r$ and the true regression coefficients to be 0.1, 0.5 or 1). However, the positive effect of increasing the number of subjects also encounters some exceptions when we look at the exact fit rate. For example, fixing the number of repeated measurement to be $r/2$ and the true coefficients to be 1, the highest exact fit rate occurs when the number of subjects is $2n$ rather than $5n$; again in this case if we look at the the sensitivity, we can see that increasing the number of subjects can increase the sensitivity until it reaches 1 without any decrease in the sensitivity.

When we compare Table 4.2 and Table 4.3, it is can be seen that the Bayesian LASSO variable selection method performs better for the Poisson component compared with the zero-inflation component. With the same settings of the three simulation factors, the results for the Poisson component have higher sensitivity in all the cases compared with for the zero-inflation component, although specificity for the zero-inflation component also remains close to 1. For the three simulation factors, the findings about the zero-inflation component are similar to those about the Poisson component. Now with a basic and general understanding of the roles of our simulation factors, I will elaborate the findings below in detail.

Table 4.3: Simulation results for the Zero-Inflation Component (with respective levels of true coefficients, R, and N in the brackets)

	Exact Fit	Specificity	Sensitivity
Case (0.1,r,n)	0	1	0
Case (0.5,r,n)	0.005	0.989	0.341
Case (1,r,n)	0.650	0.994	0.895
Case (0.1,r,2n)	0	1	0.003
Case (0.5,r,2n)	0.170	0.981	0.680
Case (1,r,2n)	0.945	0.988	0.999
Case (0.1,r,5n)	0	0.998	0.014
Case (0.5,r,5n)	0.830	0.971	0.980
Case (1,r,5n)	0.905	0.974	1
Case (0.1,r/2,n)	0	1	0
Case (0.5,r/2,n)	0	0.996	0.053
Case (1,r/2,n)	0.075	0.994	0.583
Case (0.1,r/2,2n)	0	1	0
Case (0.5,r/2,2n)	0.005	0.985	0.363
Case (1,r/2,2n)	0.625	0.986	0.895
Case (0.1,r/2,5n)	0	1	0.003
Case (0.5,r/2,5n)	0.345	0.984	0.781
Case (1,r/2,5n)	0.905	0.973	1
Case (0.1,2r,n)	0	1	0.001
Case (0.5,2r,n)	0.175	0.986	0.654
Case (1,2r,n)	0.905	0.980	0.995
Case (0.1,2r,2n)	0	0.999	0.008
Case (0.5,2r,2n)	0.730	0.981	0.935
Case (1,2r,2n)	0.865	0.958	1
Case (0.1,2r,5n)	0	0.989	0.079
Case (0.5,2r,5n)	0.925	0.980	1
Case (1,2r,5n)	0.930	0.981	1

4.2.1 Sensitivity

Sensitivity for the Poisson Component

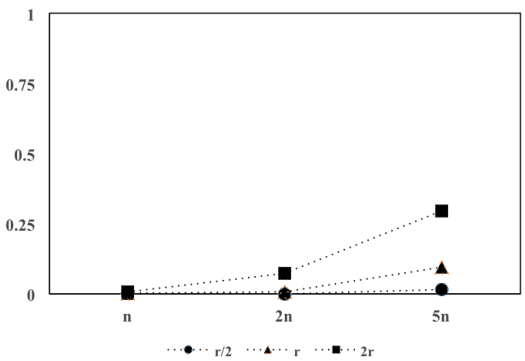
As it has been mentioned in Subsection 4.1.3, sensitivity is defined as the proportion of non-zero coefficients that have credible intervals not covering zero. Therefore, sensitivity is an indicator that represents the variable selection method's ability to correctly identify the variables that should be included in the model. Summarizing the results with the sensitivity, I construct Figure 4.2 to see how each simulation factor can affect the results without averaging out the other two factors. The sensitivity for the Poisson component is always 1 when the true coefficients are set to be 1, no matter how the number of repeated measurements or the number of subjects are changed. Therefore, only fixing the true coefficients to be at their lower levels (especially at 0.5) will allow us to observe more changes in the sensitivity. In other words, when the data collectors know the magnitudes of the true coefficients are large, they do not necessarily need large numbers of subjects or repeated measurements and the variable selection for Poisson regression component would not be greatly affected regarding sensitivity.

On the other hand, if the data collectors have the scientific background knowledge that the magnitudes of the true coefficients are small, they will need to collect data for more more subjects with more repeated measurements; otherwise the variable selection process would probably be unable to pick up the variables that should be included in the model. Figure 4.2 clearly indicates that increasing the number of subjects can lead to higher sensitivity when the true coefficients are 0.1 or 0.5. Moreover, when the true coefficients are 0.5, setting the number of subjects to $2n$ or $5n$ can lead the sensitivity to be close to or even reach 1.

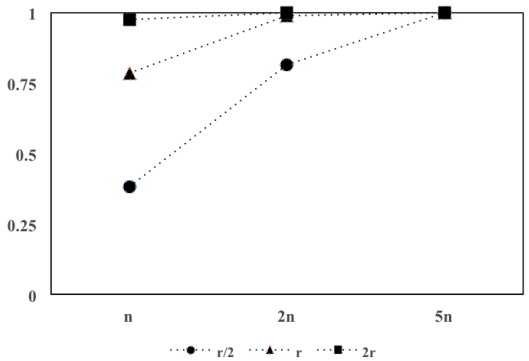
From my simulation results, when the true coefficients' absolute values are 0.1 (small effect size), the sensitivity remains close to zero if only one of the other two simulation factors is changed. The only way to increase the sensitivity to be around 0.3 is to set both of the other two simulation factors at their highest levels. Because of the time limitation of my thesis, I did not run the simulation to test whether keeping the number of subjects N and the number of repeated measurements R further increasing can make the sensitivity grow to 1. However, based on the tendency as Figure 4.2.A shows, increasing both of the simulation factors at

the same time could lead to large increase in the sensitivity if the sensitivity is not already close to 1 before increasing N and R.

