Nonparametric Rank Based Inferences for Generalized Linear Models, Longitudinal Data Analysis, and Variable Selection

by

Guy-vanie Marcias Miakonkana

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 3, 2013

Keywords: Robustness, Nonparametric, Generalized Linear Models, Longitudinal Data Analysis, Variable Selection.

Approved by

Asheber Abebe, Chair, Associate Professor of Mathematics & Statistics
Geraldo S. De Souza, Professor of Mathematics & Statistics
Peng Zeng, Associate Professor of Mathematics & Statistics
Mark D. Carpenter, Professor of Mathematics & Statistics

Abstract

Many relevant data sets from environmental sciences, biomedical sciences, finance, insurance, engineering, and many other disciplines have high-dimensionality, difficult to model dependence structure, outliers , and heavy tailed and asymmetric noise distribution. These challenges posed by the data require the use of robust statistical techniques in order to make reliable inferences. Many robust nonparametric statistical methods have been developed to address these challenging issues. Rank estimators are among statistical methods recently developed for this purpose.However little attention has been given to Rank estimation in Generalized Linear Models, Longitudinal Data Analysis, and Variable Selection. This dissertation proposes robust nonparametric methods based on the theory of rank for inferences in Generalized Linear Models, Longitudinal Data Analysis, and Group Variable Selection in Linear Models.

Acknowledgments

This research project leaves some memories of people to whom I would like to express my gratitude. First and foremost, I would like to thank God Almighty for his numerous graces, including the good health that I enjoy today and the successful completion of this project. Words cannot express how grateful I am towards my parents, brothers, and sisters including my late father Miakonkana Paul for their love and support throughout my life.I would like to express my deepest gratitude to my advisor Dr. Asheber Abebe for his amiability, encouragement, guidance, patience throughout this work and my graduate studies in general . The completion of this work would not have been possible without the endless support of my advisor. I am very grateful to Professor Charles E. Chidume for his continued encouragement and fatherly advice, including his recommendation to the mathematics PhD program at Auburn University. I would like to thank Dr.Geraldo S. De Souza, Dr. Mark D. Carpenter, and Dr. Peng Zeng for consenting to serve as my PhD committee members. I am very grateful to Dr. Geraldo S. De Souza, not only for serving as my committee member but also for his encouragement, amiability and endless support throughout my graduate studies. I also want to express my gratitude to Dr. Floyd Woods for agreeing to serve as the University reader of my dissertation. Last, but not least, special thanks to all my friends, who in a way or another continued to encourage me during the years I spent in graduate school, they made this journey easier.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Generalized linear models (Nelder and Wedderburn, 1972) provide a unified approach to many of the most common statistical procedures used in applied statistics. They have applications in disciplines ranging as widely as agriculture, demography, ecology, economics, education, engineering, environmental studies and pollution, geography, geology, history, medicine, political science, psychology, sociology and many others. In many studies, in the above mentioned fields, measurements are made over time yielding dependent observations. Dependence among observations in a data set may also occur when measurements are made at nearby locations in space. The purpose of longitudinal data analysis (Liang and Zeger, 1986) is to develop statistical models that take into account the presence and the nature of the dependence among the measurements. Another challenge encountered in many current data sets is the so-called high dimensionality, i.e data sets with a massive number of variables, usually far exceeding the number of observations. Examples of these data sets include microarray gene expression data and data sets from image and signal processing. The emergence of high dimensional data has, more than ever, driven researchers to extensive development of methods for simultaneous estimation and variable selection (Tibshirani, 1996).

There has been continued interest in the development of the theory and methodology related to generalized linear models, longitudinal data analysis, and variable selection methods. Since the fundamental work of Nelder and Wedderburn (1972), many statistical methods have been proposed for the generalized linear models. Wedderburn (1974) proposed the least squares and the quasi-likelihood estimators. These estimators are asymptotically efficient

