Zero-inflated Models for Semi-continuous Transportation Data

Tianshu Feng

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Linda Ng Boyle, Chair

Archis Ghate

Youngjun Choe

Program Authorized to Offer Degree:
Industrial & Systems Engineering

University of Washington

Abstract

Zero-inflated Models for Semi-continuous Transportation Data

Tianshu Feng

Chair of the Supervisory Committee:
Professor and Chair Linda Ng Boyle
Industrial & Systems Engineering

Zero-inflated models have been widely studied and are commonly used in the transportation safety area. Despite the success of zero-inflated models to analyze static data with counting outcomes, challenges remain in the examination of zero-inflated data with semi-continuous and auto-correlated longitudinal outcomes. This dissertation aims to explore different approaches to tackle the existing challenges in semi-continuous zero-inflated data analysis. The dissertation begins with a discussion of challenges with existing zero-inflated models in modelling semi-continuous data and time-series data. Then, our recent works on a variable selection method for semi-continuous zero-inflated models and a dynamic model for semi-continuous zero-inflated time-series data are presented. Simulated data was used to validate the proposed models. And finally, we demonstrate the proposed models using two different transportation datasets. These datasets are from a driving simulator study and a field operational test. The results suggest that the proposed models can capture the differences in driving behavior between individuals and between different driving situations, which have implications for the design of in-vehicle assistance systems.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my deepest thanks to my advisor, Dr. Linda Boyle, for being an incredible advisor. I am thankful for her generous mentoring and advice during my Ph.D. studies at the University of Washington. Our regular meetings and ad-hoc discussions have been invalueable for me to accumulate experiences in research. Dr. Boyle's enthusiasm, intelligence, leadership, and rich and extensive experience in research greatly inspired me and helped me to become a better researcher and collaborator in the future.

I would also like to acknowledge other members of my dissertation committee. I enjoyed the optimization and stochastic courses from Dr. Archis Ghate, and would like to thank him for his suggestions on the optimization methods used in the thesis. The quality control class from Dr. Youngjun Choe greatly inspired my early research works, and his detailed and insightful comments and suggestions helped me to recognize areas where greater clarity were needed. Information theory from Dr. Sreeram Kannan changed my way of thinking, and his suggestions were inspiring and helpful for me to improve the content of the thesis.

I am thankful for the wonderful faculty members, staffs, and students in the Department of Industrial & Systems Engineering at the University of Washington for creating such a lovely department. I am particularly grateful for Ms. Jennifer Tsai and Ms. Sheila Prusa, who are always trustworthy and reliable when I need help.

I would like to thank Dr. Chen Wang for his kind help and the inspiring discussions since my internship at Mayo Clinic. I greatly appreciate my friends met during the interns for the fruitful interdisciplinary discussions and chats.

I would like to acknowledge my family and friends for their encouragement and confidence. Their support kept me going regardless of the challenges I faced.

Chapter 1

INTRODUCTION

1.1 Motivation

Zero-inflated counting data is ubiquitous in the transportation domain to examine crash frequency at intersections (Dong et al., 2014), on roadway segments (Liu et al., 2018), freeways (Lord et al., 2005), and for weather-related conditions (Carson and Mannering, 2001). These datasets usually include excessive zeros that partially follows a Poisson or negative binomial distribution (Lambert, 1992; Miaou, 1994; Mwalili et al., 2008) and can be analyzed with zero-inflated counting models, e.g., zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models. However, with more temporal and spatial data available (Stathopoulos and Karlaftis, 2001; Yu et al., 2019), it becomes necessary to consider the distribution of data as more than a static zero-inflated counting distribution. A semi-continuous zero-inflated framework can capture the continuous variations observed when one deviates within a nonzero state. For example, the action related to braking in the car impacts driving safety (McGehee et al., 2016). Sudden and hard presses of the brake pedal usually imply emergency events, while soft presses are often associated with normal driving if the driver is attending to the roadway. However, in certain driving situations, such as on the highway without traffic, the drivers would not be expected to depress the brake pedal frequently or at all, which leads to inflation of zeros. When they do press on the brake pedal, the actual trajectory of force applied to the pedal would be important to examine. This kind of data, with a combination of a positive continuous distribution and a point mass at zero, is called point-mass mixture data (Taylor and Pollard, 2009), or semi-continuous zero-inflated (SCZI) data.

Recent studies have demonstrated the application of statistical models on the semi-

continuous zero-inflated data in different areas (Liu et al., 2016). Mills (2013) performed statistical tests for zero-inflated Gamma (ZIG) and zero-inflated log-normal (ZING) models and applied the two models to lane departure datasets. Many zero-inflated time-series count models have also been developed to analyze accident frequencies and crashes (Malyshkina et al., 2009; Malyshkina and Mannering, 2010; Xiong et al., 2014). For example, Malyshkina et al. (2009) proposed a zero-state Markov switching count-data model to analyze the influence of weather-condition variables on accident frequencies based on observations across three consecutive years.

Despite the success of zero-inflated models in analyzing SCZI data, many questions remain unsolved. This dissertation focuses on the following questions:

Question 1. When many covariates are available in a semi-continuous zero-inflated dataset, how can we identify these covariates that are related to the responses and can improve model fitting? Previous works tried to answer this question with hypothesis testings (Mills, 2013). This dissertation introduces a new variable selection method for SCZI models that is more efficient in selecting a subset of covariates and considers the relationship between different coefficients.

Question 2. How to incorporate the auto-correlation between semi-continuous zero-inflated responses to improve the model fitting of SCZI models? For example, data used in Mills (2013) is collected at consecutive time points, and it can be inappropriate to use static models ignoring the potential auto-correlation between responses. To answer this question, we propose a time-series model for semi-continuous zero-inflated data. A numerical study is conducted to assess the developed model.

Question 3. How to apply the SCZI models to analyze data from transportation studies improve driving assistance systems? In this dissertation, we use data from a driving simulator study (Chang, 2016) and a field operation test (Ervin et al., 2005) to demonstrate how the SCZI models can capture the differences in driving behavior between individuals and between different driving situations.

1.2 Zero-inflated data examples

Zero-inflated data exists in many areas. In finance, the number of claims and claimed loss of an insurance policy are zero-inflated (Duan et al., 1983; Chowdhury et al., 2018). For example, Chowdhury et al. (2018) applied zero-inflated models to study the claimed loss of auto insurance policies where 60.7% of the policies have no claims.

In medical care, examples include the number of physician visits (Pizer and Prentice, 2011; Chatterjee et al., 2018), the number of prescriptions (Street et al., 1999), and daily cigarette use (Pittman et al., 2018). For example, Neelon et al. (2015) studied the relationship between emergency department expenditures and patient demographics.

In transportation studies, crash frequency at intersections (Dong et al., 2014), on roadway segments (Liu et al., 2018), freeways (Lord et al., 2005), and for weather-related conditions (Carson and Mannering, 2001) is often zero-inflated. One example that motivates this dissertation comes from the study of lane departures. Studies in lane departure often consider the binary outcome of departing a lane or not (Albousefi et al., 2014), but it is also important to understand how far off they have departed (Mills, 2013). The intuition is that a driver can drift far from the center of the driving lane for an extended time. However, in most cases, the drivers remain in the lane; this leads to a non-negative outcome that is primarily continuous but has a point mass at zero. This kind of data with a combination of a positive continuous distribution and a point mass at zero is semi-continuous zero-inflated. Another example showed the impact of the brake pedal force on driving safety McGehee et al. (2016). Sudden and hard presses of the accelerator or brake pedal usually imply emergency events, while soft presses are considered common driving behaviors. However, in normal driving situations with light traffic, the drivers will not be touching the brake pedal frequently, which leads to inflation of zeros. When they do brake, the actual force applied to the pedal is important to be examined. This kind of data with a combination of a positive continuous distribution and a point mass at zero is often called mass mixture data, or semi-continuous zero-inflated data.

Chapter 2

REVIEW OF ZERO-INFLATED MODELS

Zero-inflated models are for data with excess zeros. They often assume that with probability p, the only observation is 0, and with probability $1-p$, a non-negative random variable is observed. This chapter reviews the established zero-inflated models.

2.1 Discrete zero-inflated models

Discrete zero-inflated models, e.g., zero-inflated Poisson and negative binomial models, assume that the discrete responses are from two zero generating parts (Lambert, 1992). The first part (zero part) is governed by a Bernoulli distribution that generates zeros. The second part (nonzero part) is governed by a Poisson or negative binomial distribution that generates zeros and positive counts. Therefore, although a positive count comes from the nonzero part, zero counts can come from either the zero or nonzero part.

Mathematically, suppose that we have data $(\boldsymbol{y}, \boldsymbol{X})$, where $\boldsymbol{y} = (y_1, \ldots, y_n)$ is a vector of n responses, y_i is the response of the ith subject, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ is a $n \times p$ matrix of $p-1$ variables where the first column of X represents the intercept, and $\boldsymbol{x}_i$ is the vector of the intercept and $p-1$ variables corresponding to the ith subject for $i = 1, \ldots, n$. For the zero-inflated Poisson (ZIP) model, we have:

$$\begin{aligned} P(Y_i = k) =& (1 - p_{\beta^{(0)},i}) I(k = 0) + p_{\beta^{(0)},i} g_{\beta^{(1)}}(k|\boldsymbol{x}_i, k > 0) \\ =& \begin{cases} (1 - p_{\beta^{(0)},i}) + p_{\beta^{(0)},i} e^{-\mu_i}, & \text{if } k = 0 \\ p_{\beta^{(0)},i} \frac{e^{-\mu_i} \mu_i^k}{k!}, & \text{if } k = 1, \ldots \end{cases} \end{aligned} \tag{2.1}$$

where $p_{\beta^{(0)},i} = P(y_i > 0|\boldsymbol{x}_i; \beta^{(0)})$ denotes the probability of $y_i > 0$, $I(\cdot)$ is the indicator function, i.e., $I(k = 0) = 1$ if $k = 0$ and $I(k = 0) = 0$ otherwise, $g(k|\boldsymbol{x}_i, k > 0; \beta^{(1)})$ represents the conditional distribution of the positive responses, and μ_i denotes the expectation of the

Poisson distribution. The subscript $\beta^{(0)} \in \mathbb{R}^p$ of $p_{\beta^{(0)}}$ and $\beta^{(1)} \in \mathbb{R}^p$ of $g(y|\boldsymbol{x}, y > 0; \beta^{(1)}, \theta)$ represent the regression coefficients, where

$$\begin{aligned} \text{logit}(p_{\beta^{(0)},i}) &= \log\left(\frac{p_{\beta^{(0)},i}}{1-p_{\beta^{(0)},i}}\right) = \boldsymbol{x}_i^T\beta^{(0)} \\ \log(\mu_i) &= \boldsymbol{x}_i^T\beta^{(1)}. \end{aligned} \tag{2.2}$$

For n independent observations, the negative log-likelihood function of ZIP is

$$\begin{aligned} l(\beta) =& \sum_{i=1}^{n} \log\left(1 + e^{\boldsymbol{x}_i^T\beta^{(0)}}\right) + \sum_{y_i>0} \log(y_i!) - \sum_{y_i=0} \log\left(e^{\boldsymbol{x}_i^T\beta^{(0)}} + e^{-e^{\boldsymbol{x}_i\beta^{(1)}}}\right) \\ & - \sum_{y_i>0}\left(y_i\boldsymbol{x}_i\beta^{(1)} + e^{-\boldsymbol{x}_i\beta^{(1)}}\right) \end{aligned} \tag{2.3}$$

Similarly, we can define the zero-inflated negative binomial (ZINB) model as follows:

$$\begin{aligned} P(Y_i = k) =& (1 - p_{\beta^{(0)},i})I(k=0) + p_{\beta^{(0)},i}g_{\beta^{(1)}}(k|\boldsymbol{x}_i, k>0) \\ =& \begin{cases} (1-p_{\beta^{(0)},i}) + p_{\beta^{(0)},i}\left(\dfrac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\frac{1}{\alpha}}, & \text{if } k=0 \\ p_{\beta^{(0)},i}\dfrac{\Gamma(\alpha^{-1}+k)}{\Gamma(k+1)\Gamma(\alpha^{-1})}\left(\dfrac{\mu_i}{\alpha^{-1}+\mu_i}\right)^k\left(\dfrac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}, & \text{if } k=1,\ldots \end{cases} \end{aligned} \tag{2.4}$$

The negative log-likelihood function of ZINB is

$$\begin{aligned} l(\beta,\alpha) =& -\sum_{y_i=0}\log\left(p_i + (1-p_i)\left(\frac{\alpha^{-1}}{\mu_i+\aleph^{-1}}^{\frac{1}{\alpha}}\right)\right) \\ & - \sum_{y_i>0}\log\left((1-p_i)\frac{\Gamma(\alpha^{-1}+y_i)}{\Gamma(y_i+1)\Gamma(\alpha^{-1})}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\right). \end{aligned} \tag{2.5}$$

Originally, parameters of the ZIP model can be estimated with EM algorithm (Lambert, 1992). Later works showed that the maximum likelihood estimates of both ZIP and ZINB can be efficiently found with gradient approaches with line searches (Greene, 1994).

2.2 Semi-continuous zero-inflated models

The responses of semi-continuous data consist of a point-mass at zero and a positively skewed distribution for the remaining positive responses. The responses from zero-inflated data are

often difficult to examine using traditional positive continuous distributions partly because the underlying distribution of responses is not continuous.

Semi-continuous zero-inflated (ZI) models (Mills, 2013; Zhou and Tu, 2000) provide us with approaches to evaluate this type of data by incorporating the point-mass at zero and the positive continuous values with a two-part modeling framework. Specifically, suppose that we have data $(\boldsymbol{y}, \boldsymbol{X})$, where $\boldsymbol{y} = (y_1, \ldots, y_n)$ is a vector of n responses, y_i is the response of the ith subject, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ is a $n \times p$ matrix of $p-1$ variables where the first column of X represents the intercept, and $\boldsymbol{x}_i$ is the vector of the intercept and $p-1$ variables corresponding to the ith subject for $i = 1, \ldots, n$. Under the two-part model framework, the mixture probability distribution can be formulated as:

$$f(y) = (1 - p_{\beta^{(0)}})I(y = 0) + p_{\beta^{(0)}} g(y|\boldsymbol{x}, y > 0; \beta^{(1)}, \theta), \tag{2.6}$$

where $p_{\beta^{(0)}} = P(y > 0|\boldsymbol{x}; \beta^{(0)})$ denotes the probability of $y > 0$, $I(\cdot)$ is the indicator function, i.e., $I(y = 0) = 1$ if $y = 0$ and $I(y = 0) = 0$ otherwise, $g(y|\boldsymbol{x}, y > 0; \beta^{(1)}, \theta)$ represents the conditional distribution of the positive responses, and θ denotes the parameters of $g(y|\boldsymbol{x}, y > 0; \beta^{(1)}, \theta)$. The subscript $\beta^{(0)} \in \mathbb{R}^p$ of $p_{\beta^{(0)}}$ and $\beta^{(1)} \in \mathbb{R}^p$ of $g(y|\boldsymbol{x}, y > 0; \beta^{(1)}, \theta)$ represent the regression coefficients.

Given the probability distribution $f(y)$, we have the likelihood function:

$$L(\beta, \theta) = \prod_{y_i=0} (1 - p_{\beta^{(0)},i}) \prod_{y_i>0} p_{\beta^{(0)},i} g(y_i|\boldsymbol{x}_i, y_i > 0; \beta^{(1)}, \theta), \tag{2.7}$$

where $\beta = (\beta^{(0)}, \beta^{(1)}) \in \mathbb{R}^{2p}$, and $p_{\beta^{(0)},i}$ is the probability of $y_i > 0$ for subject i. The likelihood function can be effectively segmented into two parts, a zero part $L_0(\beta^{(0)})$ and a non-zero part $L_1(\beta^{(1)}, \theta)$, that are exclusively related to either $\beta^{(0)}$ or $\beta^{(1)}$. This segmentation is feasible assuming that the responses y_i, $i = 1, \ldots, n$, are i.i.d, and $\beta^{(0)}$ and $\beta^{(1)}$ are independent of each other given data $(\boldsymbol{y}, \boldsymbol{X})$ Mills (2013). Rewriting the likelihood function (2.7) with segmentation leads to

$$L(\beta, \theta) = L_0(\beta^{(0)}) L_1(\beta^{(1)}, \theta) = \prod_{i=1}^{n} \left[(1 - p_{\beta^{(0)},i})^{I(y_i=0)} p_{\beta^{(0)},i}^{I(y_i>0)} \right] \prod_{y_i>0} g(y_i|\boldsymbol{x}_i, y_i > 0; \beta^{(1)}, \theta). \tag{2.8}$$

The probability $p_{\beta^{(0)}}$ and conditional distribution $g(y|\boldsymbol{x}, y > 0; \beta^{(1)}, \theta)$ in Eq.(2.6) and Eq.(2.8) can be parameterized with different link functions. For example, Duan et al. (1983) used probit regression for $p_{\beta^{(0)}}$ and log-normal distribution for $g(y|\boldsymbol{x}, y > 0; \beta^{(1)}, \theta)$. In (Zhou and Tu, 2000; Nobre et al., 2017), $p_{\beta^{(0)}}$ was modeled with logistic regression, and the conditional distributions are log-normal and gamma distributions, respectively.

In this thesis, logistic regression is used for the zero part and gamma regression is used for the nonzero part. For a zero-inflated gamma (ZIG) model of the form:

$$\text{logit}(p_{\beta^{(0)},i}) = \log\left(\frac{p_{\beta^{(0)},i}}{1 - p_{\beta^{(0)},i}}\right) = \boldsymbol{x}_i^T \beta^{(0)} \tag{2.9a}$$

$$g(y_i|\boldsymbol{x}_i, y_i > 0; \beta^{(1)}, \nu^{-1}) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i}\right)^{\nu} y_i^{\nu-1} \exp\left(-\frac{\nu y_i}{\mu_i}\right). \tag{2.9b}$$

where $\log(\mu_i) = \boldsymbol{x}_i^T \beta^{(1)}$. Note that it can be adapted to a zero-inflated log-normal model if we assume the nonzero responses are from a log-normal distribution:

$$\text{logit}(p_{\beta^{(0)},i}) = \log\left(\frac{p_{\beta^{(0)},i}}{1 - p_{\beta^{(0)},i}}\right) = \boldsymbol{x}_i^T \beta^{(0)} \tag{2.10a}$$

$$g(y_i|\boldsymbol{x}_i, y_i > 0; \beta^{(1)}, \sigma^2) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log(y_i) - \boldsymbol{x}_i^T \beta^{(1)})^2}{2\sigma^2}\right). \tag{2.10b}$$

While we used a gamma distribution for the non-zero part, other distributions can be used depending on the data, such as the generalized gamma distribution (Manning et al., 2005), log-skew-normal distribution (Chai and Bailey, 2008), and normal distribution after Box-Cox transformation (Liu et al., 2019). In practice, the distribution of positive values can be selected based on a modified Park's test (Manning and Mullahy, 2001).

Besides the two-part framework introduced above, many other models have been proposed to assess data with semi-continuous responses. For example, Tobit model deals with zero-inflation with a censored regression model (Tobin, 1958; Amemiya, 1984). The Tobit model assumes an underlying normal distribution $Y^* \sim N(\mu, \sigma^2)$ and an observation y_i is defined as

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases},$$

where the value is not observed when $y_i^* \leq 0$.

Compared with the two-part model we used, the Tobit model is often suitable for censored responses. However, when zeros are actual response values, the Tobit model is doubtful. In comparison, the two-part model has a better numerical properties, whose likelihood function has a unique maximum (Duan et al., 1984) and more appropriate interpretations (Min and Agresti, 2002). Some other models, such as the compound Poisson exponential dispersion model (Jørgensen, 1987) and ordinal response models (Saei et al., 1996), are able to include both zero and nonzero parts within one model. However, they are usually relatively more difficult to fit and have stronger assumptions on data, which restrict their application in practice. Compared with these models, the two-part model can be a better choice for many applications because of its simplicity in fitting and to interpretation.

2.3 Zero-inflated models for time-series data

In general, there are two types of time-series counting models: "observation-driven" and "parameter-driven" models (Cox et al., 1981). These two types of models are different in how they handle the auto-correlation between observations. For observation-driven models, the temporal correlation between observations is captured as a function of previous outcomes. On the other hand, parameter-driven models assume that the temporal correlation is from an underlying latent process, and responses are assumed to be conditionally independent of each other given the latent process.

These two types of models have their advantages and disadvantages. For example, prediction is straightforward with observation-driven models, and the likelihood is easy to calculate. However, the estimated coefficients are less interpretable compared with those from parameter-driven models. While regression parameters are easier to interpret under parameter-driven models, model estimation is challenging, and it is hard to predict future outcomes.

In this section, we introduce a parameter-driven approach originally designed for time-series data with discrete zero-inflated responses (Yang et al., 2015). This time-series zero-

inflated framework is flexible to account for auto-correlation, zero-inflation and interpretation and have been successfully applied to diverse areas, such as healthcare (Yang et al., 2015; MacDonald and Bhamani, 2018; Al-Wahsh and Hussein, 2019; da Silva et al., 2019) and geography (Alghamdi and Harrington Jr, 2018).