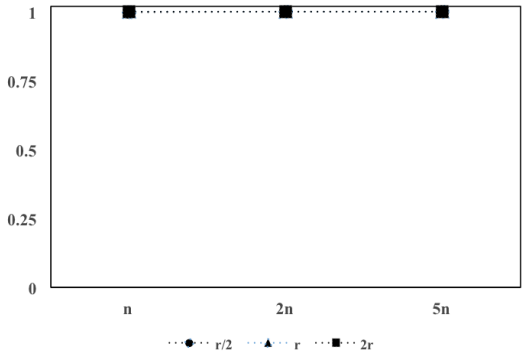
Figure 4.2 also shows the effects from changing the number of repeated measurements under different magnitude of regression coefficients. In Figure 4.2, each line stands for one level of repeated measurements, and the lines representing higher level of repeated measurements are above those representing lower levels when the true coefficients are 0.1 or 0.5 (shown in Figure 4.2.A and Figure 4.2.B). Moreover, as clearly indicated by Figure 4.2.B, when the true coefficients are 0.5 and the number of subject is at its smallest value n, increasing the number of repeated measurements from $r/2$ to $2r$ can cause the sensitivity to increase drastically from 0.38 to 0.97, which is the biggest change of the sensitivity for the Poisson component caused by merely changing the amount of the repeated measurements.



(a) True Coefficients: 0.1

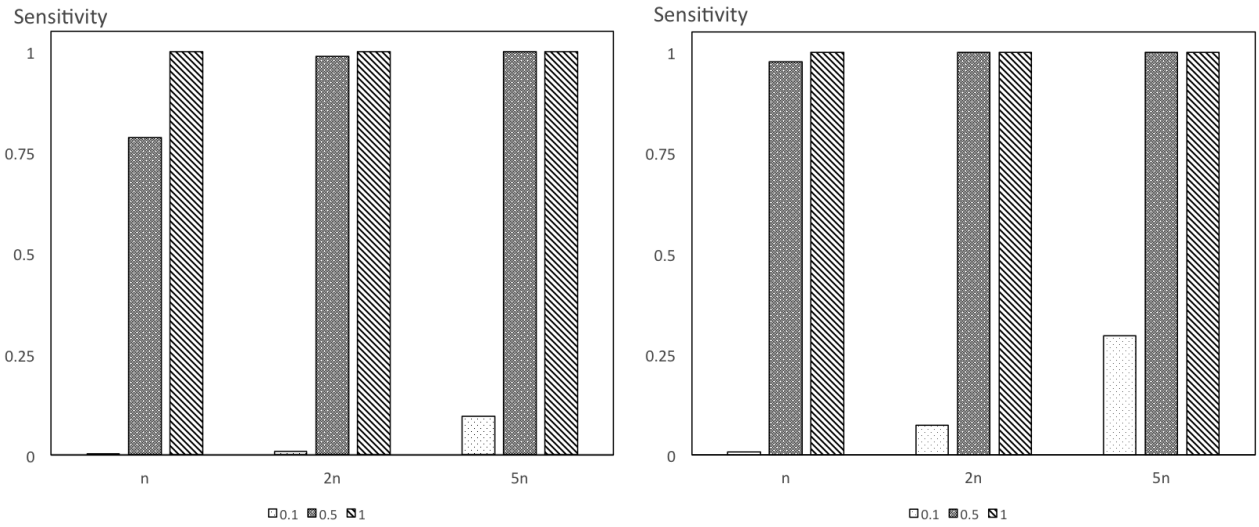


(b) True Coefficients: 0.5



(c) True Coefficients: 1

Figure 4.2: Sensitivity for the Poisson Component: Number of Subjects v.s. Number of Repeated Measurements v.s. True Coefficients



(a) Repeated Measurements: r **(b)** Repeated Measurements: $2r$

Figure 4.3: Sensitivity for the Poisson Component: Number of Subjects v.s. True Coefficients (Repeated Measurement being r and 2r)

Being similar to the number of repeated measurements, the number of subjects also has different effects under different magnitudes of the true coefficients. Figure 4.3.B shows how increasing the number of subjects can change the sensitivity under different sizes of variable effect, with the number of repeated measurements fixed to be $2r$. It can be seen that changing the number of subjects causes the sensitivity to increase when the true coefficients are 0.1. Compared with Figure 4.3.B, Figure 4.3.A fix the number of repeated measurements at a lower level r. In this case, increasing the number of subjects is more effective to increase the sensitivity when the true coefficients are 0.5 (the sensitivity's increase is from 0.79 to 1 as N goes from n to $5n$). From the above observations, we can see that changing the number of subjects is less effective under large magnitudes of coefficients (close to 1). Moreover, increasing the number of subjects cannot always improve the sensitivity, especially when the true size of variable effect is close to zero; with smaller regression coefficients, merely changing the number of subjects cannot result in higher sensitivity unless we also have higher levels of repeated measurements.

Sensitivity for the Zero-Inflation Component

The simulation factors' influence on the sensitivity for the zero-inflation component is similar to the previous discussion about the Poisson component. Firstly, it can be clearly indicated from Figure 4.5.C that if the true coefficients are 1, then having less amount of subjects or repeated measurements will still guarantee the sensitivity is above 0.55 (even though this is still far below 1 thus cannot be a satisfactory result, it is still much higher than the sensitivity when we only have regression coefficients as small as 0.1 and increase both N and R to their highest levels $5n$ and $2r$). If the true coefficients are 0.5, we can see the sensitivity will change drastically in response to the changes in the other two simulation factors. However, when the magnitudes of the true coefficients are as small as 0.1, increasing the number of subjects and the repeated measurements at the same time will increase the sensitivity but only in a fairly small range.