in the sense that that the limiting variance-covariance matrix attains a Cramer-Rao-type lower bound. Recently, Gao et al. (2012) have developed asymptotic properties of maximum quasi-likelihood estimators in generalized linear models with adaptive designs. Other recent developments in generalized linear models include Song et al. (2012) who have proposed a method for hypothesis testing in generalized linear models with functional coefficient autoregressive processes and Liu and Yuan (2012) on combining quasi and empirical likelihoods in generalized linear models with missing responses. Other recent works in this area, to name a few, are She (2012), Hardin and Hilbe (2012),Augustin et al. (2012), Mbachu et al. (2012), Abarin and Wang (2012), and Klar and Meintanis (2012). We also refer to McCullagh and Nelder (1989) for a comprehensive account of the generalized linear models and quasi-likelihood based inference procedures. Following the development of the approach of generalized estimating equations (GEE) (Liang and Zeger, 1986) for longitudinal data analysis, extensive research on longitudinal data analysis has given rise to a rich literature. Wang and Carey (2004) provided a method to supplement and enhance the GEE by constructing unbiased estimating equations from working correlation models for irregularly timed repeated measures. Hin et al. (2007) developed a criteria for selection of working-correlation-structure in GEE. Zhang (2011)studied generalized estimating equations and Gaussian estimation for longitudinal data analysis. Other significant contributions in longitudinal data analysis are the works of Tang and Leng (2011), Bandyopadhyay et al. (2011), Nakai and Ke (2011), Wang and Hin (2010), Copas and Seaman (2010), Cheng et al. (2013), and Tsai et al. (2011), just to mention a few.

Lasso regression (Tibshirani, 1996) for simultaneous estimation and variable selection has triggered extensive continued research on methods for penalized regression models. Penalized regression methods have received a lot of attention and popularity among statisticians, recently. The SCAD proposed by Fan and Li (2001) is another popular work in simultaneous variable selection and estimation in regression . Knight and Fu (2000) studied the asymptotic properties of the Lasso. Zou (2006) studied the adaptive Lasso, and showed

that the adaptive Lasso has the so-called oracle property. Zou and Hastie (2005) proposed the elastic-net penalty for variable selection in regression with dependent predictors. Zou and Zhang (2009) enhanced the elastic-net with the adaptive elastic-net. The group Lasso (Yuan and Lin, 2006) is a natural extension of the Lasso in a regression model with grouped variables. The group variable selection method given in Yuan and Lin (2006) and its ramifications have been further studied by many. These include Simon and Tibshirani (2012), Chen and Hero (2012), Hirose and Konishi (2012), Wang and Leng (2008), and Nardi and Rinaldo (2008).

Many of the statistical methods described above for generalized linear models, longitudinal data analysis, as well as variable selection are based on maximum likelihood estimation or least squares technique for parameter estimation. Despite their many good properties, it is well known that such least squares or maximum likelihood based methods may have very poor performance in data set containing outliers or heavy tailed asymmetric noise distribution. In this situation, it is desirable to use a robust estimation procedure. One way to address the lack of robustness in regression models is to employ the so-called M-estimator (Huber, 1981). Properties of the M-estimator and its multiple refinements (Klein and Yohai (1981), Collins and Portnoy (1981), Prakasa Rao (1981), and many others) have been extensively studied over the years for robust estimation of parameters in regression models. An alternative approach to develop robust estimators in regression is to use the theory of rank estimation. Rank estimation for a simple linear models was originally proposed by Adichie (1967) based on simple Hodges-Lehmann type location estimators. Jurečková (1971) and Jaeckel (1972) later generalized this to multiple regression.A comprehensive treatment of rank estimation for linear models can be found in Hettmansperger and McKean (1998). Rank estimator for nonlinear regression models have been studied by many others, among which, Abebe and McKean (2007), and Bindele and Abebe (2012). Johnson and Peng (2008) as well as Wang and Li (2009) have recently proposed rank based procedures for penalized

regression. However, not much attention has been given to rank estimation in generalized linear models and longitudinal data analysis.

1.2 Contribution of the Dissertation

In this dissertation we develop an iterative rank based estimator for parameter estimation in generalized linear models and its extension to longitudinal data analysis. In addition, we propose a rank based variable selection method for linear regression models with grouped variables. This generalizes the methods described in Johnson and Peng (2008) and Wang and Li (2009) to linear models with either categorical predictors or other type of grouped variables.

1.3 Outline of the Dissertation

Chapter 2 contains a brief review of generalized linear model and maximum likelihood estimation, generalized estimating equations and maximum quasi-likelihood estimation, penalized linear regression methods, and the theory of Rank based estimation for linear models; organized in four sections.We develop the iterative rank based procedure for parameter estimation in generalized linear models in Chapter 3. The chapter also contains results on asymptotic results of the estimator as well simulation studies and real world data examples that illustrate the theoretical results. Chapter 4 extends the procedure described in Chapter 3 to the case of dependent responses, giving rise to rank based estimation procedure for longitudinal data. Simulation studies and a data example are also provided. In Chapter 5, we study penalized linear regression with grouped predictors. We penalize a rank based objective function with the group adaptive Lasso type of penalty function. The oracle property of the estimator is established. Simulation studies confirm this.