Let $\{z_t\}$ be a stationary auto-regressive process of order p (AR(p)) such that

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \cdots + \phi_p z_{t-p} + \epsilon_t \tag{2.11}$$

where ϵ_t is from a normal distribution with mean 0 and variance σ^2, and $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_p\}$ is the vector of auto-regressive coefficients. For a stationary AR(p), it is necessary that

$$\phi_1 + \cdots + \phi_p < 1 \text{ and } |\phi_p| < 1.$$

Conditioning on the current state z_t, the probability mass function of observation y_t following a ZINB distribution can be defined as:

$$P(y_t = k) = (1 - p_{\beta^{(0)},t}) I(k = 0) + p_{\beta^{(0)},t} \frac{\Gamma(\alpha^{-1} + k)}{\Gamma(k+1)\Gamma(\alpha^{-1})} \left(\frac{\mu_t}{\alpha^{-1} + \mu_t}\right)^k \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_t}\right)^{\alpha^{-1}} \tag{2.12}$$

where $\frac{\alpha^{-1}}{\alpha^{-1}+\mu_t}$ is the probability of success in a negative binomial distribution, α is the dispersion parameter, and μ_t is an intensity parameter. Given covariate vector at time t, $\boldsymbol{x}_t$, μ_t is characterized by log-linear model:

$$\log \mu_t = \boldsymbol{x}_t^T \beta^{(1)} + z_t.$$

As in Section 2.2, $p_{\beta^{(0)},t}$ is the probability of $y_t > 0$ and

$$\text{logit}(p_{\beta^{(0)},t}) = \log\left(\frac{p_{\beta^{(0)},t}}{1 - p_{\beta^{(0)},t}}\right) = \boldsymbol{x}_t^T \beta^{(0)}.$$

Because negative binomial distribution can be treated as a Poisson-gamma mixture (Lawless,

1987), the above dynamic ZINB model can be rewritten in the hierarchical form:

$$\begin{aligned} \boldsymbol{s}_t|\boldsymbol{s}_{t-1} &\sim N(\Phi\boldsymbol{s}_{t-1}, \Sigma), \\ u_t &\sim \text{Bernoulli}(p_{\beta^{(0)},t}), \\ v_t &\sim \text{Gamma}(1/\alpha, \alpha), \\ y_t|\boldsymbol{s}_t, u_t, v_t &\sim \text{Poisson}((1-u_t)v_t\lambda_t), \end{aligned} \tag{2.13}$$

where $\boldsymbol{s}_t = (z_t, \ldots, z_{t-p+1})^T$ is the latent state vector, and Φ and Σ are defined as

$$\Phi = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}.$$

As discussed in Yang et al. (2015), the framework introduced above is general and includes many models as special cases. For example, as α goes to 0, this dynamic ZINB reduces to a dynamic ZIP model. Both the dynamic ZINB and dynamic ZIP models degenerate to ZINB and ZIP models if $\sigma = 0$ and auto-correlation does not exist.

Different methods have been proposed to estimate the framework (2.12), such as data cloning (Lele et al., 2007, 2010). This method will be introduced in Chapter 4, and we will demonstrate how to apply these methods to estimate the time-series models.

Chapter 3

SPARSE GROUP REGULARIZATION FOR SEMI-CONTINUOUS ZERO-INFLATED TRANSPORTATION DATA

This chapter considers the problem of variable selection for semi-continuous zero-inflated (SCZI) models. These models include two parts: a zero-inflated state and a non-zero state that is of a continuous nature. A special group regularization is designed to accommodate the unique structure of two-part SCZI models, and a type of Bayesian information criterion is proposed to select tuning parameters. We evaluate the effectiveness of our model using a zero-inflated dataset coming from a driving simulator study. The data shows drivers are able to stay in the intended lane for the majority of their drive (zero-inflated state). On occasion, some drivers do drift out of their intended driving lane (non-zero continuous state). This data is used to illustrate the variable selection process of the proposed model. Our findings show that individual differences can be captured, which have implications for the design of in-vehicle alerting systems.

3.1 Introduction

Traffic crash-related injuries and deaths are considered a major public health problem in the United States (Miaou et al., 2003). Between 2008 and 2018, more than 30,000 fatal crashes were reported, and 2,000,000 people were injured each year due to vehicle crashes (Administration et al., 2019). Many methods have been proposed to analyze crash data, such as generalized additive models and gamma model (Lord and Mannering, 2010). Among them, the zero-inflated count model is one of the most ubiquitous models and has been widely used in the transportation domain to examine crash frequency (Vangala et al., 2015)

at intersections (Dong et al., 2014), on roadway segments (Liu et al., 2018), freeways (Lord et al., 2005), and for weather-related conditions (Carson and Mannering, 2001).

Given that car crashes are considered rare events, other surrogate measures have been used to predict the likelihood of a crash before it occurs. These measures include lane deviation (Donmez et al., 2007), abrupt braking (Bagdadi, 2013), and pedal misapplications (McGehee et al., 2016). These surrogate measures provide a context in terms of spatial and temporal information for why a crash may occur (Stathopoulos and Karlaftis, 2001). However, most of these measures contain an abundance of zeros for the majority of observations but are of a continuous nature. This added complexity limits the use of the traditional zero-inflated count models. For example, lane deviation is considered one of the most common errors related to car crashes (Donmez et al., 2007). Traditionally, the binary outcome and count of lane departures are used to measure unsafe driving (Albousefi et al., 2014; Uc et al., 2009) and can be evaluated with the zero-inflated count models. However, it is also important to understand how far off a driver may drift off their intended lane (Mills, 2013). The further a person departs the center of the driving lane, the higher the likelihood is to crash into a vehicle or object in an adjacent lane. In most cases, the driver will not deviate to an adjacent lane; this leads to a non-negative outcome that is primarily continuous but has a point mass at zero.

Another example showed the impact of the accelerator and brake pedal force on driving safety McGehee et al. (2016) . Sudden and hard presses of the accelerator or brake pedal usually imply emergency events, while soft presses are considered common driving behaviors. However, with cruise control activated, the drivers will not be touching either pedal, which leads to inflation of zeros. When they do press on the pedals, the actual trajectory of movement is important to be examined. This kind of data with a combination of a positive continuous distribution and a point mass at zero is often called mass mixture data, or semi-continuous zero-inflated data.

Recent studies have demonstrated the application of statistical models on the semi-continuous zero-inflated data in different domains (Liu et al., 2016, 2019), such as biomedicine

(Hyndman and Grunwald, 2000; Wang et al., 2015; Chatterjee et al., 2018), economics (Kawakatsu, 2019), and transportation (Mills, 2013; Chang, 2016). For example, one study analyzed statistical tests for zero-inflated gamma (ZIG) and zero-inflated log-normal (ZILN) models and applied the two models to lane departure datasets Mills (2013). Some have also proposed Bayesian zero-inflated models for spatial semi-continuous emergency department expenditures data (Neelon et al., 2015), while others have shown the efficiency of a Bayesian semi-continuous zero-inflated model for longitudinal data (Ghosh and Albert, 2009). These semi-continuous zero-inflated models usually incorporate a two-part model framework, which can be viewed as a combination of two sub-models: the zero part, which indicates whether the outcome is zero or not, and the nonzero part, which determines the actual value of the response if it is nonzero (Duan et al., 1983).

Despite the success of zero-inflated models in analyzing semi-continuous zero-inflated data, a challenge arises as to which variables should be included in the model. For example, variables such as driver demographics, drivers' interactions with vehicles, weather, traffic and road conditions, can all have an impact driver safety (Chang, 2016; McGehee et al., 2016).However, potential sparsity and multicollinearity within these variables can affect estimation and prediction accuracy if all variables are included in the model. Efforts have been made to solve this problem by integrating zero-inflated models with variable selection methods, such as lasso and group lasso (Wang et al., 2015; Chatterjee et al., 2018). However, two major issues remain. First, while these models achieve satisfying performance, it cannot perform group variable selection for semi-continuous zero-inflated data. Second, one of the unique features of zero-inflated models is that coefficients from different parts of the model (i.e., the zero part, nonzero part) can correspond to the same variable and be related (Moulton et al., 2002). In other words, a variable can be associated with both the likelihood and magnitude of the response deviating from zero. In this case, treating coefficients from different parts equally in the variable selection process ignores the underlying connection between coefficients corresponding to the same variable, which can lead to less interpretable results.

Hence, a method to better select the variables related to the unique features of zero-inflated semi-continuous models is required. In this chapter, we propose to incorporate sparse group lasso regularization (Simon et al., 2013) with semi-continuous zero-inflated (SCZI) models, where the groups are carefully constructed to address the underlying relationship between coefficients corresponding to the same variable. An empirical Bayesian inference criterion is studied to select tuning parameters efficiently. We also demonstrate the proposed model by applying it to lane departure data that was collected as part of a driving simulator study.

3.2 Related works

Different approaches have been proposed to interpret and evaluate "zero-inflated" responses from different perspectives and for different scenarios. One of the established methods that can be used to assess "zero-inflated" responses is the discrete/continuous model (Washington et al., 2010), which has been widely adopted in transportation to study household vehicle choice, vehicle use (Fang, 2008; Spissu et al., 2009; Cirillo et al., 2017), and commuters' route and departure time choices (Abu-Eisheh and Mannering, 1987; Habib et al., 2009; Habib, 2013). The discrete/continuous model focuses on data generated from a selection process where a nonrandom sample of data is observed in discrete categories (Washington et al., 2010). For example, this kind of model is useful to examine whether a driver decides to stay in their driving lane or move to an adjacent lane. In this case, the magnitude of lane deviation when changing lanes can only be observed when lane changes are made intentionally by the drivers. Thus, it is unknown how much drivers can depart from the lane when changing lanes if they have chosen to stay within the lane, which leads to the selectivity-bias problem (Mannering and Hensher, 1987). Here, the zero-inflated responses result from subjects' discrete behavior choices, e.g., whether or not to stay within a line.

Another line of works is related to the zero-inflated models. This includes zero-inflated count models (Lambert, 1992; Miaou, 1994) (e.g., ZIP, ZINB), Zero-state Markov switching count-data model (Malyshkina and Mannering, 2010), and semi-continuous zero-inflated

(SCZI) models (Mills, 2013; Liu et al., 2019), such as ZIG and ZILN. These models often assume the existence of two states: (1) a zero state where the probability of a nonzero observation is low, and (2) an imperfect state where observations are from a certain probability distribution (Malyshkina and Mannering, 2010). For example, we consider a lane departure problem where drivers are required to stay in one lane but can drift out of the driving lane due to various reasons that include inattention and neurological impairment, such as Parkinson's disease (Mills, 2013). In such cases, the zero and imperfect states can be treated as whether or not a driver is able to keep the vehicle within the intended lane. The difference between the zero-inflated models and the discrete/continuous model is that, under the assumption of zero-inflated models, subjects do not purposely make choices but switch between different states, which usually does not lead to the selective-bias problem.

The Tobit model (Tobin, 1958) is another model used for semi-continuous outcomes. Under the assumption of the Tobit model, zero values are treated as "censored" observations. Compared with SCZI models, the Tobit model is more appropriate for data with a detection limit. However, if a meaningful definition of the detection limit does not exist, the Tobit model is not plausible. For example, zeros in the lane departure problem are considered to be true zeros and denote that the driver has not drifted outside of their intended lane. Hence, the Tobit and SCZI models answer different questions. Leung and Yu (1996)

The model proposed in this chapter follows the idea of the SCZI models. The proposed model is demonstrated with data from a driving simulator study with semi-continuous zero-inflated lane departures (Chang, 2016). In this study, drivers were told to follow a vehicle that stayed in the same lane for the entire drive. However, while driving, the drivers were not always centered on the leading vehicle. In fact, a driver may drift into an adjacent lane. The goal of this chapter is to develop a variable selection method for an SCZI model related to these lane departures. This includes estimating the model using an approximation method and providing a Bayesian inference criterion to tune parameters.

3.3 Methodology

3.3.1 Sparse group regularization

Extensive literature can be found on variable selection methods for zero-inflated count outcomes (Zeng et al., 2014). However, few studies have focused on developing variable selection methods for semi-continuous zero-inflated models. While it is possible to combine semi-continuous methods with lasso (Kogure, 2015), directly adding the lasso penalties to semi-continuous zero-inflated models ignores the potential relationship between coefficients $\beta_j^{(0)}$ and $\beta_j^{(1)}$ for the same jth variable under the zero and nonzero parts, respectively, by assuming $\beta_j^{(0)}$ and $\beta_j^{(1)}$ are independent and treating them separately in the regularization. In variable selection, we are interested in whether a variable contributes to the overall model, and if it does, to which part of the model it contributes. For this reason, the independence assumption on the coefficients can be inappropriate, and we propose to construct a sparse group regularization that carefully addresses the relationship between coefficients.

Specifically, as discussed in Moulton et al. (2002); Mills (2013), for $\beta_j^{(0)}$ and $\beta_j^{(1)}$ arising from the common variable j, the variable j does not influence the response if both $\beta_j^{(0)} = 0$ and $\beta_j^{(1)} = 0$. On the contrary, the variable j is related to the response if either $\beta_j^{(0)}$ or $\beta_j^{(1)}$ is not equal to 0. In this case, variable selection typically amounts to the selection of the group of $\beta_j^{(0)}$ and $\beta_j^{(1)}$ for the variable j, rather than selecting $\beta_j^{(0)}$ or $\beta_j^{(1)}$ individually.

This underlying grouping property of coefficients motivates us to integrate the group lasso regularization with the semi-continuous ZI models where coefficients corresponding to the same variable are grouped. Without loss of generality, we assume the p variables appear in both zero and nonzero parts. For simplicity of notation, let $\beta_j = (\beta_j^{(0)}, \beta_j^{(1)}) \in \mathbb{R}^2$ be the vector of coefficients corresponding to the variable j for $j = 1, \ldots, p$. Following the group lasso model (Yuan and Lin, 2006), we consider the regularization problem:

$$\min_{\beta,\theta} l(\beta, \theta) + \lambda \left(\sum_{j=1}^{p} \sqrt{2} \|\beta_j\|_2 \right), \tag{3.1}$$

where $l(\beta, \theta) = -\log L(\beta, \theta)$ is the negative log likelihood function in Eq.(2.8), $\|\cdot\|_i$, $i = 1, 2$

denote the l_i norms, and $\lambda \in \mathbb{R}^+$ is the tuning parameter.

While group lasso in Eq.(3.1) can select groups of coefficients that are related to the outcome, it does not yield sparsity within a selected group of coefficients. For example, if a group of coefficients β_j is selected, both $\beta_j^{(0)}$ and $\beta_j^{(1)}$ corresponding to variable j can be nonzero, and deciding if the variable j contributes to the zero part or the nonzero part becomes difficult. We therefore add extra l_1 lasso penalties to β to obtain both sparsity of groups and within each group (Friedman et al., 2010a). Further, since $\beta^{(0)}$ and $\beta^{(1)}$ are from different models (e.g., logistic regression and gamma regression in zero-inflated gamma model), equally penalizing them is unfair and can produce biased estimates for the relatively large coefficients (Fan and Li, 2001). We address this problem by incorporating the adaptive lasso model (Zou, 2006). As shown in previous studies (Zou, 2006), adaptive lasso methods are able to consistently identify the true model by assessing the relative importance of coefficients. This is especially favorable in the proposed model, where coefficients from different parts may have different scales. Hence, we focus on the following problem:

$$\min_{\beta,\theta} l(\beta,\theta) + \lambda \sum_{j=1}^{p} \left\{ (1-\alpha)\gamma_j \|\beta_j\|_2 + \alpha \left(\sum_{k=0}^{1} \xi_j^{(k)} |\beta_j^{(k)}| \right) \right\}, \tag{3.2}$$

where $\alpha \in [0,1]$ suggests a convex combination of the lasso and group lasso penalties, and γ and ξ are the adaptive weights. When $\alpha = 1$, the problem degenerates to the group lasso problem; when $\alpha = 0$, the problem reduces to the original lasso problem. One straightforward observation is that, if no variable appears in both parts, i.e., $\beta_j^{(0)} \times \beta_j^{(1)} = 0$ for all $j = 1, \ldots, p$, the regularization problem in (3.2) degenerates to an ordinary lasso regularization problem. The adaptive weights γ and ξ are estimated as $\tilde{\gamma}_j = 1/\|\tilde{\beta}_j\|^t$ and $\tilde{\xi}_j^{(k)} = 1/|\tilde{\beta}_j^{(k)}|^t$ for $k = 0, 1$, $j = 1, \ldots, p$, where $\tilde{\beta}_j^{(k)}$ is the maximum likelihood estimate (MLE) of $\beta_j^{(k)}$, and t is a tuning parameter (Zou, 2006). In this chapter, we set $t = 1$ for all the experiments.

Note that, because $l(\beta,\theta) = l_0(\beta^{(0)}) + l_1(\beta^{(1)},\theta)$ given β, and θ is not in the penalty term, we can always find the estimate of θ by minimizing $l_1(\beta^{(1)},\theta)$ with regard to θ for a given $\beta^{(1)}$. Therefore, for the rest of the chapter, we suppress θ and write $l(\beta,\theta)$ as $l(\beta)$ for ease of notation and focus on estimating β with the sparse group lasso penalty.

3.3.2 Least squares approximation algorithm

Solving (3.2) directly can be challenging due to the complicated loss functions in (2.9) and (2.10). Instead, we consider an approximation of (3.2) with the least squares approximation (LSA) method (Wang and Leng, 2007) .

The basic idea of LSA is to approximate the loss function $l(\beta)$ with its second order Taylor expansion of $l(\beta)$ at the MLE $\tilde{\beta}$. The Taylor expansion gives

$$l(\beta) \approx l(\tilde{\beta}) + \frac{\partial l(\tilde{\beta})}{\partial \beta}(\beta - \tilde{\beta}) + \frac{1}{2}(\beta - \tilde{\beta})^T \frac{\partial^2 l(\tilde{\beta})}{\partial \beta^2}(\beta - \tilde{\beta}) \triangleq \tilde{l}(\beta). \tag{3.3}$$

Because $\tilde{\beta}$ minimizes $l(\beta)$, we have $\frac{\partial l(\tilde{\beta})}{\partial \beta} = 0$. Meanwhile, we have $E\left(\frac{\partial^2 l(\tilde{\beta})}{\partial \beta^2}\right) \approx \Sigma^{-1}$, where $\Sigma = Var(\tilde{\beta})$ is the covariance matrix of $\tilde{\beta}$. Therefore, $\tilde{\Sigma}^{-1} = \frac{\partial^2 l(\tilde{\beta})}{\partial \beta^2}$ is a natural estimate of Σ^{-1} (Wang and Leng, 2007). By ignoring the constant $l(\tilde{\beta})$ and $1/2$, we have the following asymptotically equivalent least squares problem:

$$\hat{\beta} = \arg\min_{\beta}(\beta - \tilde{\beta})^T \tilde{\Sigma}^{-1}(\beta - \tilde{\beta}) + \lambda \sum_{j=1}^{p} \left\{ (1-\alpha)\tilde{\gamma}_j \|\beta_j\|_2 + \alpha \left(\sum_{k=0}^{1} \tilde{\xi}_j^{(k)} |\beta_j^{(k)}| \right) \right\}, \tag{3.4}$$

which can be efficiently solved by established algorithms that have already been developed for linear models (Simon et al., 2013).

In practice, the MLE of β can be obtained via established methods (Nelder and Wedderburn, 1972), and the covariance matrix $\tilde{\Sigma}$ can be calculated with bootstrap or data perturbation method (Shen and Ye, 2002). To reduce the computational cost, we consider the simplified problem assuming $\beta^{(0)}$ and $\beta^{(1)}$ are independent in calculating $\tilde{\Sigma}$. This allows us to express the negative log-likelihood function $l(\beta)$ as the sum of two negative log likelihood functions of the zero and nonzero parts, i.e., $l(\beta) = l_0(\beta) + l_1(\beta) = l_0(\beta^{(0)}) + l_1(\beta^{(1)})$. The estimates of covariances of $\beta^{(0)}$ and $\beta^{(1)}$ can be derived directly from the estimated generalized linear models (Nelder and Wedderburn, 1972). Consequently, the approximate $\tilde{\Sigma}$ can be calculated efficiently as

$$\frac{\partial^2 l(\tilde{\beta})}{\partial \beta^2} = \frac{\partial^2 l_0(\tilde{\beta})}{\partial \beta^2} + \frac{\partial^2 l_1(\tilde{\beta})}{\partial \beta^2} = \begin{pmatrix} \tilde{\Sigma}_0 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \tilde{\Sigma}_1 \end{pmatrix} = \begin{pmatrix} \tilde{\Sigma}_0 & 0 \\ 0 & \tilde{\Sigma}_1 \end{pmatrix},$$

where $\tilde{\Sigma}_0 = \frac{\partial^2 l_0(\tilde{\beta}^{(0)})}{\partial(\beta^{(0)})^2}$ and $\tilde{\Sigma}_1 = \frac{\partial^2 l_1(\tilde{\beta}^{(1)})}{\partial(\beta^{(1)})^2}$. Note that the assumption that $\beta^{(0)}$ and $\beta^{(1)}$ are independent is not necessarily valid, especially when there are coefficients corresponding to the same variables, and is only for simplifying the calculation of the approximate $\tilde{\Sigma}$. The forthcoming section uses simulated and empirical data to show that good results can still be achieved with this approximate $\tilde{\Sigma}$ nonetheless, where the relationship between coefficients can be captured by the proposed sparse group lasso penalty.

Although we can estimate the MLE of β with sufficient observations and a moderate number of variables, if the number of variables is much larger than the number of observations, these methods may not work. This leads to difficulty in implementing the proposed method. As discussed in Wang and Leng (2007), one possible approach is to use the l_2-penalized models instead and substitute the $\tilde{\beta}$ and $\tilde{\Sigma}$ in (3.4) for the regularized estimates $\tilde{\beta}_\kappa$ and $\tilde{\Sigma}_\kappa$ for some tuning parameter κ, where κ should be set to small values to avoid double-shrinkage effect (Zou and Hastie, 2005). Established packages are available for estimating $\tilde{\beta}_\kappa$, e.g., glmnet (Friedman et al., 2010b) for penalized logistic and log-normal regressions and HDtweedie (Qian et al., 2016) for penalized gamma regression.

3.3.3 Tuning parameter selection

Practically, we need to choose the values of the tuning parameters α and λ according to the prediction accuracy. While cross-validation and generalized cross-validation have been widely used (Tibshirani, 1996), they are computationally intensive and tend to overfit models (Wang et al., 2007). In this section, we introduce a type of revised Bayesian information criterion (BIC_r) to select the tuning parameter.