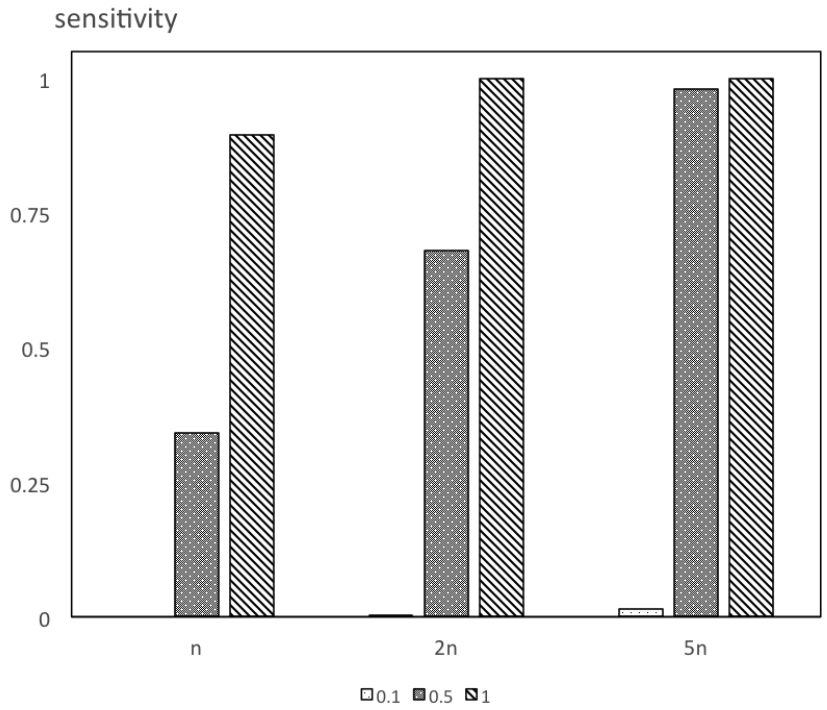
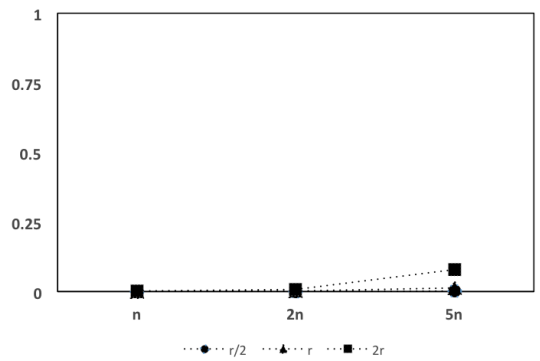
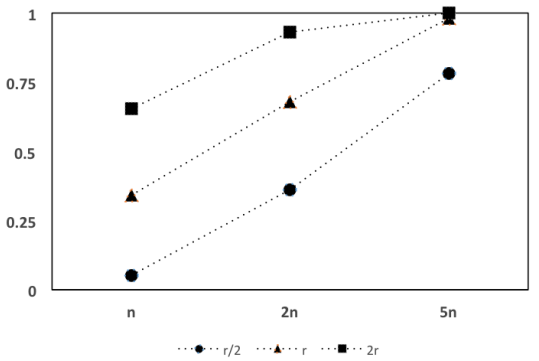


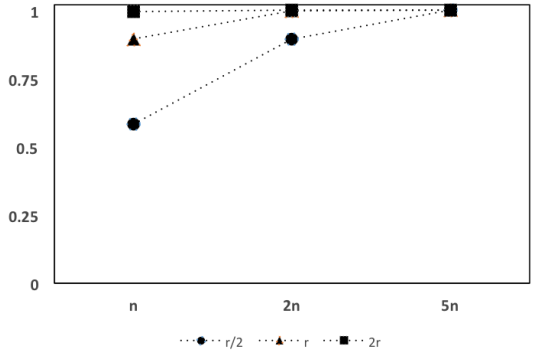
Figure 4.4: Sensitivity for the Zero-Inflation Component: Number of Subjects v.s. True Coefficients (Repeated Measurement being r)



(a) True Coefficients: 0.1



(b) True Coefficients: 0.5



(c) True Coefficients: 1

Figure 4.5: Sensitivity for the Zero-Inflation Component: Number of Subjects v.s. Number of Repeated Measurements v.s. True Coefficients

Being similar to the discussion for the Poisson component, increasing the number of subjects can increase sensitivity for the zero-inflation component, or at least make the sensitivity unchanged (setting the other two simulation factors at any levels). Again, under different magnitudes of true coefficients, increasing the number of subjects will make the sensitivity's increase ranges to be different. Setting the number of repeated measurements to be r (if setting the repeated measurements to be at different levels we can observe a similar tendency as well), increasing N will lead to the largest increase in the sensitivity when the true coefficients are 0.5 as it is indicated by Figure 4.4. This observation is similar to what has been found for the Poisson component–changing the simulation factors would not be necessary if we have the knowledge that the variables indeed have large effects in the model; on the contrary, when the true coefficients of the variables are not quite large, increasing the number of subjects can help to increase the sensitivity in the variable selection process; moreover, if we know the variables' true effects are very close to zero, larger number of subjects with more repeated measurement would be necessary to guarantee the sensitivity rate is not close to zero.

4.2.2 Specificity

Specificity for the Poisson Component

As it has been mentioned in Subsection 4.1.3, specificity is defined as the proportion of zero coefficients have credible intervals covering zero. It is worth mentioning that the specificity for the Poisson component remains almost unchanged no matter how we change the three simulation factors. Even though some small changes can be observed, the extent of the change in the specificity is quite small (smaller than 0.04), and the specificity is always larger than 0.95 (as it can be seen in Table 4.2). Therefore, most of the zero coefficients can have credible intervals covering zero and changing the simulation factors does not seem to affect the specificity for the Poisson component.