Chapter 2

Preliminaries

2.1 Generalized Linear Models

2.1.1 Exponential Family of Distribution and Generalized Linear Models

Let y be a response variable and $\mathbf{x}$ be a vector of predictors. Assume that both $\mathbf{x}$ and y are random and that we have a random sample $(\mathbf{x}_i, y_i),\ i = 1, \ldots, n$. Recall the linear regression model

$$E(y_i|\mathbf{x}_i = \mathbf{x}) = \mu_i = \mathbf{x}_i^t \boldsymbol{\theta}_0 \ , \quad y_i \sim N(\mu_i, \sigma^2) \tag{2.1.1}$$

where $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta} \subset \mathbb{R}^p$ is an unknown vector of parameters.
Advances in theoretical and computational statistics have allowed us to use method analogous to those developed for linear regression models in the following more general situations:

1. Response variables, y, have distributions other than the normal distribution, they may even be categorical rather than continuous

2. Relationship between the response and explanatory variables need not be of the simple linear form in (2.1.1).

These more general models are referred to as *generalized linear models*(GLM). This term was coined by Nelder and Wedderburn (1972). They proposed a generalization of the linear regression model in (2.1.1) as follows:

$$h[E(y_i|\mathbf{x}_i = \mathbf{x})] = h(\mu_i) = \mathbf{x}_i^t \boldsymbol{\theta}_0 \ , \tag{2.1.2}$$

where

1. h is some monotone differentiable function, called the *link function*

2. y_i, $i = 1, \cdots, n$ are independent random variables with mean μ_i, each. They share the same distribution from the exponential dispersion family. That is, the probability density function of y_i has the form

$$f(y_i; \beta_i, \phi) = exp\Big[\frac{y_i\beta_i - C(\beta_i)}{\phi} + B(y_i, \phi)\Big];$$

where $B(.)$ and $C(.)$ are known functions, and the range of y_i does not dependent on β or ϕ. In this formulation, the parameter β_i is called the *canonical parameter* and is a function of the mean, that is $\beta_i = d(\mu_i)$ for some function d. When $h \equiv d$, h is called the *canonical link function.* The parameter ϕ is called the *dispersion parameter.*Many well known distributions such as Normal, Binomial, Poisson, Gamma are members of the exponential family of distribution. The choice of the functions B and C determines the particular member of the exponential family of distribution.The exponential family of distribution has very "nice" properties. Among others, the mean and the variance of y_i are given by

$$E(y_i) = \dot{C}(\beta_i) \qquad Var(y_i) = \phi\ddot{C}(\beta_i) \tag{2.1.3}$$

where $\dot{C}$ and $\ddot{C}$ are, respectively, the first and second derivative of the function C.

For fixed $\mathbf{x}_i$, we have

$$\ddot{C}(\beta_i) = \frac{\partial\dot{C}(\beta_i)}{\partial\beta_i} = \frac{\partial\mu_i}{\partial\beta_i} \equiv V(\mu_i)$$

that is

$$Var(y_i) = \phi V(\mu_i). \tag{2.1.4}$$

V is called the *variance function.* It relates the mean to the variance of y_i.

Table 2.1: Exponential Family of Distributions

Distribution	β	$C(\beta)$	ϕ	$E(y)$	$V(\mu)$
$\mathbf{B}(n,\pi)$	$\ln\Big(\frac{\pi}{1-\pi}\Big)$	$n\ln(1+e^{\beta})$	1	$n\pi$	$n\pi(1-\pi)$
$\mathbf{P}(\mu)$	$\ln\mu$	e^{β}	1	μ	μ
$\mathbf{N}(\mu,\sigma^2)$	μ	$\frac{1}{2}\beta^2$	σ^2	μ	1
$\mathbf{G}(\mu,\nu)$	$-\frac{1}{\mu}$	$-\ln(-\beta)$	$\frac{1}{\nu}$	μ	μ^2
$\mathbf{IG}(\mu,\sigma^2)$	$-\frac{1}{2\mu^2}$	$-\sqrt{-2\beta}$	σ^2	μ	μ^3
$\mathbf{NB}(\mu,\kappa)$	$\ln\Big(\frac{\kappa\mu}{1+\kappa\mu}\Big)$	$-\frac{1}{\kappa}\ln\Big(1-\kappa e^{\beta}\Big)$	1	μ	$\mu(1+\kappa\mu)$