Wang and Leng (2007) suggested that the BIC_r for LSA approximation takes the form:

$$\text{BIC}_r = (\hat{\beta} - \tilde{\beta})^T \tilde{\Sigma}^{-1} (\hat{\beta} - \tilde{\beta}) + \log n \times df, \tag{3.5}$$

where df is the degrees of freedom. Previous research has shown that for the classic lasso model, the degrees of freedom is approximately equal to the number of nonzero coefficients

(Zou et al., 2007), i.e.,

$$\hat{df}^{l} = \sum_{j=1}^{p} \sum_{k=0}^{1} I(|\hat{\beta}_j^{(k)}| > 0).$$

For the group lasso model, the degrees of freedom can be estimated as

$$\hat{df}^{g} = \sum_{j=1}^{p} \left\{ I(\|\hat{\beta}_j\|_2 > 0) + \frac{\|\hat{\beta}_j\|_2}{\|\tilde{\beta}_j\|_2} \right\}.$$

However, to the best of our knowledge, few works explicitly define the degrees of freedom for the sparse group lasso (Matsui, 2017). Here, instead of directly calculating the degrees of freedom for the model proposed in (3.2), we take the strategy of deriving simple formulae for the approximation of (3.4) instead in a special case of orthogonality (Tibshirani, 1996; Yuan and Lin, 2006). We propose the following degrees of freedom for the proposed model (3.4):

$$\hat{df} = \sum_{j=1}^{p} \sum_{k=0}^{1} \frac{\partial \hat{\beta}_j^{(k)}}{\partial \tilde{\beta}_j^{(k)}}, \tag{3.6}$$

where

$$\begin{aligned} \frac{\partial \hat{\beta}_j^{(k)}}{\partial \tilde{\beta}_j^{(k)}} =& I\left(\|S(\tilde{\beta}_j, \alpha\lambda)\|_2 > (1-\alpha)\lambda \right) I(|\tilde{\beta}_j^{(k)}| > \alpha\lambda) \\ & \times \left\{ 1 - (1-\alpha)\lambda \frac{\|S(\tilde{\beta}_j, \alpha\lambda)\|_2^2 - S^2(\tilde{\beta}_j, \alpha\lambda)_k}{\|S(\tilde{\beta}_j, \alpha\lambda)\|_2^3} \right\}, \end{aligned}$$

and $S(\cdot)$ is the element-wise soft thresholding operator:

$$(S(z, \alpha\lambda))_k = \text{sign}(z_k) \left(|z_k| - \alpha\lambda \right)_+ .$$

The numerical tests show that this approximate BIC_r achieves good overall performance compared with a five-fold cross-validation and outperforms cross-validation in terms of variable selection.

Proposition 1. *Consider the model proposed in (3.4) and $\hat{df}$ in (3.6). Under a set of conditions, we have $df = E[\hat{df}]$.*

Proof. In this proof, we consider the special case where the covariance matrix Σ is diagonal. For the rest of the proof, we suppress the adaptive weights $\tilde{\gamma}_j$ and $\tilde{\xi}_j^{(k)}$ for ease of notation. Then the model (3.4) can be rewritten as:

$$\hat{\beta} = \arg\min_{\beta} \{\tilde{\Sigma}^{-\frac{1}{2}}(\beta - \tilde{\beta})\}^T \tilde{\Sigma}^{-\frac{1}{2}}(\beta - \tilde{\beta}) + \lambda \sum_{j=1}^{p} \left\{ (1-\alpha)\|\beta_j\|_2 + \alpha \left(\sum_{k=0}^{1} |\beta_j^{(k)}| \right) \right\}, \quad (3.7)$$

which can be viewed as a regularized weighted normal linear problem, assuming $\hat{\beta}$ and $\tilde{\beta}$ are jointly normally distributed. For an estimate $\hat{\beta}$, it is known that

$$df = \sum_{j=1}^{p} \sum_{k=0}^{1} cov(\hat{\beta}_j^{(k)}, \tilde{\beta}_j^{(k)}) / var(\tilde{\beta}_j^{(k)}).$$

As shown by in Simon et al. (2013), for group j's solution $\hat{\beta}_j$ of problem (3.7), considering the subgradient conditions, we have $\hat{\beta}_j = 0$ if

$$\|S(\tilde{\beta}_j, \alpha\lambda)\|_2 \leq (1-\alpha)\lambda,$$

otherwise, $\hat{\beta}_j$ satisfies

$$\left(1 + \frac{(1-\alpha)\lambda}{\|\hat{\beta}_j\|_2}\right) \hat{\beta}_j = S(\tilde{\beta}_j, \alpha\lambda). \quad (3.8)$$

Taking l_2 norm on both side of (3.8) and plugging it back into (3.8), we have the solution to (3.7):

$$\hat{\beta}_j = \left(1 - \frac{(1-\alpha)\lambda}{\|S(\tilde{\beta}_j, \alpha\lambda)\|_2}\right)_+ S(\tilde{\beta}_j, \alpha\lambda), \quad (3.9)$$

and

$$\hat{\beta}_j^{(k)} = \left(1 - \frac{(1-\alpha)\lambda}{\|S(\tilde{\beta}_j, \alpha\lambda)\|_2}\right)_+ \text{sign}(\tilde{\beta}_j^{(k)})(|\tilde{\beta}_j^{(k)}| - \alpha\lambda)_+. \quad (3.10)$$

Following (3.9) and (3.10), we have

$$\begin{aligned}\frac{\partial\hat{\beta}_j^{(k)}}{\partial\tilde{\beta}_j^{(k)}} &= I\left(\|S(\tilde{\beta}_j,\alpha\lambda)\|_2 > (1-\alpha)\lambda\right) I(|\tilde{\beta}_j^{(k)}| > \alpha\lambda) \\ &\quad \times \left(1 - 2\alpha\lambda\delta(\tilde{\beta}_j^{(k)})\right)\left\{\left(1 - \frac{(1-\alpha)\lambda}{\|S(\tilde{\beta}_j,\alpha\lambda)\|_2}\right) + \frac{(1-\alpha)\lambda}{\|S(\tilde{\beta}_j,\alpha\lambda)\|_2^3} S^2(\tilde{\beta}_j,\alpha\lambda)_k\right\} \\ &= I\left(\|S(\tilde{\beta}_j,\alpha\lambda)\|_2 > (1-\alpha)\lambda\right) I\left(|\tilde{\beta}_j^{(k)}| > \alpha\lambda\right) \\ &\quad \times \left\{1 - (1-\alpha)\lambda\frac{\|S(\tilde{\beta}_j,\alpha\lambda)\|_2^2 - S^2(\tilde{\beta}_j,\alpha\lambda)_k}{\|S(\tilde{\beta}_j,\alpha\lambda)\|_2^3}\right\},\end{aligned}$$

where $\delta(\cdot)$ is the Dirac delta function. The second equation holds because $\delta(\tilde{\beta}_j^{(k)}) = 0$ for $|\tilde{\beta}_j^{(k)}| > \alpha\lambda$.

As a result, we define

$$\hat{df} = \sum_{j=1}^{p}\sum_{k=0}^{1}\frac{\partial\hat{\beta}_j^{(k)}}{\partial\tilde{\beta}_j^{(k)}}.$$

Now, Stein's lemma yields

$$df = \sum_{j=1}^{p}\sum_{k=0}^{1} cov(\hat{\beta}_j^{(k)},\tilde{\beta}_j^{(k)})/var(\tilde{\beta}_j^{(k)}) = E\left[\sum_{j=1}^{p}\sum_{k=0}^{1}\frac{\partial\hat{\beta}_j^{(k)}}{\partial\tilde{\beta}_j^{(k)}}\right] = E[\hat{df}],$$

and this concludes the proof of Proposition 1. □

Lemma 1. *If $\alpha = 0$, the $\hat{df}$ proposed in (3.6) degenerates to the $\hat{df}^g$ of the group lasso.*

Proof. Note that if $\alpha = 0$, the inequality $|\tilde{\beta}_j^{(k)}| > \alpha\lambda$ always holds for $k = 0, 1$ and $j = 1,\ldots,p$, and we have

$$\hat{df} = \sum_{j=1}^{p} I\left(\|S(\tilde{\beta}_j,\alpha\lambda)\|_2 > (1-\alpha)\lambda\right) + \sum_{j=1}^{p}\left\{1 - \frac{(1-\alpha)\lambda}{\|S(\tilde{\beta}_j,\alpha\lambda)\|_2}\right\}_+.$$

Furthermore, when $\alpha = 0$, $(S(z,0))_k = z_k$. Consequently, we have

$$\hat{df} = \sum_{j=1}^{p} I\left(\|\tilde{\beta}_j\|_2 > \lambda\right) + \sum_{j=1}^{p}\left(1 - \frac{\lambda}{\|\tilde{\beta}_j\|_2}\right)_+ = \sum_{j=1}^{p} I\left(\|\hat{\beta}_j\|_2 > 0\right) + \sum_{j=1}^{p}\frac{\|\hat{\beta}_j\|_2}{\|\tilde{\beta}_j\|_2},$$

which yields Lemma 1. □

Lemma 2. *If $\alpha = 1$, the $\hat{df}$ proposed in (3.6) degenerates to the $\hat{df}^l$ of the original lasso.*

Proof. If $\alpha = 1$, $\|S(\tilde{\beta}_j, \alpha\lambda)\|_2 > 0$ always holds for $k = 0, 1$ and $j = 1, \ldots, p$ if $|\tilde{\beta}_j^{(k)}| > \lambda$. Then, based on (3.9), we have

$$\hat{df} = \sum_{j=1}^{p} \sum_{k=0}^{1} I(|\tilde{\beta}_j^{(k)}| > \alpha\lambda) = \sum_{j=1}^{p} \sum_{k=0}^{1} I(|\hat{\beta}_j^{(k)}| > 0),$$

which yields Lemma 2. □

3.3.4 Asymptotic property

Following the proof in Chatterjee et al. (2012) it can be shown that the proposed method is statistically consistent.

Proposition 2. *Let β^* be the true coefficient vector from which the data samples are generated. Under a set of conditions, for some constant c and $d > 0$, if*

$$\lambda \geq 2c \left\{ \frac{2(1+\sqrt{2})}{\sqrt{n}} + \frac{\sqrt{2(d+1)(2\log p + \log 2)}}{\sqrt{n}} \right\},$$

we have

$$\|\hat{\beta} - \beta^*\|_2^2 = O_p\left(\frac{\log 2p}{n}\right).$$

Proof. In this proof, we focus on the approximate loss function while ignoring the constants:

$$\tilde{l}(\beta) = \{\tilde{\Sigma}^{-\frac{1}{2}}(\beta - \tilde{\beta})\}^T \tilde{\Sigma}^{-\frac{1}{2}}(\beta - \tilde{\beta}).$$

The proof of this proposition is an immediate consequence of applying Theorem 2 and Corollary 1 of Chatterjee et al. (2012) under the same conditions. To formally establish the theory, let A be any subspace of $\mathbb{R}^p$. We consider the following conditions:

(1) For some constant c and $d > 0$,

$$\lambda \geq 2c \left\{ \frac{2(1+\sqrt{2})}{\sqrt{n}} + \frac{\sqrt{2(d+1)(2\log p + \log 2)}}{\sqrt{n}} \right\}.$$

(2) The true coefficient vector β^* is in the subspace A.

The key requirements to meet Theorem 2 of Chatterjee et al. (2012) for the proposed problem (3.4) are:

(1) The regularizer is decomposable.

(2) The loss function $\tilde{l}(\beta)$ should satisfy the Restricted Strong Convexity property.

The first requirement is a property of the sparse group regularizer

$$r(\beta) = \lambda \sum_{j=1}^{p} \left\{ (1-\alpha)\tilde{\gamma}_j \|\beta_j\|_2 + \alpha \left(\sum_{k=0}^{1} \tilde{\xi}_j^{(k)} |\beta_j^{(k)}| \right) \right\}.$$

Following Chatterjee et al. (2012) a regularizer r is decomposable with regard to a subspace pair $A \subseteq B \subseteq \mathbb{R}^p$, if, for any $a \in A$ and $b \in B^{\perp}$, where $B^{\perp}$ is the orthogonal space of B, we have $r(a+b) = r(a) + r(b)$. For the sparse group lasso, the regularizer $r(\beta)$ is decomposable over the subspace spanned by each group given that the $\|\beta_j\|_2$ is over disjoint groups (Chatterjee et al., 2012).

The second requirement is on the loss function $\tilde{l}(\beta)$. Instead of going into mathematical details of the definition of restricted strong convexity, we focus on showing that under assumptions, our approximate loss function satisfies an equivalent condition of the restricted strong convexity. Negahban et al. explains that the restricted strong convexity with respect to the l_2 norm is equivalent to requiring that the design matrix $\boldsymbol{X}$ satisfies a type of restricted eigenvalue condition (Negahban et al., 2012). If each row of the design matrix $\boldsymbol{X}$ is independently sampled from a normal distribution with 0 expectation, then with high probability, the restricted eigenvalue condition holds. In (3.4), we may treat $\tilde{\Sigma}^{-\frac{1}{2}}$ as the design matrix. If we assume that $\tilde{\Sigma}$ is from an inverse-Wishart distribution and each row of $\tilde{\Sigma}^{-\frac{1}{2}}$ is from a multivariate normal distribution with 0 mean, as discussed above, the restricted strong convexity holds with high probability for $\tilde{l}(\beta)$.

Then, as an application of Theorem 2 and Corollary 1 of (Chatterjee et al., 2012), under the two conditions above, because the two requirements are satisfied, for some constant $c_{\tilde{l}}$ and

function $\tau_{\tilde{l}}(\beta)$ related to the loss function $\tilde{l}(\beta)$ in (3.3), with probability at least $1 - \frac{1}{2^{d-1}p^{2d}}$, we have

$$\|\hat{\beta} - \beta^*\|_2^2 \leq \frac{9\lambda^2}{c_{\tilde{l}}^2} s_A + \frac{2\lambda}{c_{\tilde{l}}} \tau_{\tilde{l}}^2(\beta^*) = O_p\left(\frac{\log 2p}{n}\right).$$

Details of the definitions of the restricted strong convexity, restricted eigenvalue condition, $c_{\tilde{l}}$, $\tau_{\tilde{l}}(\beta)$ and a general proof are provided in (Chatterjee et al., 2012).

□

3.3.5 Group effect among variables

In previous sections, we focus on the group effect among coefficients from the zero and nonzero parts assuming that there is no group effect among variables. However, as shown in previous studies, grouped variables exist in practical applications, such as in healthcare (Chatterjee et al., 2018). In this section, we extend the method proposed above to capture the group effects among variables.

Following previous notations, we assume the p variables can be divided into G groups where the size of the group g is p_g. We let $\beta_{g,j}$ be the coefficient of the j-th variable from group g for $j = 1, \ldots, p_g$, and $\beta_{g,\cdot}$ is the vector of all coefficients corresponding to variables in group g. Then, the loss function (3.2) can be revised as

$$\min_{\beta,\theta} l(\beta, \theta) + \lambda \sum_{g=1}^{G} \left\{ (1-\alpha)\gamma_g \|\beta_{g,\cdot}\|_2 + \alpha \left(\sum_{j=1}^{p_g} \sum_{k=0}^{1} \xi_{g,j}^{(k)} |\beta_{g,j}^{(k)}| \right) \right\}. \tag{3.11}$$

Similarly, notice that the log-likelihood function remains the same as in (3.2) and the least squares approximation algorithm introduced in Section 3.3.2 makes no assumption on the regularization part of the loss function, we can effectively find the approximate solution of (3.11) with LSA. Furthermore, following the proof of Proposition 1, we have the BIC_r for the revised model as:

$$\text{BIC}_r = (\hat{\beta} - \tilde{\beta})^T \tilde{\Sigma}^{-1} (\hat{\beta} - \tilde{\beta}) + \log n \times \sum_{g=1}^{G} \sum_{j=1}^{p_g} \sum_{k=0}^{1} \frac{\partial \hat{\beta}_{g,j}^{(k)}}{\partial \tilde{\beta}_{g,j}^{(k)}},$$

where

$$\frac{\partial \hat{\tilde{\beta}}_{g,j}^{(k)}}{\partial \tilde{\beta}_{g,j}^{(k)}} = I\left(\|S(\tilde{\beta}_{g,\cdot}, \alpha\lambda)\|_2 > (1-\alpha)\lambda\right) I(|\tilde{\beta}_{g,j}^{(k)}| > \alpha\lambda) \times \left\{1 - (1-\alpha)\lambda \frac{\|(\tilde{\beta}_{g,\cdot}, \alpha\lambda)\|_2^2 - S^2(\tilde{\beta}_{g,j}, \alpha\lambda)_k}{\|S(\tilde{\beta}_{g,\cdot}, \alpha\lambda)\|_2^3}\right\}.$$

3.4 Numerical tests

Prior to applying the proposed method to data from human subjects, we validate the feasibility of the model performance for different scenarios using simulated data.

3.4.1 Evaluation and testing method

Randomly simulated data is used to assess the proposed adaptive sparse group lasso SCZI (ASGL) model by comparing it with the adaptive group lasso SCZI (AGL) model and the adaptive lasso SCZI (AL) model. In addition, we compared the ASGL model with a flexible, semi-parametric cumulative probability model with the logit link and lasso penalty (CPM) that does not make assumptions on the distribution of the responses (Liu et al., 2017).

For the comparison of ASGL, AGL, and AL, the performance of the methods is evaluated in terms of variable selection and prediction. Specifically, the accuracy of coefficients correctly identified (ACCI), the number of identified coefficients (NC), and the root mean squared error (RMSE) are employed to compare the variable selection performance. RMSE and the correlation between true and predicted response values (COR) are used to compare the prediction performance.

We compare ASGL and CPM in terms of variable selection and prediction. Specifically, ACCI and NC are used to compared the variable selection performance. And RMSE and the percentage of zeros in predictions (PZ) are used to compare the prediction performance. The CPM is fitted with the R package *ordinalNet* and tuned based on the BIC (Wurm et al., 2017). Note that if one variable is selected based on CPM, we assume this variable affects both the likelihood and magnitude of the responses that deviate from zero.

The definitions of these metrics can be found in Table 3.1, where TP is the number of coefficients that are set to nonzeros and selected by the models, TN is the number of coefficients that are set to zeros and not selected by the models, $\|.\|_0$ is the l_0 norm denoting the number of nonzero elements, and $Cor(x, y)$ represents the correlation between x and y.

Table 3.1: Definitions of performance metrics.

	Metric	Definition
Variable Selection	ACCI	$\frac{TP+TN}{p}$
	NC	$\|\hat{\beta}\|_0$
	RMSE	$1/\sqrt{2p}\|\hat{\beta} - \beta\|_2$
Prediction	RMSE	$1/\sqrt{n}\|\hat{y} - y\|_2$
	COR	$Cor(\hat{y}, y)$
	PZ	$\|\hat{y}\|_0/n \times 100\%$

We conduct a comparison across scenarios with different numbers of observations n and variables p. First, for some n and p, we randomly generate the matrix $\boldsymbol{X}$ where each entry is from a normal distribution with mean 0 and variance 1. Then, for the zero and nonzero parts, we randomly generate the coefficient vector β from normal distributions with different means and variances. Some coefficients are manually set to zeros in order to evaluate the variable selection performance. Given $\boldsymbol{X}$ and β, the response vectors $\boldsymbol{y}$ are generated following the models described in (2.9).

3.4.2 Results

The experiments are conducted with $n = 300, 400, 500$ and $p = 20, 30, 40$. Under each combination of n and p, 8, 12, and 16 pairs of coefficients, $(\beta_j^{(0)}, \beta_j^{(1)})$, are randomly selected

and set to zeros. Furthermore, 3, 4 and 5 variables with nonzero coefficients are selected, and for each variable, one of its coefficients is set to zero.

We follow the experimental setup in the previous section to generate the datasets. Tuning parameters λ and α are selected based on the BIC_r proposed in Section 3.3.3 and five-fold cross-validation (CV). The tuning parameter α uses values between 0.05 and 0.95 for ASGL so that it does not degenerate to AGL or AL. In this experiment, we focused on the semi-continuous zero-inflated gamma model. Numerical studies show that the semi-continuous zero-inflated log-normal model shares similar results.

The results are shown in Tables 3.2, 3.3, 3.4, and 3.5. Although α is restricted to $(0.05, 0.95)$, ASGL achieves the best performance most of the time from the aspects of variable selection and prediction with the highest accuracy of correctly identified coefficients (ACCI) and the lowest root mean square error (RMSE). The numbers of selected coefficients (NC) across all the experiments suggest that the AGL model tends to select more coefficients compared with the other two models because it lacks variable selection within groups. The AL model often selects fewer coefficients than the other two models because it does not consider the relationship between coefficients from different parts of the model. On the other hand, ASGL achieves a balance between AGL and AL.

A comparison of the results with BIC_r and cross-validation implies that BIC_r is as efficient as cross-validation (CV) in terms of tuning parameters with less computational requirement. Meanwhile, the numbers of selected coefficients imply that CV is more likely to select more coefficients than the proposed BIC_r. One possible explanation is that compared with BIC_r, CV focuses more on improving prediction accuracy by selecting more coefficients.

A comparison between ASGL and CPM implies that ASGL achieves better performance in terms of both variable selection and prediction (Table 3.5). We notice that on average, 86.7% of the predictions are zeros based on the fitted CPMs. In comparison, on average, based on the fitted ASGL models, 50.7% of the predictions are zeros, and based on the ground truth, 49.3% of the responses are zeros. This implies that the CPM model is highly influenced by the zero inflation, which may lead to the less efficient variable selection and

prediction performance than ASGL. This may result from the fact that CPM treats semi-continuous zero-inflated responses as ordered categorical responses (Liu et al., 2017), and thus the frequency of zero can be much larger than that of other values.