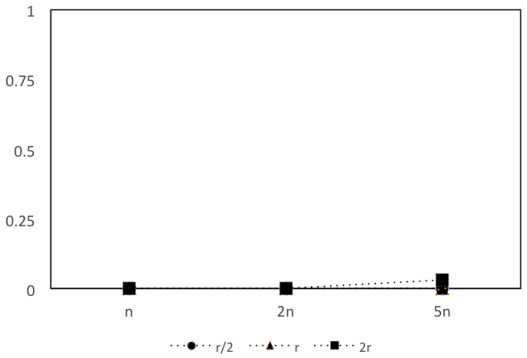
Specificity for the Zero-Inflation Component

Actually what we can find in the simulation results as for the specificity for the zero-inflation component is very similar to the previous discussion for the Poisson component. To be more specific, increasing the three parameters will result in very slight descent (with biggest decreased amount around 0.04) of the specificity, but the lowest specificity value is still larger than 0.95, which means that using the Bayesian LASSO method for variable selection, most of the zero coefficients could be successfully detected and thus the corresponding variables could be excluded from the model regardless how the simulation factors are changed. This phenomenon can be seen from Table 4.3.

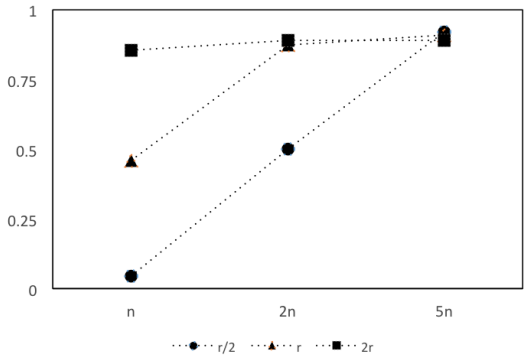
4.2.3 Exact Fit Rate

Exact Fit Rate for the Poisson Component

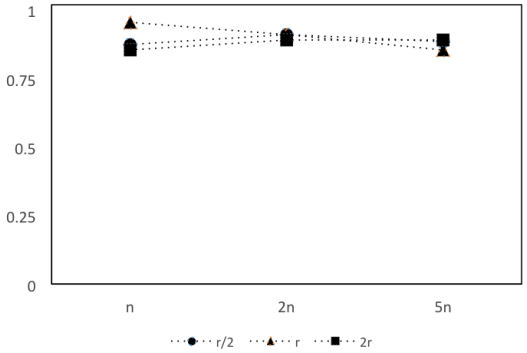
In my thesis, exact fit rate is defined as the probability of selecting the correct model (i.e. all the zero coefficients have credible intervals covering zero while all the non-zero coefficients have credible intervals not covering zero). We can investigate the influence of each simulation factor by constructing similar figures as in the discussions before. The pattern we can find in Figure 4.6 is very similar to our discussion for the sensitivity for the Poisson component. However, it does not mean that the exact fit rate has exactly the same changing pattern as the sensitivity when we choose different levels of the three simulation factors. The differences in the influence from changing the simulation factors, though quite small, would still be noticeable if we compare Figure 4.2 and Figure 4.6.



(a) True Coefficients: 0.1



(b) True Coefficients: 0.5



(c) True Coefficients: 1

Figure 4.6: Exact fit for the Poisson Component: Number of Subjects v.s. Number of Repeated Measurements v.s. True Coefficients

There are similarities between Figure 4.2 and Figure 4.6. Firstly, the exact fit rate is always close to 1 when the true coefficients are set to be 1; while the exact fit rate is close to 0 when the true coefficients are 0.1. Increasing only one of the other two simulation factors when the true coefficients are 0.1 only increase the exact fit rate up to 0.03. Therefore, only fixing the true coefficients to be at its medium levels (i.e., 0.5) will allow us to observe substantial changes in the exact fit rate. Secondly, Figure 4.6 also shows the different effects from changing the number of repeated measurements under each magnitude of regression coefficients: when the true coefficients are 0.5 and the number of subject is at its smallest value n, increasing the number of repeated measurements from $r/2$ to $2r$ can cause the exact fit rate to increase most drastically from 0.045 to 0.855. Thirdly, changing the number of subjects also has different effects with different magnitudes of the true coefficients, which can be shown by Figure 4.7: fixing the number of repeated measurements at r (the other two levels would lead to the similar pattern as well), increasing the number of subjects can lead to larger increase in the exact fit rate when the true coefficients are 0.5. For smaller coefficients, only changing the number of subjects will not be able to yield substantial increase for the exact fit rate, no matter which level of repeated measurements we choose.

However, there are some differences between the changing patterns of the sensitivity and the exact fit rate with respect to the three simulation factors. One obvious difference is about the effects of changing the number of subjects. Previously, when I discussed the findings of the sensitivity for the Poisson component, it is easy to find that increasing the number of subjects will lead the sensitivity to increase, or at least stay unchanged. However, when it comes to the exact fit rate, we can observe from Figure 4.6.C (with the true regression coefficients to be 1) that increasing the number of subjects actually result in a slight decrease in the exact fit rate when we set the number of repeated measurements to be $r/2$ or r. Moreover, increasing the repeated measurement cannot guarantee the exact fit rate to be higher either: seeing From Figure 4.6.C, fixing the number of subjects to be n, the highest exact fit occurs when the level of repeated measurements is r, instead of its highest level $2r$; similarly, fixing the number of subjects to be $2n$, the highest level of repeated measurements leads the exact fit rate to be 0.89, which is lower than the exact fit rate (0.91) when we set the number of repeated measurements to be $r/2$ or r. A similar pattern for the effects of changing the

number of repeated measurements can be found in Figure 4.6.B as well. If we compare Figure 4.2 and Figure 4.6, which are respectively for sensitivity and exact fit rate for the Poisson component, increasing the number of subjects or the number of repeated measurements may lead the exact fit rate to decrease slightly when the corresponding sensitivity has already reached 1.

Exact Fit Rate for the Zero-Inflation Component

To see how each simulation factor can affect the performance of the Bayesian LASSO, Figure 4.8 is constructed.