As equation (2.1.2) indicates, in *generalized linear models* the mean of y_i is related to the predictor $\mathbf{x}_i$. Thus the mean varies with the explanatory variables. As the mean varies so does the variance, trough $V(\mu_i)$. So the model in equation (2.1.2) also gives a relationship between the predictor $\mathbf{x}_i$ and the variance of y_i. However there are many mean-variance relationship that cannot be captured with an exponential family density, even distribution for which the theory of *generalized linear models* is valid. This issue is addressed by the *quasi-likelihood* method discussed below.

Table 2.1.1 gives a summary of different choices of the functions B, C and d leading to different distributions in the exponential family. For simplicity of the presentation, we will drop the index i in the table, that is we write μ in stead of μ_i, β in stead of β_i, and y in stead of y_i.

2.1.2 Maximum Likelihood Estimation

Consider the model in (2.1.2), where y_i is from an exponential family of distribution; that is the log-likelihood function of $y_1, \cdots, y_n$ is given by.

$$l(\boldsymbol{\theta},\phi) = \sum_{i=1}^{n} lnf(y_i,\boldsymbol{\theta},\phi) = \sum_{i=1}^{n}\Big(\frac{y_i d(\mu_i) - C(d(\mu_i))}{\phi} + B(y_i,\phi)\Big) \tag{2.1.5}$$

where

$$\mu_i = h^{-1}(\mathbf{x}_i^t \boldsymbol{\theta}).$$

The maximum likelihood estimation (MLE) of $\boldsymbol{\theta}$ and ϕ are obtained by maximizing the log-likelihood $l(\boldsymbol{\theta}, \phi)$. Consider the MLE of $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}_j$ denotes the jth component of $\boldsymbol{\theta}$. The MLE of $\boldsymbol{\theta}$ is a solution to the system of equations

$$\frac{\partial l}{\partial \boldsymbol{\theta}_j} = \sum_{i=1}^{n} \frac{\partial l}{\partial \beta_i} \frac{\partial \beta_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\theta}_j} = 0 \ , \tag{2.1.6}$$

that is

$$\sum_{i=1}^{n} \frac{\partial \beta_i}{\partial \eta_i} (y_i - \dot{C}(\beta_i)) x_{ij} = \sum_{i=1}^{n} \frac{\partial \beta_i}{\partial \eta_i} (y_i - \mu_i) x_{ij} = 0 \ , \tag{2.1.7}$$

with $\eta_i = \mathbf{x}_i^t \boldsymbol{\theta}$; $\beta_i = d(\mu_i) = d(h^{-1}(\eta_i))$; and x_{ij} is the jth element of $\mathbf{x}_i$.

Noting that

$$\Big(\frac{\partial \beta_i}{\partial \eta_i}\Big)^{-1} = \frac{\partial \eta_i}{\partial \beta_i} = \frac{\partial \eta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_i} = \dot{h}(\mu_i) V(\mu_i)$$

and

$$\frac{\partial \mu_i}{\partial \boldsymbol{\theta}_j} = \frac{x_{ij}}{\dot{h}(\mu_i)}$$

the equation (2.1.7) can be rewritten as

$$\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \boldsymbol{\theta}_j} \frac{(y_i - \mu_i)}{V(\mu_i)} = 0 \ . \tag{2.1.8}$$

The equations (2.1.8) are usually referred to as estimating equations. They depend on $\boldsymbol{\theta}$ through $\mu_i = h^{-1}(\mathbf{x}_i^t \boldsymbol{\theta})$.

The MLE requires full specification of the density of y_i. However, in some practical situations the distribution of y_i may be completely unknown. Estimation of $\boldsymbol{\theta}$ is still possible with the *maximum quasi-likelihood* if the mean-variance relationship of the type (2.1.4) is

specified. The *maximum quasi-likelihood* estimator of $\boldsymbol{\theta}$ is the minimizer of the function

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V(t)} dt \ . \tag{2.1.9}$$

It is straightforward to show that

$$\frac{\partial Q}{\partial \boldsymbol{\theta}} = \frac{\partial l}{\partial \boldsymbol{\theta}}, \tag{2.1.10}$$

that is, the MLE is identical to the *maximum quasi-likelihood* when the response y_i is from an exponential family of distribution.