Table 3.2: Numerical test results for variable selection under different combinations of n, p, and $\|\beta\|_0$. The numbers in parentheses are standard deviations based on the 20 runs.

		$n = 300$, $p = 20$, $\|\beta\|_0 = 21$			$n = 400$, $p = 30$, $\|\beta\|_0 = 32$			$n = 500$, $p = 40$, $\|\beta\|_0 = 43$		
Criteria		ACCI	NC	RMSE	ACCI	NC	RMSE	ACCI	NC	RMSE
BIC_r	ASGL	**0.862**	19.35	**0.302**	**0.791**	29.44	0.417	**0.836**	40.34	**0.435**
		(0.053)	(3.253)	(0.119)	(0.052)	(5.704)	(0.145)	(0.042)	(3.662)	(0.076)
	AGL	0.829	23.18	0.322	0.781	33.25	0.424	0.813	45.73	0.484
		(0.055)	(3.664)	(0.121)	(0.058)	(5.912)	(0.134)	(0.041)	(4.08)	(0.106)
	AL	0.803	17.06	0.315	0.769	26.38	**0.397**	0.791	31.82	0.465
		(0.088)	(3.705)	(0.147)	(0.053)	(5.931)	(0.151)	(0.059)	(4.926)	(0.105)
CV	ASGL	**0.750**	24.71	0.301	**0.694**	34.53	**0.350**	**0.808**	44.78	**0.393**
		(0.114)	(7.104)	(0.154)	(0.079)	(9.417)	(0.183)	(0.060)	(11.91)	(0.168)
	AGL	0.723	27.88	**0.295**	0.647	37.12	0.401	0.786	52.78	0.394
		(0.126)	(6.304)	(0.156)	(0.102)	(9.205)	(0.185)	(0.112)	(11.54)	(0.164)
	AL	0.718	21.65	0.324	0.683	29.88	0.371	0.774	37.11	0.406
		(0.101)	(7.261)	(0.167)	(0.067)	(9.165)	(0.192)	(0.093)	(11.99)	(0.195)

3.4.3 Stability analysis

As shown in previous works (Simon et al., 2013; Chatterjee et al., 2012), the tuning parameters α and λ plays an important role in variable selection. While we introduced a type of

Table 3.3: Numerical test results for prediction under different combinations of n, p, and $\|\beta\|_0$. The numbers in parentheses are standard deviations based on the 20 runs.

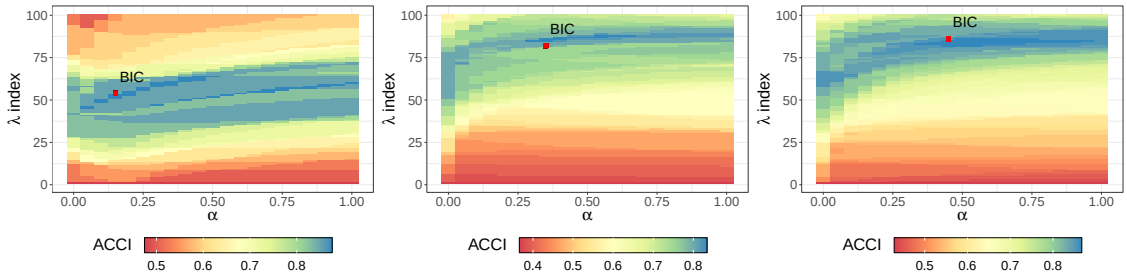
		$n = 300, p = 20, \|\beta\|_0 = 21$		$n = 400, p = 30, \|\beta\|_0 = 32$		$n = 500, p = 40, \|\beta\|_0 = 43$	
Criteria		RMSE	COR	RMSE	COR	RMSE	COR
BIC_r	ASGL	**18.64**	0.898	**56.71**	0.917	**184.7**	0.956
		(5.664)	(0.077)	(15.51)	(0.097)	(50.28)	(0.125)
	AGL	24.17	0.896	57.43	0.915	194.1	0.954
		(5.765)	(0.080)	(17.79)	(0.099)	(52.88)	(0.126)
	AL	18.98	0.898	57.39	0.914	210.8	0.955
		(7.217)	(0.079)	(17.39)	(0.096)	(68.54)	(0.134)
CV	ASGL	20.39	0.899	**57.37**	0.915	**150.7**	0.955
		(7.505)	(0.076)	(16.93)	(0.096)	(50.10)	(0.117)
	AGL	**18.71**	0.900	59.31	0.918	151.9	0.951
		(7.813)	(0.083)	(17.22)	(0.096)	(51.95)	(0.119)
	AL	20.64	0.898	64.91	0.914	172.2	0.944
		(7.383)	(0.085)	(16.95)	(0.093)	(53.69)	(0.114)

Table 3.4: Accuracy of correctly identified coefficients of ASGL for different combinations of n and p on randomly simulated data. The numbers in parentheses are standard deviations based on the 20 runs.

n	$p = 20$	$p = 30$	$p = 40$
300	0.862	0.758	0.671
	(0.053)	(0.036)	(0.071)
400	0.864	0.791	0.743
	(0.059)	(0.052)	(0.051)
500	0.879	0.853	0.836
	(0.050)	(0.045)	(0.042)

Table 3.5: Numerical test results for the comparison of ASGL and CPM in terms of variable selection (ACCI, NC) and prediction (RMSE, PZ) based on the 20 runs.

	$n = 300, p = 20, \|\|\beta\|\|_0 = 21$				$n = 400, p = 30, \|\|\beta\|\|_0 = 32$				$n = 500, p = 40, \|\|\beta\|\|_0 = 43$			
	ACCI	NC	RMSE	PZ	ACCI	NC	RMSE	PZ	ACCI	NC	RMSE	PZ
ASGL	**0.862**	19.35	18.64	49.4	**0.791**	29.44	56.71	49.9	**0.836**	40.34	184.7	49.8
	(0.053)	(3.253)	(5.664)	(2.91)	(0.052)	(5.704)	(15.51)	(2.17)	(0.042)	(3.662)	(50.28)	(2.20)
CPM	0.580	37.70	29.27	84.9	0.559	58.32	72.38	82.5	0.572	77.00	416.6	85.2
	(0.064)	(2.616)	(29.25)	(5.59)	(0.028)	(1.796)	(40.15)	(2.77)	(0.037)	(2.679)	(386.0)	(3.05)
Ground truth	-	21	-	49.5	-	32	-	50.0	-	43	-	50.3
	-	-	-	(2.37)	-	-	-	(2.53)	-	-	-	(1.91)



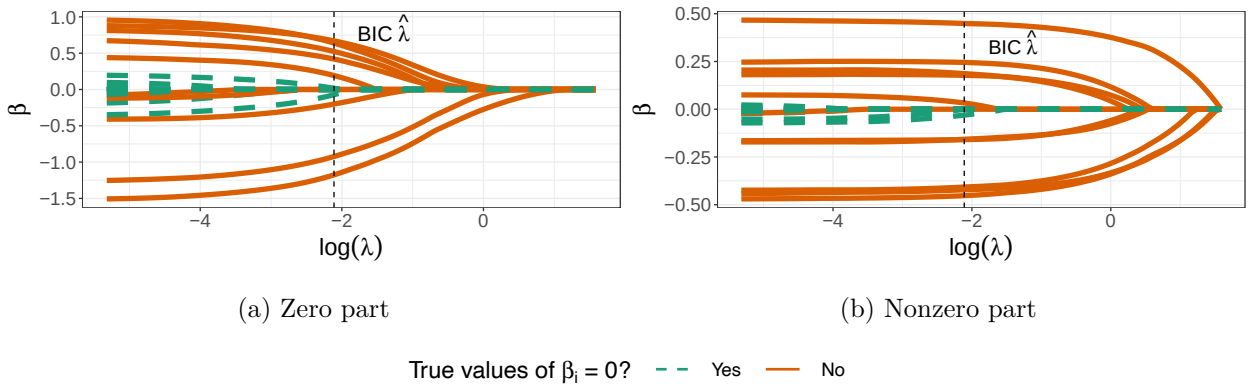
(a) $n = 300$, $p = 20$ (b) $n = 400$, $p = 30$ (c) $n = 500$, $p = 40$

Figure 3.1: Variable selection performance under different combinations of α and λ.

BIC to select the tuning parameters, it is worth analyzing how variable selection changes over different α and λ. We conducted two computer simulations to examine the accuracy of coefficients correctly identified (ACCI) and the changes in the coefficients given different α and λ.

In the first simulation, we use the same randomly generated data as in Section 3.4.2. For each combination of n and p, we created heat maps of ACCI versus α and λ. We choose α from 0 to 1 by 0.05 and fit the model for a path of 100 λ values. Note that when $\alpha = 0$, the model reduces to the group lasso method, and when $\alpha = 1$, the model reduces to the lasso method. We can observe from Fig.3.1 that the proposed ASGL achieves the best performance when compared to the group lasso and lasso method. The performance of variable selection is stable in terms of α and relatively less robust in terms of λ. The choice of α affects the selection of λ, and vice versa. The proposed BIC_r can then help us select the appropriate tuning parameters α and λ.

In the second simulation, we fix the α based on the proposed BIC_r method and compute the regularization path across different λ values. For simplicity, we focus on the regularization paths with $n = 300$, $p = 20$. Our simulation shows that the regularization paths are similar with different combinations of n and p. The results are shown in Fig.3.2. We observe that most coefficients with nonzero values can persist over large ranges of λ under both the zero



(a) Zero part

(b) Nonzero part

Figure 3.2: Regularization path of coefficients of zero and nonzero part for $n = 300$, $p = 20$.

and nonzero parts of the model, while all the coefficients with zero values are not included in the model after a penalty increase. However, some coefficients with values close to zero are not selected as λ gets larger.

3.5 Analysis of lane departure data

In previous studies, unsafe driving among those with the onset of Parkinson's disease, Alzheimer's disease, and depression has been examined using lane deviation as an indicator (Uc and Rizzo, 2008; Bulmash et al., 2006; Mills, 2013). Therefore, it is of interest to examine differences in drivers' behavior related to lane deviation. As the first step, in this section, we apply the proposed semi-continuous zero-inflated model with adaptive sparse group lasso (ASGL) to examine the association between lane departure and drivers' behaviors.

The dataset used was originally collected to assess the cognitive workload of in-vehicle voice control systems (Chang, 2016). Here, we focus on the variables related to lane departure and drivers' behaviors to demonstrate the proposed method. In the original study, 48 participants were recruited (Chang, 2016). There were 6 males and 6 females across four different age groups (18-24, 25-39, 40-54, and 55-75). In the study, participants were asked to drive a NADS MiniSim driving simulator and follow a leading vehicle, which was always

located in one lane. The simulated environment was a four-lane straight flat road with a solid double yellow line down the center. The lead vehicle never changed lanes. The study was approved through the UW Internal Review Board (#45851).

In this chapter, a lane departure is defined as exceeding $\pm 2\sigma$ from the center of the lane, where σ is the standard deviaion. This is a reasonable threshold as $\pm 2\sigma$ is often viewed as the point where an operator should be provided a warning (Green et al., 2003). Finally, given the offsets and threshold, we defined the lane departure responses as $\max\{0, |\text{offset}| - 2\sigma\}$.

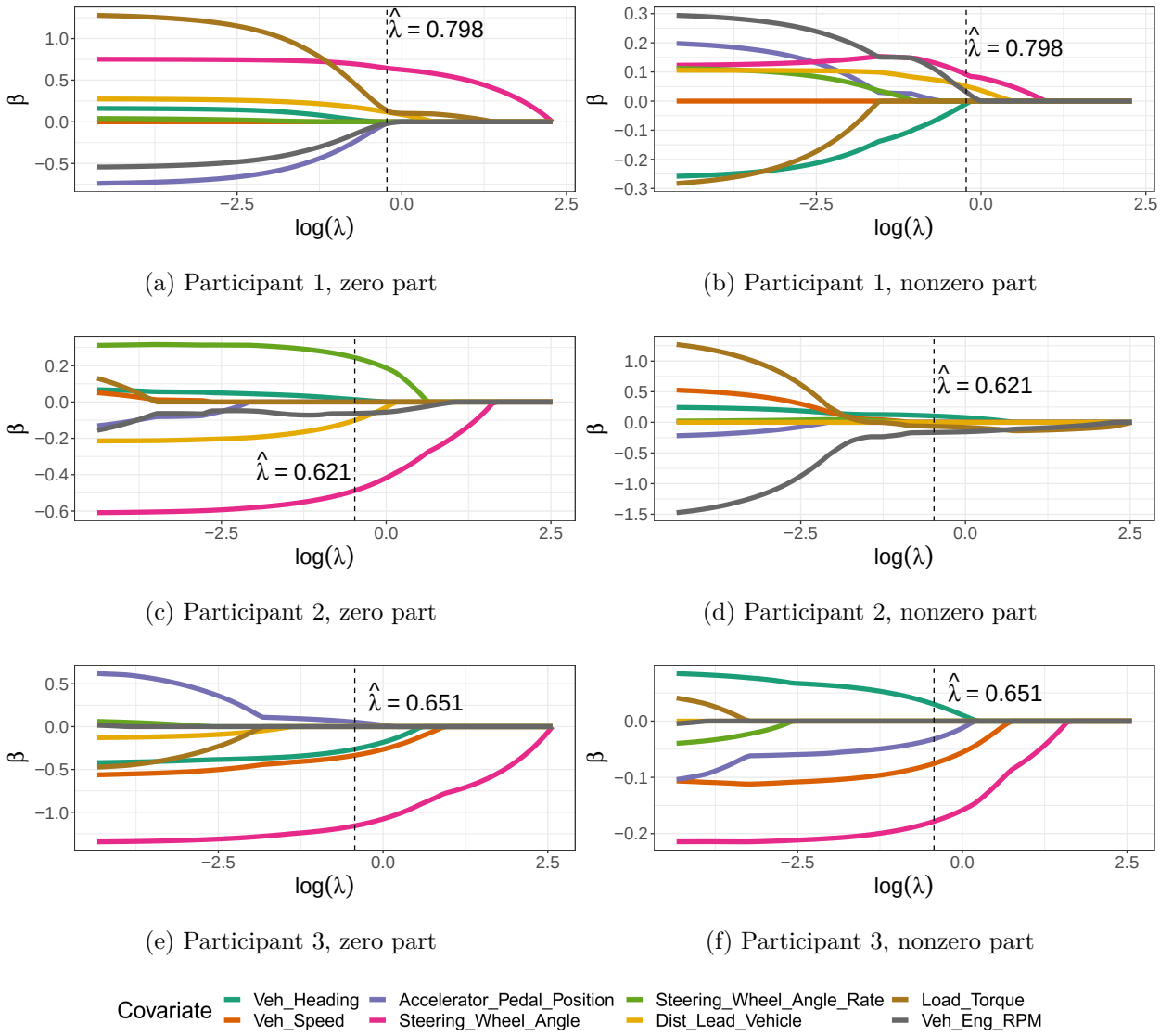
This analysis focused on six out of the 48 participants who departed the lanes more than 5% of the entire drive. The other participants had very few to no lane departures. The data was recorded at 60Hz, and each observation was aggregated up to a two-second window. The length of the drive ranged from 10 and 20 minutes. Eight variables were included in the analysis (Table 3.6).

Table 3.6: Description of variables in the lane departure dataset.

Variable	Description
Veh_Heading	The degree of vehicle heading away from 0 degree
Veh_Speed	The speed of vehicle
Accelerator_Pedal_Position	Accelerator pedal position
Steering_Wheel_Angle	Steering wheel angle in degrees
Steering_Wheel_Angle_Rate	Steering wheel angle rate in degrees/second
Dist_Lead_Vehicle	Distance to lead vehicle in feet
Load_Torque	Wheel torque due to external forces
Veh_Eng_RPM	Engine revolutions per minute

We applied the proposed method to the 6 participants separately to simultaneously assess the influence of variables on the likelihood and magnitude of lane departure. For each participant, we chose α and λ based on the BIC_r and then analyzed how coefficients changed

across different λ for the selected α. The results are shown in Figs. 3.3 (for subject 1-3) and 3.4 (for subjects 4-6). The vertical dashed lines in the graphs represent the values of $\hat{\lambda}$ selected by the proposed BIC_r. If a coefficient is not zero by the time it reaches $\hat{\lambda}$, the corresponding variable is selected and can have an impact on lane departure for that person.



(a) Participant 1, zero part

(b) Participant 1, nonzero part

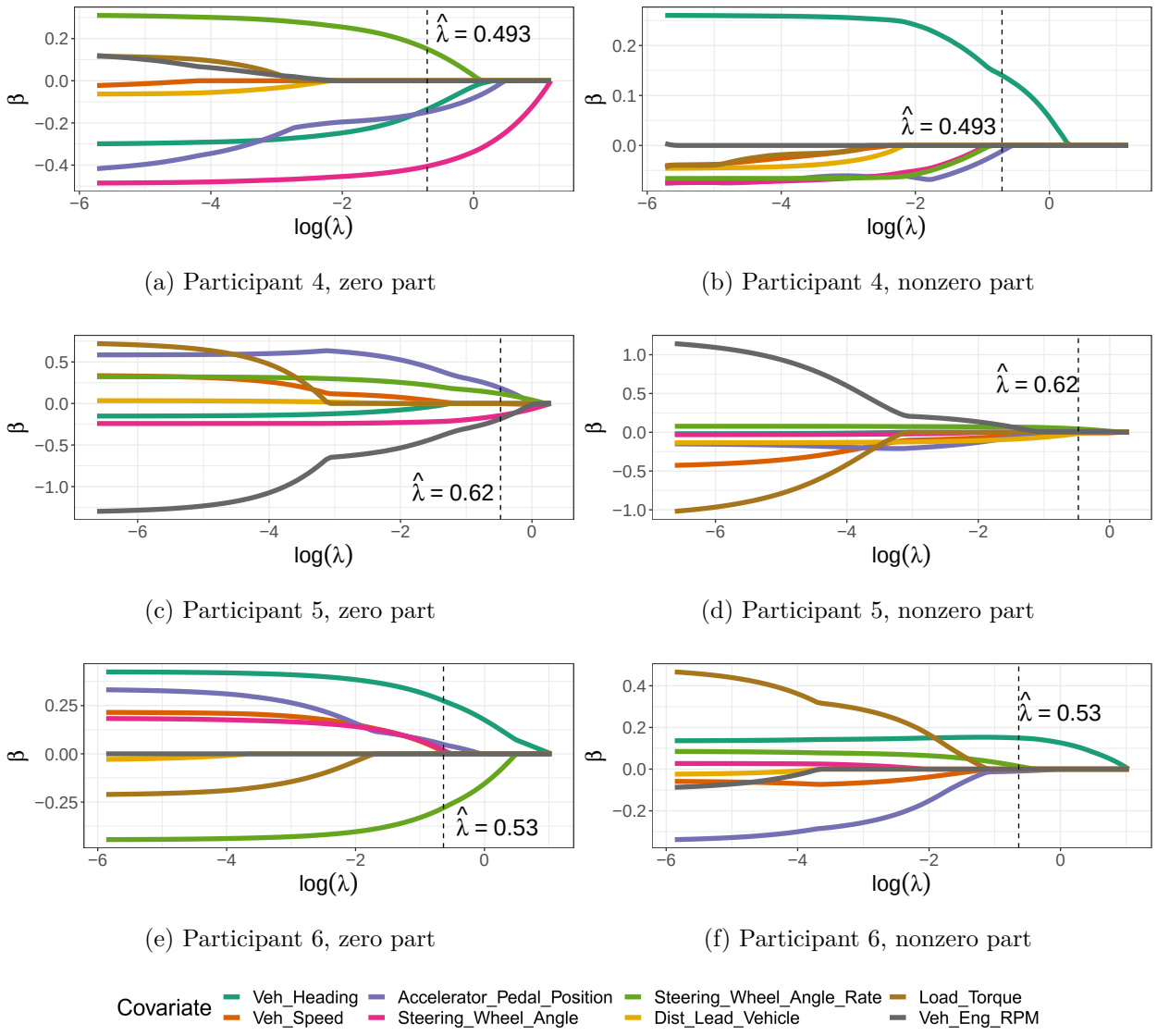
(c) Participant 2, zero part

(d) Participant 2, nonzero part

(e) Participant 3, zero part

(f) Participant 3, nonzero part

Figure 3.3: Regularization path of coefficients of zero and nonzero parts for participants 1-3.



(a) Participant 4, zero part

(b) Participant 4, nonzero part

(c) Participant 5, zero part

(d) Participant 5, nonzero part

(e) Participant 6, zero part

(f) Participant 6, nonzero part

Figure 3.4: Regularization path of coefficients of zero and nonzero parts for participants 4-6.

These graphs indicated that for most of the participants, there existed some variables that share influences on both the zero and nonzero portions of the model. For example, for participants 1 and 3, the steering wheel angle was related to both the likelihood and magnitude of lane departure, but the steering wheel angle rate was much less related to either portion of the model. And for participants 4 and 6, vehicle heading was related to both portions of the model, but the load torque was less related to either portion of the model. These observations coincide with our assumption on the underlying relationship between coefficients corresponding to the same variables. On the other hand, the graphs showed that some variables played different roles in the zero and nonzero portions of the model. For example, for the zero part, the steering wheel angle was strongly associated with lane departure for participants 2, 4, and 6. However, for the nonzero part, the heading of the vehicle was more related to the magnitude of lane departure for these participants. Therefore, the sparse group regularization was necessary for the in-group selection.