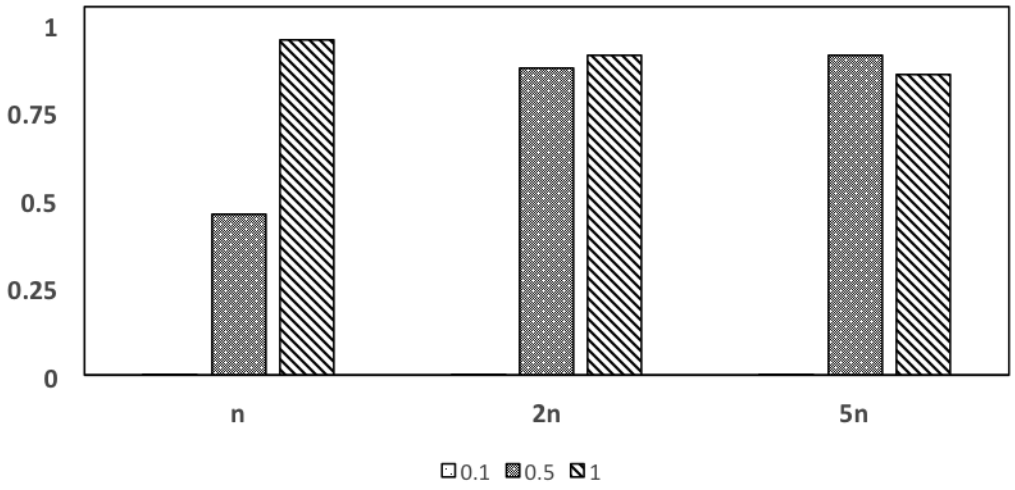
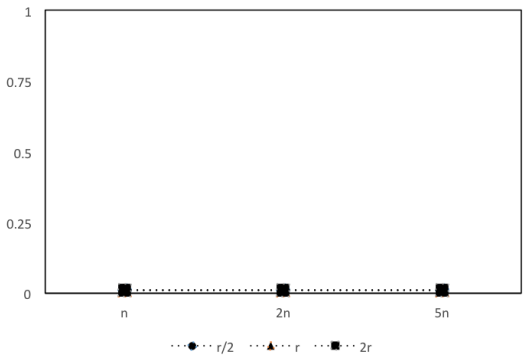
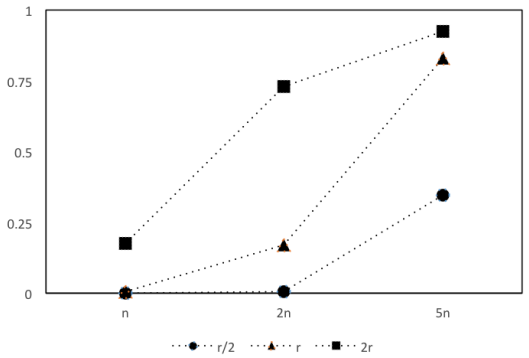


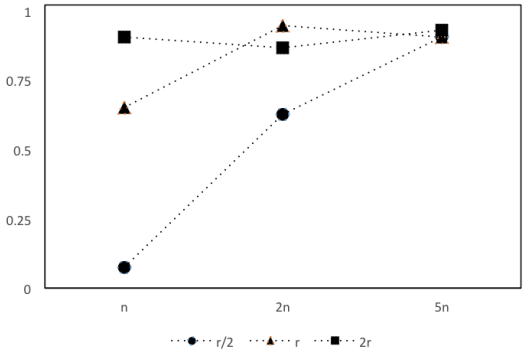
Figure 4.7: Exact fit for the Poisson Component: Number of Subjects v.s. True Coefficients (Repeated Measurement being r)



(a) True Coefficients: 0.1



(b) True Coefficients: 0.5



(c) True Coefficients: 1

Figure 4.8: Exact Fit Rate for the Zero-Inflation Component: Number of Subjects v.s. Number of Repeated Measurements v.s. True Coefficients

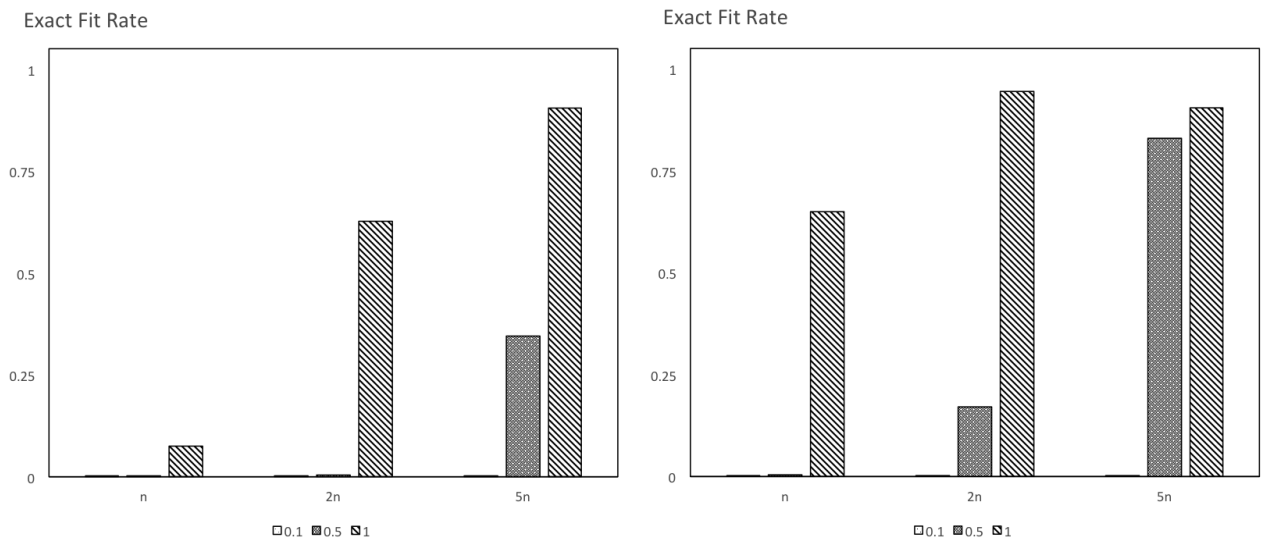
In Figure 4.8, we can observe some similar findings as the previous discussion for the sensitivity and the exact fit rate for the Poisson component. As I have mentioned in the general discussion before, the exact fit of the zero-inflation component remains to be zero without any variation when the true coefficients are 0.1. Also, the exact fit when the true coefficients are 1 is higher than the exact fit rate when the true coefficients are set at 0.5, fixing the other two simulation factors at any of their levels. Moreover, when the true coefficients are set to be 1 and the number of repeated measurements is set to be at its medium and highest level, we can observe smaller changes in the exact fit rate compared with the cases when the true coefficients are set to be 0.5. Therefore, when the true coefficients are 0.5, which is its medium level, changing the other two simulation factors would be mostly effective to change the exact fit rate.