We noticed that the scales of the coefficients, β, was different for the zero and nonzero parts. This observation supported the necessity of using adaptive weights in the proposed model. Moreover, the graphs implied that the proposed BIC_r was able to find a reasonable λ. This is based on the fact that we are able to select some variables for the model. For example, the steering wheel angle was highly related to the likelihood of lane departure across all the participants. For most participants (1, 3, 4, 6), the vehicle heading was an important factor in deciding the likelihood and magnitude of lane departure. It is also interesting to note that for some participants (1, 2), the distance to the leading vehicle impacted their lane departure. Finally, each participant has a different coefficient with different magnitude of change; some change was positive, and others were negative. This difference is most likely due to different driving styles and underscores the need to consider individual differences.

3.6 Summary

In this chapter, we constructed a novel adaptive sparse group lasso regularization for semi-continuous zero-inflated (SCZI) models. This model considers the relationship between co-

efficients from the zero and nonzero parts of the SCZI model when selecting variables. We also developed a revised BIC for the proposed model and showed the asymptotic property of the model under a set of conditions. Extensive numerical tests showed that the proposed regularization performed better than the original lasso and group lasso regularizations in terms of variable selection.

We applied the proposed model to data from a driving simulator study that included lane departures. The metric of lane departure is semi-continuous zero-inflated, and the measure is equal zero if a subject stays inside the lane (which was the majority of the time) and greater than zero if a subject drifted outside of the lane. Our proposed method can be used to examine the relationship between driving behaviors and lane departures.

The model was able to capture individual differences in driving. More specifically, the model showed that factors impacted the likelihood as well as the magnitude of lane departure for each individual. For some drivers, lane departure occurred with an abrupt change in the steering wheel angle, while for other drivers, lane departure occurred due to a failure to correct the vehicle heading. While there is a correlation between steering wheel angle and vehicle heading ($\rho = 0.2, p < 0.05$), the differences are based on whether the driver initiated the move (changing the steering wheel) or reacting to a shift due to external conditions (changing as a result of the vehicle heading).

In practice, the proposed model can be used by car manufacturers to develop algorithms to ensure the safety of drivers. With this zero-inflated model, the system will be able to learn the operator's driving styles continuously. The learned custom model will be able to detect when drivers depart from the lane, and appropriately warn the driver in advance. It would also be of interest to use this model to examine the associations between other cognitive impairments and diseases known to impact driving and lane deviation. Note that, as shown in previous studies and the above proof, the lasso-type models achieve consistent variable estimation and selection only under conditions for large sample sizes (Leng et al., 2006; Zhao and Yu, 2006; Wang and Leng, 2008; Nardi et al., 2008; Chatterjee et al., 2012). For small sample sizes, these models can be inconsistent in terms of variable selection (Leng et al.,

2006; Wang and Leng, 2008). Therefore, for practical implementation, one usually needs to justify the variable selection results based on the domain knowledge and context.

The current model can be further improved. The current model uses LSA approximation method to estimate coefficients, and the performance of LSA highly depends on the unpenalized estimate of the covariance matrix of coefficients. This estimate can be inaccurate under certain conditions. For example, when calculating the unpenalized estimate of the covariance matrix of coefficients, we assumed coefficients from zero and nonzero parts were independent based on the log-likelihood function. However, this assumption can be violated if the two parts are highly related to each other. One possible alternative approach to solving this problem is to estimate the covariance matrix with Bayesian methods. The current model may also be improved by using other well-designed penalty functions, such as smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001) and minimax concave penalty (MCP) (Zhang et al., 2010). For real-world applications, the random effect is important to account for the systematic differences within and between subjects (Han et al., 2019). It would be worthwhile to explore ways to account for the random effects in one model (rather than multiple models) while also retaining the asymptotic properties.

Chapter 4

DYNAMIC SEMI-CONTINUOUS ZERO-INFLATED MODEL

This chapter considers the problem of modeling semi-continuous time series data with excess zeros. This kind of data have two distinctive features: semi-continuous zero-inflation and temporal correlation. A parameter-driven dynamic model is developed to accommodate these two features. We apply our method to study drivers' braking behavior in different driving situations. Our findings show that drivers' braking behavior can be influenced by the closest in-path vehicle, and the influence is different in different driving situations. This result has implications for the design of braking assistance systems.

4.1 Introduction

Time series of zero-inflated data is frequently encountered in modern transportation and healthcare studies (Malyshkina et al., 2009; DeSantis and Bandyopadhyay, 2011; Yang et al., 2015; Koh et al., 2019). As a result, the analysis of this kind of data has received growing attentions in the literature. However, previous studies focused mainly on the analysis of time series of zero-inflated count data (Malyshkina and Mannering, 2010; Xiong et al., 2014), and the time series of zero-inflated semi-continuous data was largely neglected.

The time series of zero-inflated semi-continuous data can be found in recent transportation studies. For example, Mills (2013) analyzed the lane deviation of drivers with Parkinson disease. In the analysis, it is important to understand when the driver may deviate from the lane and if lane deviation happens, how far off a driver may drift off their intended lane. Another example considers the influence of driving situations on drivers' driving behavior while developing braking assistance systems (Ervin et al., 2005; McGehee et al., 2016). Sudden and hard presses of the brake pedal usually imply emergency events, while soft presses

are considered common driving behaviors. However, in certain driving situations, such as on the highway without traffic, the drivers will not be touching the brake pedal, which leads excess zeros. When they do press on the brake pedals, the actual force applied to the pedal is important to be examined. This kind of data with a combination of a positive continuous distribution and a point mass at zero is often called mass mixture data, or semi-continuous zero-inflated data. As this kind of datasets was usually collected at consecutive time points, it is necessary to consider the auto-correlation during the data analysis to further improve the model estimation.

In general, two classes of models are widely used for non-Gaussian time series data: "observation driven" and "parameter-driven" (Cox et al., 1981). The major difference between these two classes of models comes from the way they handle auto-correlations. Under the observation-driven models, current parameters are formulated as a deterministic function of past responses. In contrast, parameter-driven models assume that the temporal correlation is from an underlying latent process, and that responses are independent of each other conditioning on the latent process.

Numerous frameworks and models have been developed for zero-inflated time series under the two classes. For example, among the parameter-driven approaches, Yang et al. (2015) proposed a state-space framework for zero-inflated count time series. In DeSantis and Bandyopadhyay (2011), the latent states were represented by the Hidden Markov model. And Zhou et al. (2020) extended the work of DeSantis and Bandyopadhyay (2011) to semi-continuous time-series data. Other examples in this line of works include Hasan et al. (2016); Zhang et al. (2016); Tang and Cavanaugh (2017); MacDonald and Bhamani (2018). Based on the observation-driven framework, Wang et al. (2018) proposed a series of two-part autoregressive models for continuous time-series data with excess zeros. Following the idea of the Gaussian generalized autoregressive conditional heteroskedasticity (GARCH) model (Bollerslev, 1986), a number of models have been developed to incorporate auto-regression and zero-inflated (Zhu, 2012; Gonçalves et al., 2016; Kawakatsu, 2019). Other examples following the observation-driven framework include models based on non-negative integer valued

auto-regressive process (Jazi et al., 2012), multiplicative error model (Hautsch et al., 2014), and censored distributions Harvey and Ito (2020).

Given the different natures of these two classes of models, both have their advantages and limitations (Koopman et al., 2016). For example, the closed-form likelihood function is available for observation-driven models, and this leads to easier estimation and prediction than parameter-driven models. However, the estimated coefficients are less interpretable compared with those from parameter-driven models. On the other hand, the parameter-driven models are more flexible and can be applied to data from complex distributions, and regression coefficients are easier to interpret under parameter-driven models. Nevertheless, compared with observation-driven models, model estimation and prediction under parameter-driven models are more challenging.

Previous study showed the effectiveness of observation-driven dynamic semi-continuous zero-inflated model in predicting stock prices (Kawakatsu, 2019). However, this model lacked the analysis of the influence of covariates on the responses. In this chapter, following the idea of (Yang et al., 2015) mentioned above, we introduce a dynamic semi-continuous zero-inflated model (DSCZI) that focuses on evaluating the relationship between time-varying covariates and auto-correlated responses. This model follows the parameter-driven framework and benefits from its interpretability and flexibility to accommodate both semi-continuous zero-inflation and auto-correlation. A novel data cloning method (Lele et al., 2007) is used for model estimation and statistical inference for its advantages in convergence and statistical inference.

4.2 Parameter-driven dynamic semi-continuous zero-inflated model

In this section, we formulate the dynamic semi-continuous zero-inflated model (DSCZI) for a semi-continuous time series with excess zeros. Suppose that we have data $(\boldsymbol{y}, \boldsymbol{X}^{(1)}, \boldsymbol{X}^{(0)})$, where $\boldsymbol{y} = (y_1, \ldots, y_T)$ is a vector of T responses, y_t is the response at time t. The $T \times m^{(0)}$ matrix $\boldsymbol{X}^{(0)}$ represents the $m^{(0)}$ time-varying variables associated with the probability of obtaining positive responses. $\boldsymbol{X}^{(1)} = (\boldsymbol{x}_1^{(1)}, \ldots, \boldsymbol{x}_T^{(1)})^T$ is a $T \times m^{(1)}$ matrix of $m^{(1)} - 1$ time-

varying variables related to the positive continuous responses, where the first column of $\boldsymbol{X}^{(1)}$ represents the intercept, and $\boldsymbol{x}_t^{(1)}$ is the vector of variables at time t for $t = 1, \ldots, T$.

To accommodate the temporal correlation of the responses, let latent states $\{s_t\}$ be a stationary autoregressive process of order p, AR(p), such that

$$s_t = \phi_1 s_{t-1} + \phi_2 s_{t-2} + \cdots + \phi_p s_{t-p} + \epsilon_t, \tag{4.1}$$

where ϵ_t is from a normal distribution with mean 0 and variance σ^2, and $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_p\}$ is the vector of auto-regressive coefficients. For a stationary AR(p), it is necessary that

$$\phi_1 + \cdots + \phi_p < 1 \text{ and } |\phi_p| < 1.$$

We use logistic regression for the zero part and Gamma regression for the nonzero part as an example to introduce the structure of the proposed dynamic semi-continuous zero-inflated model (DSCZI) model. Conditioning on the current state s_t, y_t follows a DSCZI model if its distribution is given by:

$$f(y_t|s_t) = (1 - p_{0,t})I(y_t = 0) + p_{0,t}g(y_t|s_t, y_t > 0), \tag{4.2}$$

where

$$\begin{aligned} \text{logit}(p_{0,t}) &= \log\left(\frac{p_{0,t}}{1 - p_{0,t}}\right) = (\boldsymbol{x}_t^{(0)})^T\beta^{(0)} + \omega s_t \\ g(y_t|s_t, y_t > 0) &= \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu_t}\right)^{\nu} y_t^{\nu-1}\exp\left(-\frac{\nu y_t}{\mu_t}\right). \end{aligned} \tag{4.3}$$

Here, $\beta^{(0)} = (\beta_1^{(0)}, \ldots, \beta_{m^{(0)}}^{(0)}) \in \mathbb{R}^{m^{(0)}}$ and $\beta^{(1)} = (\beta_1^{(1)}, \ldots, \beta_{m^{(1)}}^{(1)}) \in \mathbb{R}^{m^{(1)}}$ represent the regression coefficients. $p_{0,t} = P(y_t > 0|s_t)$ denotes the probability of $y_t > 0$, $I(\cdot)$ is the indicator function, i.e., $I(y = 0) = 1$ if $y = 0$ and $I(y = 0) = 0$ otherwise, $g(y_t|s_t, y_t > 0)$ represents the conditional distribution of the positive responses, ν is the shape parameter of the gamma distribution, ω is a scale parameter, and μ_t denotes the conditional expectation of y_t given $y_t > 0$. The parameter μ_t is associated with the latent state s_t and covariates at time t, $\boldsymbol{x}_t$, through a link function:

$$\log(\mu_t) = (\boldsymbol{x}_t{}^{(1)})^T\beta^{(1)} + s_t.$$

Note that while we used logistic regression and Gamma regression as an example to show the structure of the proposed DSCZI model, other distributions and models can be used depending on the data. For example, probit regression can be applied for $p_{\beta^{(0)}}$ (Duan et al., 1983). For $g(y|\boldsymbol{x}, y > 0; \beta^{(1)}, \theta)$, the log-normal distribution, generalized gamma distribution (Manning et al., 2005), log-skew-normal distribution (Chai and Bailey, 2008), and normal distribution after Box-Cox transformationLiu et al. (2019) can also be used. The flexibility of the model estimation method introduced in Section 4.3 allows us to easily adopt these distributions without heavily modifying the proposed model.

For the rest of this chapter, we assume the latent state transits follow a stationary AR(p) defined in (4.1). The stationary assumption ensures that the expected value, variance, and correlation between latent states are constant across a time period, so that we can estimate these parameters with some model estimation methods introduced in the following sections. While it is possible to consider auto-regression moving-average (ARMA) instead of AR for the latent state transition under the current model, primitive results showed that the parameter estimation performance would not improve much by adding moving average to the current model. Moreover, adding moving average requires larger sample size for model estimation and is computationally inefficient. Therefore, in this chapter, we focus on AR for latent state transition to demonstrate the proposed model.

4.3 Model estimation

The DSCZI model proposed in (4.2) has the following characteristics: 1) the response distribution is non-Gaussian, and 2) the latent states s_t are unknown and auto-correlated. Because of these characteristics, it is difficult to explicitly derive the marginal likelihood of observations $\boldsymbol{y} = (y_1, \ldots, y_t)$ and maximize the likelihood function with gradient-based methods. In this section, we introduce a novel computational method, data cloning (Lele et al., 2007, 2010), and apply it to estimate the DSCZI model.

4.3.1 Review of data cloning

Extensive literature can be found on the estimation of dynamic zero-inflated models for count data. Examples of these methods include the Monte Carlo EM algorithm (Yang et al., 2015), Gaussian Copula (Alqawba et al., 2019), and data cloning (Lele et al., 2007, 2010; Al-Wahsh and Hussein, 2019). Among these methods, the data cloning approach is attractive for its advantages over the other two methods. For example, the estimates of parameters with data cloning were proved to converge to a multivariate normal distribution with mean equal to their maximum likelihood estimates (MLE) and covariance proportional to the inverse of the Fisher information matrix. These features are especially preferred as the proposed model focuses on analyzing the relationship between covariates and responses.

In this chapter, we propose to use data cloning to estimate the proposed model (4.2). With slight abuse of notations, we introduce the basic idea of data cloning method considering the following hierarchical model:

$$\begin{aligned} &\text{Hierarchy 1: } \boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x} \sim f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_1), \\ &\text{Hierarchy 2: } \boldsymbol{X} \sim g(\boldsymbol{x}|\boldsymbol{\theta}_2), \end{aligned}$$

where $\boldsymbol{y}$ is observed and x is unknown, and the parameters of interest are $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. The goal is to estimate the parameters θ and predict the unknown states x. Then the likelihood function of this hierarchical model is

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \int f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_1) g(\boldsymbol{x}|\boldsymbol{\theta}_2) d\boldsymbol{x}.$$

However, estimating $\boldsymbol{\theta}$ is computationally challenging in terms of (1) the calculation of the likelihood function due to the high-dimensional integration; (2) obtaining the location of the maximum; (3) calculating the standard errors of the estimators (Lele et al., 2010). Data cloning method avoid these challenges in a simple fashion.

The idea of data cloning method comes from a hypothetical situation where observations can be repeated independently by K different individuals and by happenstance all these individuals obtain exactly the same set of observations $\boldsymbol{y}$, denoting as $\boldsymbol{y}_{(K)} = (\boldsymbol{y}, ..., \boldsymbol{y})$. Then

the likelihood function based on $\boldsymbol{y}_{(K)}$ is given by $[L(\boldsymbol{\theta}|\boldsymbol{y})]^K$. Assuming that the parameters are identifiable and there is a unique mode, the posterior distribution of $\boldsymbol{\theta}$ conditional on the observations $y_{(K)}$is

$$\pi_{(K)}(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{[L(\boldsymbol{\theta}|\boldsymbol{y})]^K \pi(\boldsymbol{\theta})}{C(K;\boldsymbol{y})},$$

where $C(K; y_{1:T})$ is the normalization constant. Walker (1969) proved that under conditions, if K is large, $\pi_{(K)}(\boldsymbol{\theta}|\boldsymbol{y})$ is approximately normal distributed with mean $\hat{\theta}$, the maximum likelihood estimator, and variance $\frac{1}{K}\boldsymbol{I}^{-1}(\hat{\theta})$, the Fisher information matrix.

However, data from such K independent experiments usually does not exist. Instead of considering $\pi_{(K)}(\boldsymbol{\theta}|\boldsymbol{y})$ as the posterior distribution of $\boldsymbol{\theta}$ with K independent experiments, Lele et al. (2010) treated it as another distribution consisting of a set of observations $\boldsymbol{y}$ and model components. It can be proved that, even without the independence assumption, when K becomes large, the distribution function $\pi_{(K)}(\boldsymbol{\theta}|y_{1:n})$ converges to a distribution with mean $\hat{\boldsymbol{\theta}}_{1:n}$ and variance $K^{-1}I^{-1}(\hat{\boldsymbol{\theta}}_{1:n})$. Then, we can generate random variates $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots$ from $\pi_{(K)}(\boldsymbol{\theta}|\boldsymbol{y})$ with the Markov chain Monte Carlo (MCMC) method. If K is large enough, theorems in Lele et al. (2010) guarantees that the mean of the random variates converges to the MLE of the parameter $\boldsymbol{\theta}$.

4.3.2 Estimation of parameter-driven dynamic SCZI model with data cloning

In order to apply the data cloning method, we first derive the likelihood function of the proposed DSCZI method. Let $\boldsymbol{s}_t = (s_t, \ldots, s_{t-p+1})^T$, $t = 1, \ldots, T$ be the latent state vectors, $u_t \sim$ Bernoulli$(p_{0,t})$, and assuming that the latent states $\boldsymbol{s}_{0:n}$ and $u_{1:n}$ are observable, the likelihood function of the model can be expressed and decomposed as follows:

$$\begin{aligned} L(\nu, \beta^{(0)}, \beta^{(1)}, \boldsymbol{\phi}, \sigma, \boldsymbol{s}_{1:T}) &= f(\boldsymbol{s}_{0:T}, \boldsymbol{u}_{1:T}, y_{1:n}) \\ &= f(\boldsymbol{s}_{0:T}, u_{1:T}) f(y_{1:T}|\boldsymbol{s}_{0:T}, u_{1:T}) \\ &= f(\boldsymbol{s}_{0:T}) f(u_{1:T}) f(y_{1:T}|\boldsymbol{s}_{0:T}, u_{1:T}) \\ &= f(\boldsymbol{s}_0) \prod_{t=1}^{T} f(\boldsymbol{s}_t|\boldsymbol{s}_{t-1}) \prod_{t=1}^{T} f(u_t) \prod_{t=1}^{T} f(y_t|\boldsymbol{s}_t, u_t). \end{aligned} \tag{4.4}$$

where $\boldsymbol{\theta} = (\nu, \beta^{(0)}, \beta^{(1)}, \boldsymbol{\phi}, \sigma)$ is the vector of unknown parameters. The initial state vector $\boldsymbol{s}_0$ can be assumed to be normally distributed. Our primitive experiments show that the choice of the mean and variance of the initial state has little influence on the parameter estimation, and we can set elements of $\boldsymbol{s}_0$ to be i.i.d. from the standard normal distribution.

Following the likelihood function (4.4), using logistic regression and Gamma regression as an example, assuming the latent process is a Gaussian auto-regression, we can rewrite the DSCZI method in a hierarchical form:

$$\begin{aligned}
\boldsymbol{s}_t|\boldsymbol{s}_{t-1} &\sim N(\Phi\boldsymbol{s}_{t-1}, \Sigma), \\
\log(\mu_t) &= (\boldsymbol{x}_t{}^{(1)})^T\beta^{(1)} + s_t, \\
\text{logit}(p_{0,t}) &= (\boldsymbol{x}_t^{(0)})^T\beta^{(0)} + \omega s_t \\
u_t &\sim \text{Bernoulli}(p_{0,t}), \\
y_t|u_t &\sim \begin{cases} \text{Gamma}(\nu, \frac{\mu_t}{\nu}) & \text{, if } u_t = 1, \\ 0 & \text{, if } u_t = 0, \end{cases}
\end{aligned} \tag{4.5}$$

where Φ and Σ are defined as

$$\Phi = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}.$$

This hierarchical form allows us to apply the data cloning method directly and generate random variates of $\boldsymbol{\theta}$ through MCMC. Then, the MLE and asymptotic fisher information matrix of $\boldsymbol{\theta}$ can be obtained with the mean and variance of the random variates. With the data cloning method, we avoid the evaluation of the high-dimensional integral of the likelihood function with respect to the latent states $\boldsymbol{s}_t$ and the numerical optimization of the marginal likelihood function.

4.4 Numerical study

Prior to applying the proposed method to data from the simulator study, we validate the feasibility of the model performance for different scenarios using simulated data.

We use randomly generated data to assess the proposed parameter-driven dynamic semi-continuous zero-inflated model (DSCZI). We compare DSCZI with static semi-continuous zero-inflated model (SCZI), observation-driven generalized autoregressive moving average model (Benjamin et al., 2003) with zero adjusted gamma distribution (GARMA-ZAGA) from R package *gamlss* (Stasinopoulos et al., 2007), and GARMA with binomial distribution (GARMA-Binomial).

The experiments are conducted with $T = 200, 500, 1000$, $m^{(0)} = 4$, 6, $m^{(1)} = 5, 10$ and $p = 2$. For each combination of T, $m^{(0)}$, $m^{(1)}$ and p, we randomly generate the matrices $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(0)}$ from a normal distribution with mean 0 and variance 1. The coefficients $\beta_i^{(j)}$, $i = 1, \ldots, m^{(j)}$, $j = 0, 1$ and ϕ are randomly generated from normal distributions with different means and variances. The parameters σ and ν are set to 0.15 and 2, respectively. Given $\boldsymbol{X}^{(1)}$, $\boldsymbol{X}^{(0)}$, $\beta_i^{(j)}$, $i = 1, \ldots, m^{(j)}$, $j = 0, 1$, ϕ, σ, and ν, the responses $\boldsymbol{y}$ are generated based on the model described in (4.2).

In the numerical study, the priors of the DSCZI model are chosen as follows: $\beta_i^{(j)} \sim N(0, 100)$, $i = 1, \ldots, m^{(j)}$, $j = 0, 1$, $\phi_i \sim N(0, 100)$, $i = 1, \ldots, p$, $\nu \sim lognormal(0, 100)$, and $\sigma \sim lognormal(0, 100)$. We assume the initial states are from standard normal distribution. The burn-in period is 3000, and three Markov chains are generated with 2000 iterations.

The parameter estimation performance of the models is compared based on the the root mean squared error (RMSE) of $\hat{\beta}^{(0)}$ and $\hat{\beta}^{(1)}$. Here, the RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{1}{m^{(j)}} \sum_{i=1}^{m^{(j)}} (\beta_i^{(j)} - \hat{\beta}_i^{(j)})^2}, \; j = 0, 1,$$

where $\hat{\beta}_i^{(j)}$ is the estimate of $\beta_i^{(j)}$. We repeat the experiment 20 times and report the mean and standard deviation of the final result. Note that because we cannot estimate $\beta^{(0)}$ with GARMA-ZAGA and $\beta^{(1)}$ with GARMA-Binomial, RMSE of $\beta^{(0)}$ and $\beta^{(1)}$ are omitted under

GARMA-ZAGA and GARMA-Binomial, respectively. In this experiment, we used logistic regression for the zero part and Gamma regression for the nonzero part. Numerical studies showed that the results were similar when we used other models for the zero and nonzero parts.

The results are shown in Table 4.1. Overall, the proposed DSCZI achieved the best parameter estimation performance over the other three models with the lowest RMSE. A comparison of DSCZI and two GARMA models implied that DSCZI can estimate the coefficients better than two GARMA models that fits the zero and nonzero parts separately. Comparison of SCZI and the other three models indicated the necessity of considering autocorrelation when analyzing time-series data. Comparison of standard deviations showed that DSCZI is stabler than GARMA and SCZI. Finally, the parameter estimation performance of all the models increase with the increase of sample size, and the performance difference between models become smaller with larger sample size.

We conducted extensive studies to analyze the property of estimators from DSCZI. We set $T = 500$, $m^{(0)} = 3$ and $m^{(1)} = 5$ and used the same prior and parameter settings as in the above comparison study. Parameter estimators were obtained based on the last 500 iterations of data cloning. First, we used Q-Q plots to assess the approximate normality of the random variates used to obtain the estimators. The results are shown in Figure 4.1. These figures suggests that the approximate normality holds for all the empirical distributions of the random variates. The empirical distributions of random variates for the auto-regressive coefficients ϕ_1 and ϕ_2 and standard deviation σ are light-tailed compared with other empirical distributions. Additional simulation studies suggested that this problem was less prominent with larger sample size nonetheless. Table 4.2 gives the true values of parameters as well as the mean, median, and standard errors (SE) of the corresponding random variates. We notice a small bias associated with the auto-regressive coefficients ϕ_1 and ϕ_2 and standard deviation σ. Overall, the means are close to the medians. These results imply that it is reasonable to use data cloning to obtain the MLEs and their asymptotic standard errors.

It is worth noting that under the current numerical tests, GARMA models are "misspeci-

Table 4.1: Numerical test results for model estimation under different combinations of T, $m^{(0)}$ and $m^{(1)}$. The numbers in parentheses are standard deviations based on the 20 runs.

	Estimate of $\beta^{(1)}$						Estimate of $\beta^{(0)}$					
	T = 200	T = 200	T = 500	T = 500	T = 1000	T = 1000	T = 200	T = 200	T = 500	T = 500	T = 1000	T = 1000
	$m^{(0)} = 4$	$m^{(0)} = 6$	$m^{(0)} = 4$	$m^{(0)} = 6$	$m^{(0)} = 4$	$m^{(0)} = 6$	$m^{(0)} = 4$	$m^{(0)} = 6$	$m^{(0)} = 4$	$m^{(0)} = 6$	$m^{(0)} = 4$	$m^{(0)} = 6$
	$m^{(1)} = 5$	$m^{(1)} = 10$	$m^{(1)} = 5$	$m^{(1)} = 10$	$m^{(1)} = 5$	$m^{(1)} = 10$	$m^{(1)} = 5$	$m^{(1)} = 10$	$m^{(1)} = 5$	$m^{(1)} = 10$	$m^{(1)} = 5$	$m^{(1)} = 10$
DSCZI	**0.129**	**0.099**	**0.061**	**0.054**	**0.045**	**0.046**	**0.242**	**0.332**	**0.180**	**0.189**	**0.099**	**0.132**
	(0.051)	(0.018)	(0.017)	(0.015)	(0.020)	(0.011)	(0.083)	(0.105)	(0.054)	(0.059)	(0.055)	(0.038)
GARMA	0.151	0.120	0.081	0.077	0.056	0.051	-	-	-	-	-	-
-ZAGA	(0.062)	(0.022)	(0.021)	(0.023)	(0.016)	(0.013)	-	-	-	-	-	-
GARMA	-	-	-	-	-	-	0.245	0.334	0.185	0.197	0.106	0.133
-Binomial	-	-	-	-	-	-	(0.101)	(0.109)	(0.056)	(0.069)	(0.063)	(0.045)
SCZI	0.154	0.127	0.085	0.079	0.058	0.052	0.248	0.339	0.188	0.204	0.108	0.138
	(0.056)	(0.031)	(0.023)	(0.020)	(0.017)	(0.015)	(0.086)	(0.105)	(0.069)	(0.065)	(0.057)	(0.045)

fied". That is, the simulated data is generated based on the structure of the proposed model, and it is different from GARMA models. We took this approach because the underlying structure of temporal correlation should be defined when generating auto-correlated data. However, this problem may not be prominent because the simulated latent states are not considered as part of the simulated data, and we only used the structure to generate auto-correlated data. Therefore, only the covariates and auto-correlated responses are included in the simulated data, and we do not consider latent state estimation in model comparison. Nevertheless, it is worth exploring other approaches to generate simulated data and compare models with different underlying structures. For example, we may generate data based on the GARMA framework, and compare the performance when the proposed model is misspecified. Another approach to compare models with different structure is to use real auto-regressive datasets. However, these datasets usually do not include true values of parameters as benchmark to compare against. In this case, other model comparison methods are required. For example, we may use cross-validation and compare predicted values with true responses. This requires future works to develop prediction methods under the current framework. Finally, besides misspecification, it is interesting to explore other conditions when the proposed model does not perform well with additional experiments. For example, it is interesting to evaluate the performance of DSCZI when the underlying latent states transition is not linear.

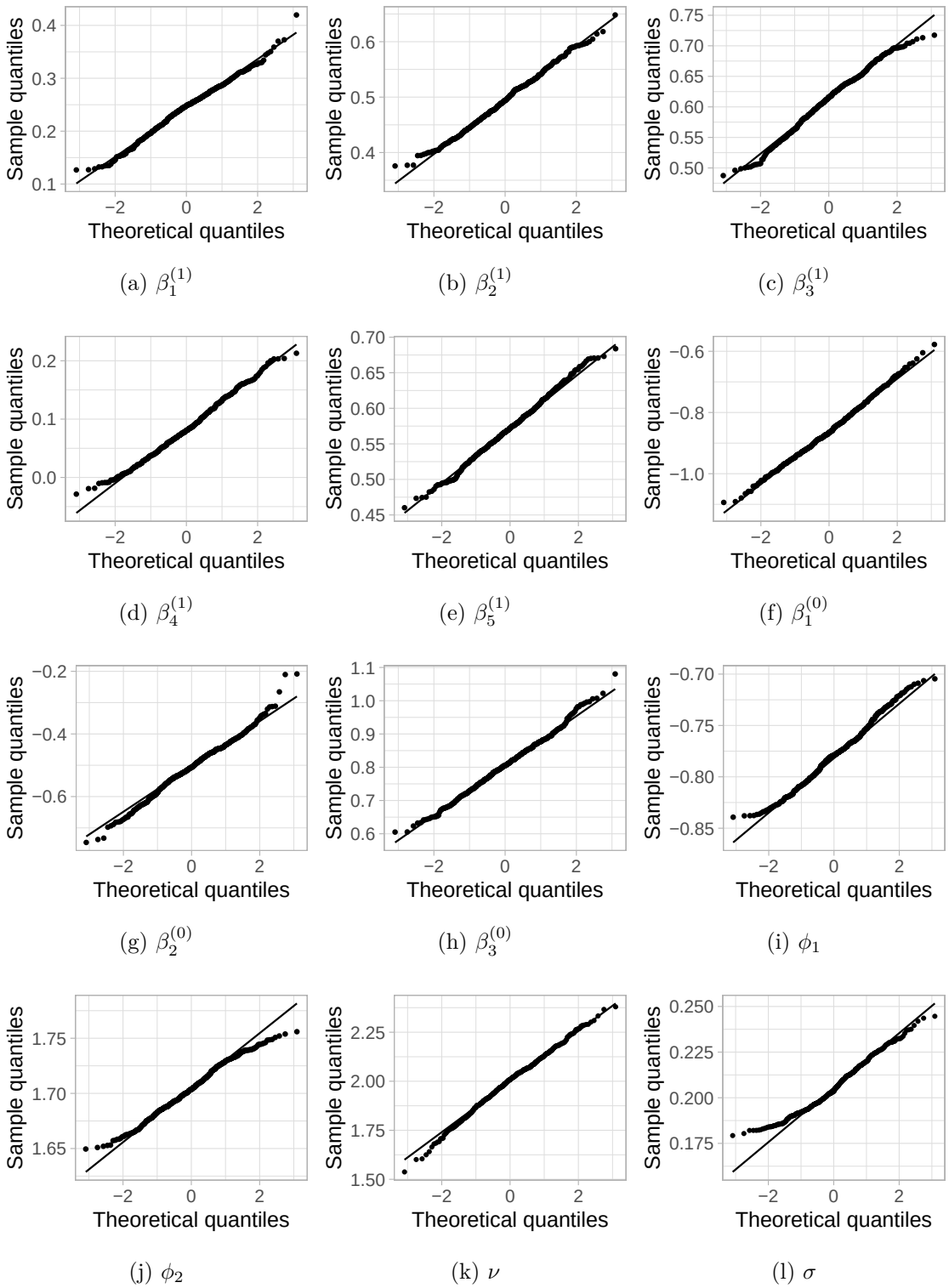
4.5 Analysis of adaptive cruise control data

Previous studies have used drivers' braking behavior as an indicator of drivers' response to emergency situations (Banks et al., 2014; Duan et al., 2017). Therefore, it is of interest to analyze differences in drivers' behaviors related to braking in different situations. In this section, we apply the proposed dynamic semi-continuous zero-inflated model (DSCZI) to examine the association between the closest in-path vehicle and drivers' braking behavior in different situations.

The dataset was originally collected by the University of Michigan Transportation Re-

Table 4.2: Summary of statistics for the estimators of the fitted model with $T = 500$, $m^{(0)} = 3$ and $m^{(1)} = 5$.

	Ture value	Mean	Median	SE
$\beta_1^{(1)}$	0.243	0.231	0.231	0.00394
$\beta_2^{(1)}$	0.417	0.510	0.510	0.00425
$\beta_3^{(1)}$	0.572	0.592	0.591	0.00389
$\beta_4^{(1)}$	0.201	0.108	0.108	0.00429
$\beta_5^{(1)}$	0.555	0.598	0.598	0.00407
$\beta_1^{(0)}$	-0.753	-0.860	-0.860	0.00724
$\beta_2^{(0)}$	-0.648	-0.513	-0.513	0.00707
$\beta_3^{(0)}$	0.338	0.813	0.814	0.00720
ν	2.000	2.184	2.167	0.01027
ϕ_1	-0.900	-0.834	-0.841	0.00367
ϕ_2	1.752	1.686	1.698	0.00434
σ	0.100	0.164	0.156	0.00291



(a) $\beta_1^{(1)}$ (b) $\beta_2^{(1)}$ (c) $\beta_3^{(1)}$

(d) $\beta_4^{(1)}$ (e) $\beta_5^{(1)}$ (f) $\beta_1^{(0)}$

(g) $\beta_2^{(0)}$ (h) $\beta_3^{(0)}$ (i) ϕ_1

(j) ϕ_2 (k) ν (l) σ

Figure 4.1: Q-Q plots for the estimated parameters with $T = 500$, $m^{(0)} = 3$ and $m^{(1)} = 5$.

search Institute (UMTRI) as part of a field operational test on an advanced collision avoidance system (Ervin et al., 2005; Xiong and Boyle, 2012). It employed 10 vehicles using a total of 66 drivers (Najm et al., 2006). There have been many studies conducted using this dataset, all with different analytical methods. For example, logistic regression was used to compare distracted normal driving (Green et al., 2007) and to examine drivers' adaption to adaptive cruise control (Xiong and Boyle, 2012). Other studies focused on the drivers' behavior using linear mixed model and decision tree (LeBlanc et al., 2013; Rosenfeld et al., 2015).

In this dissertation, we applied the proposed method to study the drivers' brake response to the closest in-path vehicle (CIPV). Our analysis focused on two objectives. First, we wanted to investigate the existence of auto-correlation in drivers' braking behavior. Second, we wished to assess and compare the influence of CIPV on drivers' braking behavior in different driving situations. The response variable was defined as follows: it was set to zero if the brake pedal was not engaged; if the driver braked, then the value of the response variable was set to the declaration of the vehicle. We used this variable to reflect the likelihood and magnitude of braking, which was not included in the original dataset.

This proposed method was examined for different road and traffic scenarios that were based on 5-minute segments from three different trips that had nonzero responses more than 0.5% of all the observations. These three trips were from the same participant in the experiment. The selected segments represented three different driving situations: local roads, highway with traffic, and highway without traffic. These driving situations were determined by the types of roads, i.e., local roads and highway, and the level of service that categorizes traffic flow.

The data was recorded at 10Hz, and observations were aggregated up to a 0.2-second window. The percentages of nonzero responses ranged from 0.8% to 6.2%. A detailed summary of the segments and their corresponding percentages of nonzero responses can be found in Table 4.3. Six variables and one optional variable were included in the analysis (Table 4.4). Note that the "No CIPV" variable is optional because in three trip segments, a

CIPV always exists. We set the order of the latent auto-regression to $p = 1, 2$ and used the deviance information criterion (DIC) (Spiegelhalter et al., 2014) to select the optimal model.

Table 4.3: Summary of trip segments and their corresponding percentages of nonzero responses.

	Local	Highway w/o traffic	Highway w/ traffic
Trip 1	4.13%	0.80%	-
Trip 33	4.27%	-	-
Trip 93	-	1.80%	6.20% (Segment 1) 4.73% (Segment 2)

Table 4.4: Description of variables in the adaptive cruise control data.

Variable name	Data type	Unit	Description
Distance to CIPV	Continuous	m	Distance to the closest in-path vehicle (CIPV)
Speed of CIPV	Continuous	m/s	Transmission speed of CIPV
CIPV Acceleration	Continuous	m/s^2	Acceleration of CIPV
CIPV Deceleration	Continuous	m/s^2	Deceleration of CIPV
Time to Crash into CIPV	Continuous	s	Time to crash into CIPV if the speeds of the vehicles do not change.
No CIPV	Binary, optional		1 – no CIPV; 0 – CIPV exists. It is not included in the model if CIPV always exists.
Lane Offset	Continuous	m	Offset from enter of lane

Table 4.5: Results for parameter estimates of the five trip segments based on the proposed DSCZI method. The numbers in parentheses are p-values.

Variables	Local				Highway w/o traffic				Highway w/ traffic			
	Trip 1		Trip 33		Trip 1		Trip 93		Trip 93 - 1		Trip 93 - 2	
	Zero	Nonzero	Zero	Nonzero	Zero	Nonzero	Zero	Nonzero	Zero	Nonzero	Zero	Nonzero
Intercept	**-3.605**	**0.705**	**-6.469**	0.236	**-19.784**	-0.497	**39.811**	**3.135**	**-21.601**	4.127	**14.421**	-0.191
	(0.005)	**(0.029)**	**(3.9E-06)**	(0.391)	**(0.007)**	(0.500)	**(0.017)**	**(0.040)**	**(0.008)**	(0.070)	**(0.041)**	(0.168)
Distance to CIPV	-0.008	0.002	**0.084**	0.000	**1.027**	0.034	**1.563**	0.117	0.159	0.014	**-0.562**	0.006
	(0.804)	(0.782)	**(0.008)**	(0.877)	**(0.030)**	(0.492)	**(0.025)**	(0.095)	(0.623)	(0.493)	**(0.004)**	(0.422)
Speed of CIPV	**-0.117**	**-0.040**	**-0.207**	-0.006	**-0.536**	0.042	**-1.942**	**-0.145**	-0.408	-0.032	**1.946**	-0.014
	(0.002)	**(0.002)**	**(0.002)**	(0.277)	**(0.031)**	(0.092)	**(0.011)**	**(0.047)**	(0.391)	(0.221)	**(1.5E-09)**	(0.153)
CIPV Acceleration	-1.311	**-1.343**	-1.506	**-1.157**	-8.883	1.018	-7.265	-3.710	0.044	**-0.329**	0.196	-20.096
	(0.821)	**(0.002)**	(0.114)	**(0.036)**	(0.204)	(0.096)	(0.326)	(0.132)	(0.968)	**(0.002)**	(0.777)	(0.473)
CIPV Deceleration	**5.150**	**0.894**	**6.392**	**0.519**	**2.619**	**0.637**	**6.419**	**0.719**	**8.740**	**0.523**	**8.102**	**0.798**
	(< 2E-16)	**(< 2E-16)**	**(< 2E-16)**	**(5.3E-10)**	**(5.9E-05)**	**(3.6E-09)**	**(2.0E-08)**	**(1.1E-04)**	**(< 2E-16)**	**(2.7E-15)**	**(< 2E-16)**	**(< 2E-16)**
Time to Crash into CIPV	-0.161	-0.146	**-0.464**	**-0.050**	**27.029**	-0.897	**-42.695**	-2.789	**2.265**	-0.075	**1.395**	**-0.047**
	(0.760)	(0.455)	**(0.013)**	**(0.024)**	**(0.031)**	(0.520)	**(0.014)**	(0.064)	**(0.001)**	(0.217)	**(4.4E-04)**	**(0.049)**
No CIPV	-	-	0.237	0.218	-	-	-	-	**11.067**	**1.123**	**12.063**	**0.564**
	-	-	(0.867)	(0.107)	-	-	-	-	**(0.009)**	**(0.014)**	**(9.7E-06)**	**(0.004)**
Lane Offset	1.025	-0.062	**1.226**	-0.003	-2.384	**-0.373**	-0.634	0.133	2.712	0.667	-12.343	**0.838**
	(0.188)	(0.381)	**(0.020)**	(0.978)	(0.238)	**(0.047)**	(0.669)	(0.283)	(0.754)	(0.309)	(0.265)	**(0.016)**
ϕ_1	0.069		**0.483**		**0.176**		**0.689**		**0.442**		**0.759**	
	(0.102)		**(5.51E-05)**		**(8.04E-05)**		**(4.07E-07)**		**(9.06E-05)**		**(<2E-16)**	
ϕ_2	**0.066**		**-0.070**		-	-	-	-	**0.329**			
	(0.031)		**(0.038)**		-	-	-	-	**(0.003)**		-	-

The results are shown in Table 4.5. We choose $\alpha = 0.05$ and calculate asymptotic p-values following the discussion in Section 4.4. Significant parameters are highlighted in bold. We observe that for all the segments, there exists auto-correlation between the responses. The deceleration of CIPV is the most significant factor affecting both the likelihood and magnitude of braking regardless of the driving situation. This implied the driver was more likely to engage brake if CIPV decelerated, and tended to apply more force on the brake pedal when CIPV decelerated faster.

The results suggested that the influence of CIPV shared similarity in the same driving situation and differed in different driving situations. For example, on local roads, the likelihood of engaging the brake depended on the speed and deceleration of CIPV. The magnitude of braking was related to the acceleration and declaration of CIPV. In comparison, on highway without traffic, the likelihood was also related to the distance to CIPV, and the magnitude of brake pedal engagement was not related to the acceleration of CIPV. On highway with traffic, the existence of CIPV was strongly associated with the likelihood of engaging the brake pedal. The declaration of and the time to crash into CIPV also impacted the likelihood of engaging the brake pedal. The force applied on the brake pedal was highly related to the declaration of CIPV, which was similar to the highway without traffic situation.

In the same driving situation, the influence of CIPV on the likelihood and magnitude of braking could be different. For example, on local roads, for both Trip 1 and 33, the acceleration of CIPV was more related to the magnitude of braking. In the highway without traffic situation, for both Trip 1 and 93, the distance to CIPV and time of crashing into CIPV were more related to the likelihood of engaging the brake pedal. And in the highway with traffic situation, for Trip 93, the time to crash into CIPV mainly affected the likelihood of engaging the brake traffic.

Finally, inconsistent influence of CIPV in the same driving situation existed. For instance, on local roads, the time to crash into CIPV was associated with the likelihood of engaging the brake pedal only in Trip 33. The time to crash into CIPV was positively related to the likelihood of braking in Trip 1, but was negatively to it in Trip 93. The speed of CIPV

impacted the magnitude of braking only in the first highway with traffic segment of Trip 93. Overall, the influence of CIPV was more consistent in the simple driving situation (highway without traffic) than the complex driving situation (local roads and highway with traffic). This inconsistency is most likely due to different driving styles in different trips and missing variables describing the complex traffic conditions.