I then set the number of repeated measurements to be r and $r/2$ (a similar pattern can be found when change r to $2r$), and investigate how changing the number of subjects will influence the exact fit rate given different true coefficients. The results can be shown in Figure 4.9. When the number of repeated measurements is $r/2$, increasing the number of subjects will lead to the biggest increase of the exact fit when the true coefficients are 1; while setting the number of repeated measurements to be r, increasing the number of subjects is mostly effective to increase the exact fit when the true coefficients are 0.5. The above patterns can be seen in Figure 4.9.

Similar to the discussion for the Poisson component's exact fit rate, increasing the number of subjects or the number of repeated measurements cannot always make the exact fit rate to increase. For example, in Figure 4.8.C, setting the level of repeated measurements to be $2r$, the exact fit rate is higher when the number of subjects is n compared with $2n$; while fixing the number of subjects to be $2n$, the zero-inflation component's exact fit rate is highest when we choose the medium level of repeated measurements.

4.3 Summary

From Subsection 4.1 to Subsection 4.2, I described the simulation results in terms of the sensitivity, the specificity and the exact fit rate respectively for the Poisson component and



(a) Repeated Measurements: $r/2$ **(b)** Repeated Measurements: r

Figure 4.9: Exact Fit Rate for the Zero-Inflation Component: Number of Subjects v.s. True Coefficients (Repeated Measurement being $r/2$ and r)

the zero-inflation component. This subsection of Chapter 4 is a brief summary of my findings and my explanation for some of the interesting findings.

Firstly, one main finding is the Bayesian LASSO has a quite good performance when the true coefficients are large, indicated by the sensitivity and the exact fit rate. Therefore, the data collectors can save their time or money in observing fewer subjects with fewer repeated measurements. On the contrary, when the true coefficients are very close to zero, the number of subjects and the number of repeated measurements have to increase at the same time, otherwise the Bayesian LASSO will have poor performance with low sensitivity and exact fit rate.

Secondly, we can see that increasing both the number of subjects and the number of repeated measurements can improve the Bayesian LASSO's performance in terms of the sensitivity, or at least make the sensitivity stay unchanged. Moreover, both the Poisson component and zero-inflation's specificity remains above 0.95 and we do not observe a clear pattern of the simulation factors' effects on the specificity.

Thirdly, increasing the number of subjects or the number of repeated measurements cannot always lead to higher exact fit rate, even though the sensitivity always response

positively to the increase of these two simulation factors. The slight decrease in the exact fit rate may result from the decrease of specificity. To be more specific, we can compare the Case (1, 2r, n) and Case (1, 2r, 2n). As it has been mentioned in the simulation results, increasing the number of subjects from n to 2n lead the exact fit rate for the zero inflation component to decrease from 0.905 to 0.865 (examining the Poisson component will allow us get the same conclusion). When examining Case (1, 2r, n), the specificity rate is pretty close to 1 (being 0.98). The specificity is not exactly one since some zero coefficients have credible intervals not covering zero. In other words, the Bayesian LASSO has picked up some variables that should not be included in the model. After examining the simulation results, we can see that among the 200 runs, there are 14 runs where the Bayesian LASSO picked up one variable that should not be included in the model; there is one run where the Bayesian LASSO picked up two wrong variables. When examining the Case (1, 2r, 2n), where the only difference in the simulation setting is the increase of the number of subjects (from n to 2n), we can see in 21 runs the Bayesian LASSO picked up one variable that should not be included in the model; meanwhile in 5 runs the Bayesian LASSO picked up 2 variables that were not used to generate the response variable. It is worth mentioning that the sensitivity, which represent Bayesian LASSO's ability to pick up the variables that should be included in the model, goes up from 0.995 to 1. However the slight increase in the sensitivity cannot lead to an increase in the exact fit rate since the decrease in the specificity is larger. We can actually observe similar patterns when we consider increasing the number of repeated measurements: the exact fit rate still goes down since the increase in the sensitivity cannot compensate the negative effect of picking up the variables that should not be in the model.

For the causes of this phenomenon (further increasing the number of subjects or repeated measurements may lead to the risk of lowering the exact fit rate when the corresponding sensitivity already reaches 1), I currently do not have a justifiable answer. My current conjecture is that with the high level of zero inflation in my generated response variable (near 80% of total zero proportion), my changing ranges for N and R are still not large enough to allow us to observe the stable patterns of how the proposed Bayesian LASSO's performance can be affected. I have this conjecture since when we have N to be n and R to be r, and if we perform the model fitting, the point estimates for the model coefficients are

not close to their true values. With more time to undertake more comprehensive simulations, it is possible to testify my conjecture and investigate if further increase N or R could enable us to find more stable patterns of the proposed Bayesian LASSO's performance.

Fourthly, the three simulation factors we consider actually interactively affect the Bayesian LASSO's performance, thus it is impossible to conclude the effects of one of them without considering the other two. For example, when applying the Bayesian LASSO into the real world problems for variable selection, increasing the number of subjects may not be equally effective under different magnitudes of true coefficients: when the number of repeated measurements is relatively small (like the cases in Figure 4.3.A and Figure 4.9.A), increasing the number of subjects will increase the sensitivity/exact fit rate for the cases with larger true coefficients first if the sensitivity/exact fit rate is not already 1; while if the repeated measurements are larger (Figure 4.3.B and Figure 4.9.B), cases with large true coefficients have already reached high levels of sensitivity/exact fit rate with lower levels of N, thus further increasing N will be more useful to increase the sensitivity/exact fit rate for the cases with smaller magnitudes of coefficients. In shorts, the three simulation factors jointly affect the simulation results, and increasing N or R will be effective to select the variables with larger true effects first.