4.6 Summary

In this chapter, we designed a novel dynamic semi-continuous zero-inflated model (DSCZI) for data with zero-inflated time-series responses. This model considers the auto-correlation between responses by introducing a latent auto-regressive process. The model estimation of DSCZI utilizes a novel Monte Carlo Markov chain tool, the data cloning. The use of data cloning guarantees the convergence of generated random variates to the MLE of parameters and provides tools for statistical inference under certain conditions. Extensive numerical tests showed that the proposed model performed better than the misspecified static zero-inflated model and models treating the zero and nonzero parts separately. We applied the proposed model to analyze drivers' braking behavior in different situations. The metric of braking is auto-correlated and zero-inflated as data was collected in a series of consecutive time points. The measure is equal zero if the driver did not brake and greater than zero if the driver braked.

The model could capture the similarity and difference of driver's braking behavior between driving situations when interacting with the closest in-path vehicle (CIPV). Specifically, the experiment showed that the conditions of CIPV impacted both the likelihood and magnitude of braking for each driving situation. For all the driving situations, braking happened when the CIPV decelerated. For highway with traffic situation, the driver was more likely to brake when CIPV presented. In comparison, the driver was less likely to brake if the speed of CIPV was higher in the highway without traffic situation, and on local roads, the driver usually applied less force on the brake pedal if the CIPV was accelerating faster.

We also noticed that the influence of CIPV on driver's braking behavior can be different in

the same driving situation. For example, on the local roads, the driver payed more attention to the distance to CIPV and the time to crash into CIPV in Trip 33 than in Trip 1. On highways without traffic, the driver adjusted the force applied to the brake pedal based on the speed of CIPV in Trip 93. But in Trip 1, the force on the brake pedal was almost only influenced by the deceleration of CIPV.

Therefore, in practice, collision avoidance system should be tailored for different driving situations. While certain variables, such as the deceleration of CIPV, impacted drivers' braking behavior in all the driving situations, other variables, e.g., acceleration of CIPV, played different roles in different driving situations. For different driving situations, variables that trigger the brake assistance and determines the brake force should also be different based on the driving situation. Finally, in complex driving situations, such as local roads and highways with heavy traffics, the systems should be able to adjust themselves adapting to drivers' recent behavior and traffic conditions.

Other models exist in the literature that focus on modeling time series data. For example, it is known that neural networks are also capable of modeling time-series data with, e.g., long short term memory (LSTM) (Hochreiter and Schmidhuber, 1997), and achieve good results in terms of prediction (Zheng et al., 2017; Rangapuram et al., 2018). However, latent state based probabilistic models, such as the model proposed in this chapter, still have their advantages. For example, Panzner and Cimiano (2016) showed that hidden Markov model is better than LSTM with smaller sample size. Furthermore, applications, such as driver's braking behavior analysis, focus on understanding the relationship between covariates and responses. In these cases, the interpretability is important. Motivated by the application of driver's braking behavior analysis, the proposed model is based on the parameter-driven framework aiming to understand how drivers respond to nearby vehicles. On the other hand, LSTM models are block boxes, and estimated models are hard to interpret.

The current model can be further improved. For example, moving average may also need to be included to model the latent state transitions. As shown in the experiments, the influence of some variables on driver's braking behavior shares similarity. Therefore, it can be

interesting to extend the current model to penal data. One possible approach is to introduce individual-specific effects and time-varying random effects. Nevertheless, current model is not computationally efficient. Techniques, such as parallel computing, to accelerate the model estimation is important and worthy of exploring. While it is possible to assume the latent state transits following a nonlinear AR model (Jones and Cox, 1978), model estimation can be more challenging and time-consuming. For example, Hwang and Shin (2011) used bootstrap to estimate stationary nonlinear auto-regressive model. Moreover, because latent states do not correspond to aspects of physical reality, linear transitions can be enough to capture the temporal correlation between responses. Further study and experiments are necessary to determine if using nonlinear auto-regressive latent states can improve the parameter estimation performance. Finally, it is critical to develop model selection methods that can help decide the optimal number of clones to be used.

In the real data analysis, the dataset is highly imbalanced, and the small number of nonzero responses in some segments may affect the parameter estimation. Thus, additional data collection can be necessary for further data analysis. Finally, it is worth comparing the Monte Carlo Expectation-Maximization method proposed in Yang et al. (2015) to the data cloning method used in the proposed model in terms of parameter estimation accuracy and computational efficiency.

Chapter 5

CONCLUSIONS

In this thesis, we reviewed the established zero-inflated models and proposed two new models for semi-continuous zero-inflated (SCZI) data. In Chapter 3, we constructed a novel adaptive sparse group lasso regularization for semi-continuous zero-inflated models (ASGL). This model assumed the coefficients from the zero and nonzero parts of the SCZI model for the same variable should not be independent. A revised BIC was derived for the proposed model and showed the asymptotic property of the model under a set of conditions. Extensive numerical tests showed that the proposed regularization performed better than original lasso regularization in terms of variable selection.

In Chapter 4, we designed a novel dynamic semi-continuous zero-inflated model (DSCZI) for time series of semi-continuous zero-inflated data. This model takes into account the semi-continuous zero-inflation of responses and the auto-correlation between responses by introducing a latent auto-regressive process to the static SCZI model. An advanced Monte Carlo Markov chain tool, the data cloning, was used to estimate the model. Compared with other estimtion methods, data cloning guaranteed the convergence of generated random variates to the MLE of parameters and provided tools for statistical inference. Extensive numerical tests showed that the proposed model performed better than the static zero-inflated model and treating the zero and nonzero parts separately.

The proposed models were applied to transportation datasets from a driving simulator study and a field operation study. We applied the ASGL model to data from a driving simulator study that included lane departures. The metric of lane departure is zero if a subject stays inside the lane (which was the majority of time) and greater than zero if a subject drifted outside of the lane. Our proposed method could be used to examine the

relationship between driving performance measures and lane departures. The model captured individual differences in driving. For example, for some drivers, lane departure occurred when the steering wheel angle changed, while for other drivers, lane departure occurred due to a failure to correct the vehicle heading. While there was a correlation between steering wheel angle and vehicle heading, the differences are based on whether the driver initiated the move (changing the steering wheel) or reacting to a shift due to external conditions (changing as a result of the vehicle heading).

We applied the DSCZI model to analyze drivers' braking behavior in different situations. The metric of braking is auto-correlated as data was collected in a series of consecutive time points. It is zero-inflated because the measure was equal zero if the driver did not brake, and greater than zero if the driver braked.

The model was able to capture the similar and different influences of the closest in-path vehicle (CIPV) on driver's braking behavior in different driving situations. More specifically, the experiment showed that the conditions of CIPV impacted both the likelihood and magnitude of braking for each driving situation. For all the driving situations, braking happened when the CIPV decelerated. For highway with traffic situation, the driver was more likely to brake if there is CIPV. In comparison, the driver was less likely to brake if the speed of CIPV was higher in the highway without traffic situation, and on local roads, the driver usually applied less force on the brake pedal if the CIPV was accelerating faster.

We also noticed that the influence of CIPV on driver's braking behavior can be different in the same driving situation. For example, on the local roads, the driver payed more attention to the distance to CIPV and the time to crash into CIPV in Trip 33 than in Trip 1. On highways without traffic, the driver adjusted the force applied to the brake pedal based on the speed of CIPV in Trip 93. But in Trip 1, the force on the brake pedal was almost only influenced by the deceleration of CIPV.

Based on the experiment results, the proposed methods can help develop in-vehicle driver-assistance systems. For example, car manufacturers can use the proposed ASGL model to develop algorithms as part of the in-vehicle warning system. Using this zero-inflated model,

the system will be able to identify critical variables and learn the operator's driving styles continuously based on these variables. The learned custom model will be able to detect when drivers depart from the lane, and appropriately warn the driver in advance.

The DSCZI method can assist tailoring the automatic emergency braking system for different driving situations. Variables exist that impacted drivers' braking behavior in all the driving situations, such as the deceleration of CIPV. On the other hand, other variables, e.g., acceleration of CIPV, played different roles in different driving situations. Therefore, in different driving situations, variables that trigger the braking and determine the brake force can be different and should be carefully selected based on the driving situations. Finally, the results suggested that in complex driving situations, such as local roads and highways with heavy traffics, drivers' braking behavior and traffic situations vary from time to time, and the systems should be able to adapt to drivers' recent behavior and traffic conditions.

The proposed models can be further improved. For ASGL, the current model uses LSA approximation method to estimate coefficients, and the performance of LSA highly depends on the unpenalized estimate of the covariance matrix of coefficients. This estimate can be inaccurate. For example, the unpenalized estimate of the covariance matrix of coefficients was obtained with the assumption that coefficients from zero and nonzero parts were independent based on the log-likelihood function. However, this assumption can be violated if the two parts are highly related to each other. One possible alternative approach to solving this problem is to estimate the covariance matrix with Bayesian methods. Nevertheless, in real applications, we face the challenge of capturing the random effect that accounts for the systematic differences between subjects. It is worthwhile to explore methods to include random effects in the current model while also retaining the asymptotic properties.

For DSCZI, moving average may also need to be included to model the latent state transitions. As shown in the experiments, the influence of some variables on driver's braking behavior shares similarity. Therefore, it can be interesting to extend the current model to penal data. One possible approach is to introduce individual-specific effects and time-varying random effects. In the real data analysis, the dataset is highly imbalanced, and the

small number of non-zero responses in some segments may affect the parameter estimation. Thus, additional data collection can be necessary for further data analysis. Finally, it is worth comparing the Monte Carlo Expectation-Maximization method proposed in Yang et al. (2015) to the data cloning method used in the proposed model in terms of parameter estimation accuracy and computational efficiency.

In terms of application, it can be interesting to explore applications of the proposed two models in other areas besides transportation. For example, as discussed in Section 1.2, zero-inflated models have been widely used in actuarial science and healthcare. For the car insurance claim data, the ASGL model can be used to identify clients' covariates related to claimed losses. If an intervention, such as change of transportation law, happens, the DSCZI can be used to evaluate the effect of the intervention by comparing the claimed loss and related covariates before and after the intervention.

In summary, this thesis aims to separate the discrete distributions from the continuous ones, design a novel variable selection method, and extend the current zero-inflated models to time-series data and transportation applications. More studies are required for a more comprehensive and flexible model that can assess and predict all drivers' actions. While this thesis used data on lane departures and braking behavior as examples, the proposed models can also be considered for other areas facing semi-continuous zero-inflation problems, such as accrual science, safety and health.

BIBLIOGRAPHY

Aarts, L. and Van Schagen, I. "Driving speed and the risk of road crashes: A review." *Accident Analysis & Prevention*, 38(2):215–224 (2006).

Abu-Eisheh, S. and Mannering, F. L. "Discrete/continuous analysis of commuters' route and departure time choices." *Transportation Research Record*, 1138:27–34 (1987).

Administration, N. H. T. S. et al. "Early estimate of motor vehicle traffic fatalities for the first 9 months of 2019." *DOT HS*, 812:874 (2019).

Al-Wahsh, H. and Hussein, A. "Estimation of zero-inflated parameter-driven models via data cloning." *Journal of Statistical Computation and Simulation*, 89(6):951–965 (2019).

Albousefi, A. A., Ying, H., Filev, D., Syed, F., Prakah-Asante, K. O., Tseng, F., and Yang, H.-H. "A support vector machine approach to unintentional vehicle lane departure prediction." In *2014 IEEE Intelligent Vehicles Symposium Proceedings*, 299–303. IEEE (2014).

Alghamdi, A. S. and Harrington Jr, J. "Time-sensitive analysis of a warming climate on heat waves in Saudi Arabia: Temporal patterns and trends." *International Journal of Climatology*, 38(7):3123–3139 (2018).

Alqawba, M., Diawara, N., and Rao Chaganty, N. "Zero-inflated count time series models using Gaussian copula." *Sequential Analysis*, 38(3):342–357 (2019).

Amemiya, T. "Tobit models: A survey." *Journal of econometrics*, 24(1-2):3–61 (1984).

Assuncao, A. N., Aquino, A. L., Santos, C. d. M., Ricardo, C., Guimaraes, R. L., and Oliveira, R. A. "Vehicle Driver Monitoring through the Statistical Process Control." *Sensors*, 19(14):3059 (2019).

Bagdadi, O. "Assessing safety critical braking events in naturalistic driving studies." *Transportation research part F: traffic psychology and behaviour*, 16:117–126 (2013).

Banks, V. A., Stanton, N. A., and Harvey, C. "What the drivers do and do not tell you: using verbal protocol analysis to investigate driver behaviour in emergency situations." *Ergonomics*, 57(3):332–342 (2014).

Bengler, K., Kohlmann, M., and Lange, C. "Assessment of cognitive workload of in-vehicle systems using a visual peripheral and tactile detection task setting." *Work*, 41(Supplement 1):4919–4923 (2012).

Benjamin, M. A., Rigby, R. A., and Stasinopoulos, D. M. "Generalized autoregressive moving average models." *Journal of the American Statistical association*, 98(461):214–223 (2003).

Bollerslev, T. "Generalized autoregressive conditional heteroskedasticity." *Journal of econometrics*, 31(3):307–327 (1986).

Borisov, A., Runger, G., Tuv, E., and Lurponglukana-Strand, N. "Zero-inflated boosted ensembles for rare event counts." In *International Symposium on Intelligent Data Analysis*, 225–236. Springer (2009).

Brijs, T., Karlis, D., and Wets, G. "Studying the effect of weather conditions on daily crash counts using a discrete time-series model." *Accident Analysis & Prevention*, 40(3):1180–1190 (2008).

Bulmash, E. L., Moller, H. J., Kayumov, L., Shen, J., Wang, X., and Shapiro, C. M. "Psychomotor disturbance in depression: assessment using a driving simulator paradigm." *Journal of affective disorders*, 93(1-3):213–218 (2006).

Carson, J. and Mannering, F. "The effect of ice warning signs on ice-accident frequencies and severities." *Accident Analysis & Prevention*, 33(1):99–109 (2001).

Chai, H. S. and Bailey, K. R. "Use of log-skew-normal distribution in analysis of continuous data with a discrete component at zero." *Statistics in medicine*, 27(18):3643–3655 (2008).

Chan, K. and Ledolter, J. "Monte Carlo EM estimation for time series models involving counts." *Journal of the American Statistical Association*, 90(429):242–252 (1995).

Chang, C.-C. "Assessing Cognitive Workload of In-Vehicle Voice Control Systems." Ph.D. thesis, University of Washington (2016).

Chatterjee, S., Chowdhury, S., Mallick, H., Banerjee, P., and Garai, B. "Group regularization for zero-inflated negative binomial regression models with an application to health care demand in Germany." *Statistics in medicine*, 37(20):3012–3026 (2018).

Chatterjee, S., Steinhaeuser, K., Banerjee, A., Chatterjee, S., and Ganguly, A. "Sparse group lasso: Consistency and climate applications." In *Proceedings of the 2012 SIAM International Conference on Data Mining*, 47–58. SIAM (2012).

Chowdhury, S., Chatterjee, S., Mallick, H., Banerjee, P., and Garai, B. "Group regularization for zero-inflated poisson regression models with an application to insurance ratemaking." *Journal of Applied Statistics*, 1–15 (2018).

Cirillo, C., Liu, Y., and Tremblay, J.-M. "Simulation, numerical approximation and closed forms for joint discrete continuous models with an application to household vehicle ownership and use." *Transportation*, 44(5):1105–1125 (2017).

Cox, D. R., Gudmundsson, G., Lindgren, G., Bondesson, L., Harsaae, E., Laake, P., Juselius, K., and Lauritzen, S. L. "Statistical analysis of time series: Some recent developments [with discussion and reply]." *Scandinavian Journal of Statistics*, 93–115 (1981).

Crum, M. R., Morrow, P. C., Olsgard, P., and Roke, P. J. "Truck driving environments and their influence on driver fatigue and crash rates." *Transportation research record*, 1779(1):125–133 (2001).

da Silva, C. Q., da Silva, P. H., Turnes, O., and Correia, L. T. "Dynamic model averaging adapted to dynamic regression models for time series of counts." *Communications in Statistics-Simulation and Computation*, 1–24 (2019).

Davis, R. A. and Wu, R. "A negative binomial model for time series of counts." *Biometrika*, 96(3):735–749 (2009).

Dempster, A. P., Laird, N. M., and Rubin, D. B. "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22 (1977).

DeSantis, S. M. and Bandyopadhyay, D. "Hidden Markov models for zero-inflated Poisson counts with an application to substance use." *Statistics in medicine*, 30(14):1678–1694 (2011).

Dong, C., Richards, S. H., Clarke, D. B., Zhou, X., and Ma, Z. "Examining signalized intersection crash frequency using multivariate zero-inflated Poisson regression." *Safety science*, 70:63–69 (2014).

Donmez, B., Boyle, L. N., and Lee, J. D. "SAfety VEhicles using adaptive Interface Technology (Task 4) Final Report: Phase 2 Distraction Mitigation Evaluation." *University of Iowa* (2007).

Duan, J., Li, R., Hou, L., Wang, W., Li, G., Li, S. E., Cheng, B., and Gao, H. "Driver braking behavior analysis to improve autonomous emergency braking systems in typical Chinese vehicle-bicycle conflicts." *Accident Analysis & Prevention*, 108:74–82 (2017).

Duan, N., Manning, W. G., Morris, C. N., and Newhouse, J. P. "A comparison of alternative models for the demand for medical care." *Journal of business & economic statistics*, 1(2):115–126 (1983).

—. "Choosing between the sample-selection model and the multi-part model." *Journal of Business & Economic Statistics*, 2(3):283–289 (1984).

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. "Least angle regression." *The Annals of statistics*, 32(2):407–499 (2004).

Ervin, R., Sayer, J., LeBlanc, D., Bogard, S., Mefford, M., Hagan, M., Bareket, Z., and Winkler, C. "Automotive collision avoidance system field operational test report: methodology and results." Technical report, University of MichiganTransportation Research Institute (2005).

Fan, J. and Li, R. "Variable selection via nonconcave penalized likelihood and its oracle properties." *Journal of the American statistical Association*, 96(456):1348–1360 (2001).

Fang, H. A. "A discrete–continuous model of households' vehicle choice and usage, with an application to the effects of residential density." *Transportation Research Part B: Methodological*, 42(9):736–758 (2008).

Farewell, V., Long, D., Tom, B., Yiu, S., and Su, L. "Two-part and related regression models for longitudinal data." *Annual review of statistics and its application*, 4:283–315 (2017).

Fokianos, K., Rahbek, A., and Tjøstheim, D. "Poisson autoregression." *Journal of the American Statistical Association*, 104(488):1430–1439 (2009).

Fowler, G. F., Larson, R. E., and Wojcik, L. A. "Driver crash avoidance behavior: Analysis of experimental data collected in NHTSA's vehicle antilock brake system (ABS) research program." Technical report, SAE Technical Paper (2005).

Freeland, R. and McCabe, B. P. "Analysis of low count time series data by Poisson autoregression." *Journal of Time Series Analysis*, 25(5):701–722 (2004).

Friedman, J., Hastie, T., and Tibshirani, R. "A note on the group lasso and a sparse group lasso." *arXiv preprint arXiv:1001.0736* (2010a).

—. "Regularization paths for generalized linear models via coordinate descent." *Journal of statistical software*, 33(1):1 (2010b).

Ghosh, P. and Albert, P. S. "A Bayesian analysis for longitudinal semicontinuous data with an application to an acupuncture clinical trial." *Computational statistics & data analysis*, 53(3):699–706 (2009).

Godsill, S. J., Doucet, A., and West, M. "Monte Carlo smoothing for nonlinear time series." *Journal of the american statistical association*, 99(465):156–168 (2004).

Gonçalves, E., Mendes-Lopes, N., and Silva, F. "Zero-inflated compound Poisson distributions in integer-valued GARCH models." *Statistics*, 50(3):558–578 (2016).

Gordon, N. J., Salmond, D. J., and Smith, A. F. "Novel approach to nonlinear/non-Gaussian Bayesian state estimation." In *IEE proceedings F (radar and signal processing)*, volume 140, 107–113. IET (1993).

Green, P., Cullinane, B., Zylstra, B., and Smith, D. "Typical values for driving performance with emphasis on the standard deviation of lane position: A summary of the literature." *University of Michigan, Transportation Research Institute (UMTRI)* (2003).

Green, P. E., Wada, T., Oberholtzer, J., Green, P. A., Schweitzer, J., and Eoh, H. "How do distracted and normal driving differ: An analysis of the ACAS naturalistic driving data." Technical report, University of Michigan Transportation Research Institute (2007).

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

Habib, K. M. N. "A joint discrete-continuous model considering budget constraint for the continuous part: application in joint mode and departure time choice modelling." *Transportmetrica A: Transport Science*, 9(2):149–177 (2013).

Habib, K. M. N., Day, N., and Miller, E. J. "An investigation of commuting trip timing and mode choice in the Greater Toronto Area: Application of a joint discrete-continuous model." *Transportation Research Part A: Policy and Practice*, 43(7):639–653 (2009).

Han, D., Liu, L., Su, X., Johnson, B., and Sun, L. "Variable selection for random effects two-part models." *Statistical methods in medical research*, 28(9):2697–2709 (2019).

Harbluk, J. L., Noy, Y. I., Trbovich, P. L., and Eizenman, M. "An on-road assessment of cognitive distraction: Impacts on drivers' visual behavior and braking performance." *Accident Analysis & Prevention*, 39(2):372–379 (2007).

Harvey, A. and Ito, R. "Modeling time series when some observations are zero." *Journal of Econometrics*, 214(1):33–45 (2020).

Hasan, M. T., Huda, S., and Sneddon, G. "A Comparative Study of Observation-and Parameter-driven Zero-inflated Poisson Models for Longitudinal Count Data." *Communications in Statistics-Simulation and Computation*, 45(10):3643–3659 (2016).

Hautsch, N., Malec, P., and Schienle, M. "Capturing the zero: a new class of zero-augmented distributions and multiplicative error processes." *Journal of Financial Econometrics*, 12(1):89–121 (2014).

Helske, J. "KFAS: Exponential Family State Space Models in R." *Journal of Statistical Software*, 78(10) (2017).

Hochreiter, S. and Schmidhuber, J. "Long short-term memory." *Neural computation*, 9(8):1735–1780 (1997).

Hu, S.-R., Li, C.-S., and Lee, C.-K. "Assessing casualty risk of railroad-grade crossing crashes using zero-inflated poisson models." *Journal of transportation engineering*, 137(8):527–536 (2010).

Hwang, E. and Shin, D. W. "Stationary bootstrapping for non-parametric estimator of nonlinear autoregressive model." *Journal of Time Series Analysis*, 32(3):292–303 (2011).

Hyndman, R. J. and Grunwald, G. K. "Applications: Generalized Additive Modelling of

Mixed Distribution Markov Models with Application to Melbourne's Rainfall." *Australian & New Zealand Journal of Statistics*, 42(2):145–158 (2000).

Jang, H., Lee, S., and Kim, S. W. "Bayesian analysis for zero-inflated regression models with the power prior: Applications to road safety countermeasures." *Accident Analysis & Prevention*, 42(2):540–547 (2010).

Jazi, M. A., Jones, G., and Lai, C.-D. "First-order integer valued AR processes with zero inflated Poisson innovations." *Journal of Time Series Analysis*, 33(6):954–963 (2012).

Johansson, P. "Speed limitation and motorway casualties: a time series count data regression approach." *Accident Analysis & Prevention*, 28(1):73–87 (1996).

Jones, D. A. and Cox, D. R. "Nonlinear autoregressive processes." *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 360(1700):71–95 (1978).

Jørgensen, B. "Exponential dispersion models." *Journal of the Royal Statistical Society: Series B (Methodological)*, 49(2):127–145 (1987).

Kawakatsu, H. "Jointly Modeling Autoregressive Conditional Mean and Variance of Non-Negative Valued Time Series." *Econometrics*, 7(4):48 (2019).

Kilpeläinen, M. and Summala, H. "Effects of weather and weather forecasts on driver behaviour." *Transportation research part F: traffic psychology and behaviour*, 10(4):288–299 (2007).

Kim, J. and Stoffer, D. S. "Fitting stochastic volatility models in the presence of irregular sampling via particle methods and the EM algorithm." *Journal of time series analysis*, 29(5):811–833 (2008).

Kogure, A. "Predicting Health Care Costs by Two-part Model with Sparse Regularization." In *The World Risk and Insurance Economics Congress* (2015).

Koh, Y., Bukhari, N., and Mohamed, I. "Parameter-driven state-space model for integer-valued time series with application." *Journal of Statistical Computation and Simulation*, 89(8):1394–1409 (2019).

Koopman, S. J., Lucas, A., and Scharth, M. "Predicting time-varying parameters with parameter-driven and observation-driven models." *Review of Economics and Statistics*, 98(1):97–110 (2016).

Lachenbruch, P. A. "Comparisons of two-part models with competitors." *Statistics in medicine*, 20(8):1215–1234 (2001).

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

Lawless, J. F. "Negative binomial and mixed Poisson regression." *Canadian Journal of Statistics*, 15(3):209–225 (1987).

LeBlanc, D. J., Bao, S., Sayer, J. R., and Bogard, S. "Longitudinal driving behavior with integrated crash-warning system: Evaluation from naturalistic driving data." *Transportation research record*, 2365(1):17–21 (2013).

Lee, S.-K. and Jin, S. "Decision tree approaches for zero-inflated count data." *Journal of applied statistics*, 33(8):853–865 (2006).

Lele, S. R., Dennis, B., and Lutscher, F. "Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods." *Ecology letters*, 10(7):551–563 (2007).

Lele, S. R., Nadeem, K., and Schmuland, B. "Estimability and likelihood inference for generalized linear mixed models using data cloning." *Journal of the American Statistical Association*, 105(492):1617–1625 (2010).

Leng, C., Lin, Y., and Wahba, G. "A note on the lasso and related procedures in model selection." *Statistica Sinica*, 1273–1284 (2006).

Leung, S. F. and Yu, S. "On the choice between sample selection and two-part models." *Journal of econometrics*, 72(1-2):197–229 (1996).

Liu, C., Zhao, M., Li, W., and Sharma, A. "Multivariate random parameters zero-inflated negative binomial regression for analyzing urban midblock crashes." *Analytic methods in accident research*, 17:32–46 (2018).

Liu, H. and Powers, D. A. "Bayesian inference for zero-inflated Poisson regression models." *Journal of Statistics: Advances in Theory and Applications*, 7(2):155–188 (2012).

Liu, H. and Zhang, J. "Estimation consistency of the group lasso and its applications." In *Artificial Intelligence and Statistics*, 376–383 (2009).

Liu, L., Shih, Y.-C. T., Strawderman, R. L., Zhang, D., Johnson, B. A., Chai, H., et al. "Statistical Analysis of Zero-Inflated Nonnegative Continuous Data: A Review." *Statistical Science*, 34(2):253–279 (2019).

Liu, L., Strawderman, R. L., Johnson, B. A., and O'Quigley, J. M. "Analyzing repeated measures semi-continuous data, with application to an alcohol dependence study." *Statistical Methods in Medical Research*, 25(1):133–152 (2016).

Liu, Q., Shepherd, B. E., Li, C., and Harrell Jr, F. E. "Modeling continuous response variables using ordinal regression." *Statistics in medicine*, 36(27):4316–4335 (2017).

Long, D. L., Preisser, J. S., Herring, A. H., and Golin, C. E. "A marginalized zero-inflated Poisson regression model with overall exposure effects." *Statistics in medicine*, 33(29):5151–5165 (2014).

Lord, D. and Mannering, F. "The statistical analysis of crash-frequency data: a review and

assessment of methodological alternatives." *Transportation research part A: policy and practice*, 44(5):291–305 (2010).

Lord, D., Washington, S. P., and Ivan, J. N. "Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory." *Accident Analysis & Prevention*, 37(1):35–46 (2005).

Ma, Y., Hu, B., Chan, C.-Y., Qi, S., and Fan, L. "Distractions intervention strategies for in-vehicle secondary tasks: An on-road test assessment of driving task demand based on real-time traffic environment." *Transportation Research Part D: Transport and Environment*, 63:747–754 (2018).

MacDonald, I. L. and Bhamani, F. "A Time-Series Model for Underdispersed or Overdispersed Counts." *The American Statistician*, 1–12 (2018).

Mahapatro, S. R., Singh, P., and Singh, Y. "How effective health insurance schemes are in tackling economic burden of healthcare in India." *Clinical Epidemiology and Global Health*, 6(2):75–82 (2018).

Malyshkina, N. V. and Mannering, F. L. "Zero-state Markov switching count-data models: An empirical assessment." *Accident Analysis & Prevention*, 42(1):122–130 (2010).

Malyshkina, N. V., Mannering, F. L., and Tarko, A. P. "Markov switching negative binomial models: an application to vehicle accident frequencies." *Accident Analysis & Prevention*, 41(2):217–226 (2009).

Mannering, F. and Hensher, D. A. "Discrete/continuous econometric models and their application to transport analysis." *Transport Reviews*, 7(3):227–244 (1987).

Manning, W. G., Basu, A., and Mullahy, J. "Generalized modeling approaches to risk adjustment of skewed outcomes data." *Journal of health economics*, 24(3):465–488 (2005).

Manning, W. G. and Mullahy, J. "Estimating log models: to transform or not to transform?" *Journal of health economics*, 20(4):461–494 (2001).

Matsui, H. "Selection of variables and decision boundaries for functional data via bi-level selection." *arXiv preprint arXiv:1702.02010* (2017).

McGehee, D. V., Roe, C. A., Boyle, L. N., Wu, Y., Ebe, K., Foley, J., and Angell, L. "The wagging foot of uncertainty: data collection and reduction methods for examining foot pedal behavior in naturalistic driving." *SAE International journal of transportation safety*, 4(2):289–294 (2016).

Miaou, S.-P. "The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions." *Accident Analysis & Prevention*, 26(4):471–482 (1994).

Miaou, S.-P., Song, J. J., and Mallick, B. K. "Roadway traffic crash mapping: a space-time modeling approach." *Journal of transportation and Statistics*, 6:33–58 (2003).

Mills, E. D. "Adjusting for covariates in zero-inflated gamma and zero-inflated log-normal models for semicontinuous data." Ph.D. thesis, University of Iowa (2013).

Min, Y. and Agresti, A. "Modeling nonnegative data with clumping at zero: a survey." *Journal of the Iranian Statistical Society*, 1(1):7–33 (2002).

Moulton, L. H., Curriero, F. C., and Barroso, P. F. "Mixture models for quantitative HIV RNA data." *Statistical Methods in Medical Research*, 11(4):317–325 (2002).

Mwalili, S. M., Lesaffre, E., and Declerck, D. "The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research." *Statistical methods in medical research*, 17(2):123–139 (2008).

Najm, W., Stearns, M., Howarth, H., Koopmann, J., Hitz, J. S., et al. "Evaluation of

an automotive rear-end collision avoidance system." Technical report, Department of Transportation, National Highway Traffic Safety (2006).

Nardi, Y., Rinaldo, A., et al. "On the asymptotic properties of the group lasso estimator for linear models." *Electronic Journal of Statistics*, 2:605–633 (2008).

Neelon, B., O'Malley, A. J., and Smith, V. A. "Modeling zero-modified count and semi-continuous data in health services research part 2: case studies." *Statistics in medicine*, 35(27):5094–5112 (2016).

Neelon, B., Zhu, L., and Neelon, S. E. B. "Bayesian two-part spatial models for semi-continuous data with application to emergency department expenditures." *Biostatistics*, 16(3):465–479 (2015).

Negahban, S. N., Ravikumar, P., Wainwright, M. J., Yu, B., et al. "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers." *Statistical Science*, 27(4):538–557 (2012).

Nelder, J. A. and Wedderburn, R. W. "Generalized linear models." *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384 (1972).

Nelson, K. P. and Leroux, B. G. "Statistical models for autocorrelated count data." *Statistics in Medicine*, 25(8):1413–1430 (2006).

Nobre, A. A., Carvalho, M. S., Griep, R. H., Fonseca, M. d. J. M. d., Melo, E. C. P., Santos, I. d. S., and Chor, D. "Multinomial model and zero-inflated gamma model to study time spent on leisure time physical activity: an example of ELSA-Brasil." *Revista de saude publica*, 51:76 (2017).

Oh, M.-S. and Lim, Y. B. "Bayesian analysis of time series Poisson data." *Journal of Applied Statistics*, 28(2):259–271 (2001).

Panzner, M. and Cimiano, P. "Comparing hidden markov models and long short term memory neural networks for learning action representations." In *International Workshop on Machine Learning, Optimization, and Big Data*, 94–105. Springer (2016).

Patton, A. J. "A review of copula models for economic time series." *Journal of Multivariate Analysis*, 110:4–18 (2012).

Pitt, M., Chan, D., and Kohn, R. "Efficient Bayesian inference for Gaussian copula regression models." *Biometrika*, 93(3):537–554 (2006).

Pittman, B., Buta, E., Krishnan-Sarin, S., O'Malley, S. S., Liss, T., and Gueorguieva, R. "Models for Analyzing Zero-Inflated and Overdispersed Count Data: An Application to Cigarette and Marijuana Use." *Nicotine & Tobacco Research* (2018).

Pizer, S. D. and Prentice, J. C. "Time is money: outpatient waiting times and health insurance choices of elderly veterans in the United States." *Journal of Health Economics*, 30(4):626–636 (2011).

Plummer, M., Stukalov, A., and Denwood, M. "rjags: Bayesian graphical models using MCMC, 2016." *URL http://CRAN. R-project. org/package= rjags. R package version*, 2:0–4 (2017).

Qian, W., Yang, Y., and Zou, H. "Tweedie's compound poisson model with grouped elastic net." *Journal of Computational and Graphical Statistics*, 25(2):606–625 (2016).

Quddus, M. A. "Time series count data models: an empirical application to traffic accidents." *Accident Analysis & Prevention*, 40(5):1732–1741 (2008).

Rangapuram, S. S., Seeger, M. W., Gasthaus, J., Stella, L., Wang, Y., and Januschowski, T. "Deep state space models for time series forecasting." In *Advances in neural information processing systems*, 7785–7794 (2018).

Ranney, T. A., Baldwin, G., Smith, L. A., Mazzae, E. N., and Pierce, R. S. "Detection response task (DRT) evaluation for driver distraction measurement application." Technical report (2014).

Rizopoulos, D. *GLMMadaptive: Generalized Linear Mixed Models using Adaptive Gaussian Quadrature* (2019). R package version 0.5-1.
URL `https://CRAN.R-project.org/package=GLMMadaptive`

Rosenfeld, A., Bareket, Z., Goldman, C. V., LeBlanc, D. J., and Tsimhoni, O. "Learning drivers' behavior to improve adaptive cruise control." *Journal of Intelligent Transportation Systems*, 19(1):18–31 (2015).

Rowden, P., Matthews, G., Watson, B., and Biggs, H. "The relative impact of work-related stress, life stress and driving environment stress on driving outcomes." *Accident Analysis & Prevention*, 43(4):1332–1340 (2011).

Rupp, G. "The tactile detection task as a method for assessing drivers' cognitive load." (2010).

Saei, A., Ward, J., and McGilchrist, C. "Threshold models in a methadone programme evaluation." *Statistics in Medicine*, 15(20):2253–2260 (1996).

Shen, X. and Ye, J. "Adaptive model selection." *Journal of the American Statistical Association*, 97(457):210–221 (2002).

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. "A sparse-group lasso." *Journal of Computational and Graphical Statistics*, 22(2):231–245 (2013).

Sólymos, P. "dclone: Data Cloning in R." *R Journal*, 2(2) (2010).

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van der Linde, A. "The deviance information criterion: 12 years on." *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 485–493 (2014).

Spissu, E., Pinjari, A. R., Pendyala, R. M., and Bhat, C. R. "A copula-based joint multinomial discrete–continuous model of vehicle type choice and miles of travel." *Transportation*, 36(4):403–422 (2009).

Stasinopoulos, D. M., Rigby, R. A., et al. "Generalized additive models for location scale and shape (GAMLSS) in R." *Journal of Statistical Software*, 23(7):1–46 (2007).

Stathopoulos, A. and Karlaftis, M. "Temporal and spatial variations of real-time traffic data in urban areas." *Transportation Research Record*, 1768(1):135–140 (2001).

Street, A., Jones, A., and Furuta, A. "Cost-sharing and pharmaceutical utilisation and expenditure in Russia." *Journal of Health Economics*, 18(4):459–472 (1999).

Sukhai, A., Jones, A. P., Love, B. S., and Haynes, R. "Temporal variations in road traffic fatalities in South Africa." *Accident Analysis & Prevention*, 43(1):421–428 (2011).

Swallow, B., Buckland, S. T., King, R., and Toms, M. P. "Bayesian hierarchical modelling of continuous non-negative longitudinal data with a spike at zero: An application to a study of birds visiting gardens in winter." *Biometrical Journal*, 58(2):357–371 (2016).

Tang, F. and Cavanaugh, J. E. "State-Space Models for Binomial Time Series with Excess Zeros." In *Time Series Analysis and Applications*. IntechOpen (2017).

Tang, Y., Xiang, L., and Zhu, Z. "Risk factor selection in rate making: EM adaptive LASSO for zero-inflated poisson regression models." *Risk Analysis*, 34(6):1112–1127 (2014).

Taylor, S. and Pollard, K. "Hypothesis tests for point-mass mixture data with application toomics data with many zero values." *Statistical applications in genetics and molecular biology*, 8(1) (2009).

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

Tobin, J. "Estimation of relationships for limited dependent variables." *Econometrica: journal of the Econometric Society*, 24–36 (1958).

Uc, E. Y. and Rizzo, M. "Driving and neurodegenerative diseases." *Current neurology and neuroscience reports*, 8(5):377 (2008).

Uc, E. Y., Rizzo, M., Anderson, S., Dastrup, E., Sparks, J., and Dawson, J. "Driving under low-contrast visibility conditions in Parkinson disease." *Neurology*, 73(14):1103–1110 (2009).

Vangala, P., Lord, D., and Geedipally, S. R. "Exploring the application of the negative binomial–generalized exponential model for analyzing traffic crash data with excess zeros." *Analytic methods in accident research*, 7:29–36 (2015).

Vincent, M. and Hansen, N. R. "Sparse group lasso and high dimensional multinomial classification." *Computational Statistics & Data Analysis*, 71:771–786 (2014).

Walker, A. M. "On the asymptotic behaviour of posterior distributions." *Journal of the Royal Statistical Society: Series B (Methodological)*, 31(1):80–88 (1969).

Wang, H. and Leng, C. "Unified LASSO estimation by least squares approximation." *Journal of the American Statistical Association*, 102(479):1039–1048 (2007).

—. "A note on adaptive group lasso." *Computational statistics & data analysis*, 52(12):5277–5286 (2008).

Wang, H., Li, R., and Tsai, C.-L. "Tuning parameter selectors for the smoothly clipped absolute deviation method." *Biometrika*, 94(3):553–568 (2007).

Wang, Y., Wang, T., and Zhuang, J. "Modeling continuous time series with many zeros and an application to earthquakes." *Environmetrics*, 29(4):e2500 (2018).

Wang, Z., Ma, S., and Wang, C.-Y. "Variable selection for zero-inflated and overdispersed data with application to health care demand in Germany." *Biometrical Journal*, 57(5):867–884 (2015).

Washington, S. P., Karlaftis, M. G., and Mannering, F. *Statistical and econometric methods for transportation data analysis*. Chapman and Hall/CRC (2010).

Wurm, M. J., Rathouz, P. J., and Hanlon, B. M. "Regularized ordinal regression and the ordinalNet R package." *arXiv preprint arXiv:1706.05003* (2017).

Xiong, H. and Boyle, L. N. "Drivers' adaptation to adaptive cruise control: Examination of automatic and manual braking." *IEEE transactions on intelligent transportation systems*, 13(3):1468–1473 (2012).

Xiong, Y., Tobias, J. L., and Mannering, F. L. "The analysis of vehicle crash injury-severity data: A Markov switching approach with road-segment heterogeneity." *Transportation research part B: methodological*, 67:109–128 (2014).

Yan, X., Wang, B., An, M., and Zhang, C. "Distinguishing between rural and urban road segment traffic safety based on zero-inflated negative binomial regression models." *Discrete Dynamics in Nature and Society*, 2012 (2012).

Yang, M., Cavanaugh, J. E., and Zamba, G. K. "State-space models for count time series with excess zeros." *Statistical Modelling*, 15(1):70–90 (2015).

Yang, M., Zamba, G. K., and Cavanaugh, J. E. "Markov regression models for count time series with excess zeros: A partial likelihood approach." *Statistical Methodology*, 14:26–38 (2013).

Yang, S., Harlow, L. L., Puggioni, G., and Redding, C. A. "A comparison of different methods of zero-inflated data analysis and its application in health surveys." *Journal of Modern Applied Statistical Methods* (2017).

Yannis, G. and Karlaftis, M. G. "Weather effects on daily traffic accidents and fatalities: a time series count data approach." In *Proceedings of the 89th Annual Meeting of the Transportation Research Board*, volume 10, 14 (2010).

Yu, S., Wang, G., Wang, L., Liu, C., and Yang, L. "Estimation and Inference for Generalized Geoadditive Models." *Journal of the American Statistical Association*, 1–27 (2019).

Yuan, M. and Lin, Y. "Model selection and estimation in regression with grouped variables." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67 (2006).

Zeng, P., Wei, Y., Zhao, Y., Liu, J., Liu, L., Zhang, R., Gou, J., Huang, S., and Chen, F. "Variable selection approach for zero-inflated count data via adaptive lasso." *Journal of Applied Statistics*, 41(4):879–894 (2014).

Zhang, C., Chen, N., and Zhang, L. "Time series of multivariate zero-inflated Poisson counts." In *2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 1365–1369. IEEE (2016).

Zhang, C.-H. et al. "Nearly unbiased variable selection under minimax concave penalty." *The Annals of statistics*, 38(2):894–942 (2010).

Zhao, P. and Yu, B. "On model selection consistency of Lasso." *Journal of Machine learning research*, 7(Nov):2541–2563 (2006).

Zheng, X., Zaheer, M., Ahmed, A., Wang, Y., Xing, E. P., and Smola, A. J. "State space LSTM models with particle MCMC inference." *arXiv preprint arXiv:1711.11179* (2017).

Zhou, H., Yang, Y., and Qian, W. "Tweedie Gradient Boosting for Extremely Unbalanced Zero-inflated Data." *arXiv preprint arXiv:1811.10192* (2018).

Zhou, X., Kang, K., and Song, X. "Two-part hidden Markov models for semicontinuous longitudinal data with nonignorable missing covariates." *Statistics in Medicine* (2020).

Zhou, X.-H. and Tu, W. "Interval estimation for the ratio in means of log-normally distributed medical costs with zero values." *Computational statistics & data analysis*, 35(2):201–210 (2000).

Zhu, F. "A negative binomial integer-valued GARCH model." *Journal of Time Series Analysis*, 32(1):54–67 (2011).

—. "Zero-inflated Poisson and negative binomial integer-valued GARCH models." *Journal of Statistical Planning and Inference*, 142(4):826–839 (2012).

Zou, H. "The adaptive lasso and its oracle properties." *Journal of the American statistical association*, 101(476):1418–1429 (2006).

Zou, H. and Hastie, T. "Regularization and variable selection via the elastic net." *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320 (2005).

Zou, H., Hastie, T., Tibshirani, R., et al. "On the "degrees of freedom" of the lasso." *The Annals of Statistics*, 35(5):2173–2192 (2007).