Lastly, we can actually see from the above figures and tables that the Bayesian LASSO has a better performance (higher sensitivity) for the Poisson component compared with the zero-inflation component. The reason for this phenomenon is probably that our likelihood function (given by Equation 2.1 to Equation 2.5) is more informative about the parameters in the Poisson component. Given that we cannot separate the Poisson zeros and the zero inflation, the Bayesian LASSO's likelihood function is less informative about the parameters in zero-inflation component. That is the possible reason that the Bayesian LASSO's performance for $\boldsymbol{\alpha}$ parameters is not as good as its performance for $\boldsymbol{\beta}$ parameters in terms of the sensitivity and the exact fit rate.

Chapter 5

Conclusions and Future Work

In this chapter, I will summarize my research objective and the main findings of my thesis. The object of my thesis is to evaluate how the proposed Bayesian LASSO variable selection method (Chapter 3) performs when applying to ZIP models. I conducted a series of simulation studies to investigate and summarize my findings from the simulation results.

My thesis introduces the Bayesian LASSO method proposed by Brown et al. (2015) for variable selection of ZIP models, gives the description of my simulation design and summarizes the results and presents some main findings. To summarize the simulation results, the first obvious conclusion is that the Bayesian LASSO has a quite good performance when the true coefficients are of larger magnitudes (close to 1). Secondly, increasing the number of subjects cannot always improve the variable selection performance: in some of the cases, increasing N might lead to slight decrease in the exact fit rate and the specificity, while the sensitivity will increase or at least stay unchanged. The slight decrease in the exact fit rate and the specificity is actually beyond my expectation. The third finding is that when the true effects of the variables are very close to zero, it would help to promote the Bayesian LASSO's performance if both the number of subjects and the number of repeated measurements are increased; Otherwise the sensitivity will be very close to zero. Fourthly, we can actually see that the Bayesian LASSO has a better performance (higher sensitivity and higher exact fit rate) for the Poisson component compared with the zero inflation component.

The results of my simulation study are also of practical meanings. For researchers in Ecology, Plant Science and some other related fields, my simulation results may imply some guidance for data collection. Firstly, in the longitudinal dataset, more subjects with larger numbers of repeated measurements would be necessary since the proposed Bayesian LASSO can have better performance with larger N and R, especially when the true effects of the

variables are not very large. Secondly, with the constraint of time and the cost of data collection, the researchers may want to collect data on more subjects, instead of more repeated measurements within a single site, especially the true effects of the variables are small or moderate. To be more specific, compared with increasing the number of repeated measurements, increasing the number of subjects can lead to larger increase in the sensitivity and the exact fit rate when the true coefficients are 0.1 and 0.5 for both the Poisson component and the zero inflation component. When the true coefficients are 1, then I can observe slight decrease in both the sensitivity and the exact fit rate when increasing the number of subjects or the number of repeated measurements, thus I cannot obtain a unified conclusion on whether larger N or R could be more influential to affect Bayesian LASSO's performance. It is possible that my current settings of N are not big enough to capture the patterns for different R when the true coefficients are 1. In this sense, more simulations with more levels of N and R may be necessary to answer the question that whether the data collectors should obtain data for more subjects or more repeated measurements within a single subject when the true effects of the coefficients are very large.

My thesis has several limitations which can be the directions for future research work. Firstly, as mentioned above, when using the exact fit rate and the specificity, there are some results that are beyond my expectation. Future researchers can develop other criteria to better evaluate variable selection methods' performance in the ZIP models. Secondly, this thesis only considers L_1 penalty, where the regression parameters are set to follow double exponential priors. However, it is possible that other functions of penalty can be employed as well. Therefore, the comparison between different penalty forms for the Bayesian LASSO applying in the ZIP model is one possible focus for future works. Thirdly, the current method is only for fixed effects selection. However, the mixed effect selection in ZIP models, which has not been covered by my thesis, is also missing from the existing literature. Lastly, my thesis only considers independent and identically distributed (iid) random effects, while future works can extend the iid random effect to dependent random effects. For example, the random effect can have spatial dependence, thus how to incorporate this dependence in the variable selection in ZIP models would be a possible direction of future work.

References

Algamal, Z. Y. and Lee, M. H. (2015), "Adjusted adaptive lasso in high-dimensional poisson regression model," *Modern Applied Science*, 9, 170.

Bae, K. and Mallick, B. K. (2004), "Gene selection using a two-level hierarchical Bayesian model," *Bioinformatics*, 20, 3423–3430.

Bohara, A. K. and Krieg, R. G. (1996), "A zero-inflated Poisson model of migration frequency," *International Regional Science Review*, 19, 211–222.

Breiman, L. (1995), "Better subset regression using the nonnegative garrote," *Technometrics*, 37, 373–384.

Brown, C. and Johnstone, J. (2011), "How does increased fire frequency affect carbon loss from fire? A case study in the northern boreal forest," *International Journal of Wildland Fire*, 20, 829–837.

Brown, C. D. and Johnstone, J. F. (2012), "Once burned, twice shy: Repeat fires reduce seed availability and alter substrate constraints on Picea mariana regeneration," *Forest Ecology and Management*, 266, 34–41.

Brown, C. D., Liu, J., Yan, G., and Johnstone, J. F. (2015), "Disentangling legacy effects from environmental filters of postfire assembly of boreal tree assemblages," *Ecology*, 96, 3023–3032.

Buu, A., Johnson, N. J., Li, R., and Tan, X. (2011), "New variable selection methods for zero-inflated count data with applications to the substance abuse field," *Statistics in medicine*, 30, 2326–2340.

Carrico, C., Gennings, C., Wheeler, D. C., and Factor-Litvak, P. (2015), "Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting," *Journal of Agricultural, Biological, and Environmental Statistics*, 20, 100–120.

Costa, M. A., de Souza Rodrigues, T., da Costa, A. G. F., Natowicz, R., and Braga, A. P. (2015), "Sequential selection of variables using short permutation procedures and multiple adjustments: An application to genomic data," *Statistical methods in medical research*, 10.1177/0962280214566262.

Czarnota, J., Wheeler, D. C., and Gennings, C. (2015), "Evaluating Geographically Weighted Regression Models for Environmental Chemical Risk Analysis," *Cancer informatics*, 14, 117.

Derksen, S. and Keselman, H. (1992), "Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables," *British Journal of Mathematical and Statistical Psychology*, 45, 265–282.

Desouhant, E., Debouzie, D., and Menu, F. (1998), "Oviposition pattern of phytophagous insects: on the importance of host population heterogeneity," *Oecologia*, 114, 382–388.

Fan, J. and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American statistical Association*, 96, 1348–1360.

Figueiredo, M. A. (2003), "Adaptive sparseness for supervised learning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25, 1150–1159.

Frank, L. E. and Friedman, J. H. (1993), "A statistical view of some chemometrics regression tools," *Technometrics*, 35, 109–135.

Greene, W. H. (1994), "Accounting for excess zeros and sample selection in Poisson and negative binomial regression models," .

Hall, D. B. (2000), "Zero-inflated Poisson and binomial regression with random effects: a case study," *Biometrics*, 56, 1030–1039.

Heilbron, D. (1989), "Generalized linear models for altered zero probabilities and overdispersion in count data," *Unpublished Technical report, University of California, San Francisco, Department of Epidemiology and Biostatistics*.

Hocking, R. R. (1976), "A Biometrics invited paper. The analysis and selection of variables in linear regression," *Biometrics*, 32, 1–49.

Hoerl, A. E. and Kennard, R. W. (1970), "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, 12, 55–67.

Hsu, D. (2015), "Identifying key variables and interactions in statistical models of building energy consumption using regularization," *Energy*, 83, 144–155.

Johnstone, J., Boby, L., Tissier, E., Mack, M., Verbyla, D., and Walker, X. (2009), "Postfire seed rain of black spruce, a semiserotinous conifer, in forests of interior Alaska," *Canadian Journal of Forest Research*, 39, 1575–1588.

Lambert, D. (1992), "Zero-inflated Poisson regression, with an application to defects in manufacturing," *Technometrics*, 34, 1–14.

Lian, H. (2012), "Variable selection in high-dimensional partly linear additive models," *Journal of Nonparametric Statistics*, 24, 825–839.

Lim, M. and Hastie, T. (2013), "Learning interactions through hierarchical group-lasso regularization," *arXiv preprint arXiv:1308.2719*.

Lord, D., Washington, S. P., and Ivan, J. N. (2005), "Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory," *Accident Analysis & Prevention*, 37, 35–46.

Miaou, S.-P. (1994), "The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions," *Accident Analysis & Prevention*, 26, 471–482.

Mitchell, T. J. and Beauchamp, J. J. (1988), "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, 83, 1023–1032.

Mortier, F., Ouédraogo, D.-Y., Claeys, F., Tadesse, M. G., Cornu, G., Baya, F., Benedet, F., Freycon, V., Gourlet-Fleury, S., and Picard, N. (2015), "Mixture of inhomogeneous matrix models for species-rich ecosystems," *Environmetrics*, 26, 39–51.

Mullahy, J. (1986), "Specification and testing of some modified count data models," *Journal of econometrics*, 33, 341–365.

Narendra, P. M. and Fukunaga, K. (1977), "A branch and bound algorithm for feature subset selection," *Computers, IEEE Transactions on*, 100, 917–922.

Neelon, B. H., O?Malley, A. J., and Normand, S.-L. T. (2010), "A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use," *Statistical Modelling*, 10, 421–439.

Park, T. and Casella, G. (2008), "The bayesian lasso," *Journal of the American Statistical Association*, 103, 681–686.

Plummer, M. (2013), "rjags: Bayesian graphical models using MCMC," *R package version*, 3.

Qian, W. and Yang, Y. (2013), "Model selection via standard error adjusted adaptive lasso," *Annals of the Institute of Statistical Mathematics*, 65, 295–318.

Ridout, M., Demétrio, C. G., and Hinde, J. (1998), "Models for count data with many zeros," in *Proceedings of the XIXth international biometric conference*, vol. 19, pp. 179–192.

Rose, C. E., Martin, S. W., Wannemuehler, K. A., and Plikaytis, B. D. (2006), "On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data," *Journal of biopharmaceutical statistics*, 16, 463–481.

Tang, Y., Xiang, L., and Zhu, Z. (2014), "Risk Factor Selection in Rate Making: EM Adaptive LASSO for Zero-Inflated Poisson Regression Models," *Risk Analysis*, 34, 1112–1127.

Team, R. C. (2014), "R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013," .

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

Van den Broek, J. (1995), "A score test for zero inflation in a Poisson distribution," *Biometrics*, 738–743.

Wang, Z., Ma, S., and Wang, C.-Y. (2015), "Variable selection for zero-inflated and overdispersed data with application to health care demand in Germany," *Biometrical Journal*, 57, 867–884.

Wheeler, D. C. (2009), "Simultaneous coefficient penalization and model selection in geographically weighted regression: the geographically weighted lasso," *Environment and Planning A*, 41, 722–742.

Xu, X., Ghosh, M., et al. (2015), "Bayesian variable selection and estimation for group lasso," *Bayesian Analysis*, 10, 909–936.

Yau, K. K. and Lee, A. H. (2001), "Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme," *Statistics in medicine*, 20, 2907–2920.

Yuan, M. and Lin, Y. (2006), "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 49–67.

Zeng, P., Wei, Y., Zhao, Y., Liu, J., Liu, L., Zhang, R., Gou, J., Huang, S., and Chen, F. (2014), "Variable selection approach for zero-inflated count data via adaptive lasso," *Journal of Applied Statistics*, 41, 879–894.

Zhao, P. and Yu, B. (2006), "On model selection consistency of Lasso," *The Journal of Machine Learning Research*, 7, 2541–2563.

Zou, H. (2006), "The adaptive lasso and its oracle properties," *Journal of the American statistical association*, 101, 1418–1429.

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