While the MLE (equivalently the *maximum quasi-likelihood*) of $\boldsymbol{\theta}$ remains widely used due to its many good properties, it is vulnerable to outlying observations in the data. This deficiency is addressed in this dissertation.

2.2 Longitudinal Data Analysis

While applications of generalized linear models are abundant, there are many situations in which repeated response measurements are made on the same unit, yielding a cluster of dependent observations. These measurements are obtained either prospectively(such as in clinical trial), following subjects forward in time, or retrospectively, by extracting multiple measurements on each subject from historical records.The defining advantage with this type of study is that one can distinguish changes over time within individuals from differences at fixed times. The assumption of correlation is not confined to observations made over time on the same individuals. Observations made at nearby locations in space may also be correlated. For example, in agriculture studies we may have observations made on the same small experimental area. The need to account for the correlation in the data gave rise to special statistical methods suited for the analysis of longitudinal data.

Liang and Zeger (1986) applied the quasi-likelihood approach to longitudinal data by proposing the generalized estimating equations (GEE) described below.

Consider the longitudinal data $(y_{ij}, \mathbf{x}_{ij})$, $j = 1, 2, \cdots, m_i$ and $i = 1, 2, \cdots, n$, where the response y_{ij} is the measurement on the ith subjects at time j, and $\mathbf{x}_{ij}$ the corresponding p-dimensional vector of predictors. Assume that the mean of y_{ij}, $\mu_{ij} = E(y_{ij})$, is related to the predictor $\mathbf{x}_{ij}$ through the model in equation (2.1.2). That is

$$h[E(y_{ij}|\mathbf{x}_{ij} = \mathbf{x})] = \mathbf{x}_{ij}^t \boldsymbol{\theta}_0. \tag{2.2.1}$$

The assumptions made in model (2.1.2) remain valid here.

Let $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ denote, respectively, the mean and the variance-covariance matrix of, $\mathbf{Y}_i = (y_{i1}, \cdots, y_{im_i})$, the response from the ith subject. Similarly, let $\mathbf{X}_i = (\mathbf{x}_{i1}, \cdots, \mathbf{x}_{im_i})$. For a $s \times 1$ vector of unknown parameters α, let $\mathbf{R}_i = \mathbf{R}_i(\alpha)$ denote a $m_i \times m_i$ matrix. Define the matrix

$$\mathbf{V}_i = \mathbf{A}_i^{1/2} \mathbf{R}_i \mathbf{A}_i^{1/2} / \phi \tag{2.2.2}$$

where $\mathbf{A}_i = diag\{\partial\mu_{i1}/\partial\beta_{i1}, \cdots, \partial\mu_{im_i}/\partial\beta_{im_i}\}$. Note that the matrix $\mathbf{V}_i$ may or may not be the variance-covariance matrix of $\mathbf{Y}_i$.Nonetheless, we refer to $\mathbf{R}_i$ as the correlation matrix. Liang and Zeger (1986) GEE approach defines the estimator of $\boldsymbol{\theta}_0$ in (2.2.1) as the solution of the equation

$$\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}} \widehat{\mathbf{V}}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0} \tag{2.2.3}$$

where the matrix $\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\theta} = \{\partial\boldsymbol{\mu}_{ij}/\partial\boldsymbol{\theta}_l\}_{jl}$, $\boldsymbol{\theta}_l$ the lth element of $\boldsymbol{\theta}$, and $\widehat{\mathbf{V}}_i$ is an estimate of $\mathbf{V}_i$. In fact the estimating equation (2.2.3) is the *quasi-score equation* (2.1.8) for correlated data. Note that in addition to the lack of robustness, both in $\mathbf{x}$ and y direction, the GEE estimator of $\boldsymbol{\theta}_0$ requires a specification and estimation of the dependence structure of the elements of $\mathbf{Y}_i$ prior to estimating $\boldsymbol{\theta}_0$. In addition, the GEE may be less efficient when the correlation structure is incorrectly specified, even though they are still consistent.The method developed in this dissertation, though, does not require the specification, nor does

it require the estimation of the correlation structure,prior to estimating $\boldsymbol{\theta}_0$. In addition, the resulting estimator is robust in the response space, y.

2.3 Variable Selection in Regression

Recall the *linear regression model* :

$$y_i = \mathbf{x}_i^t \boldsymbol{\theta}_0 + \epsilon_i \tag{2.3.1}$$

where $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})$ is a p-dimensional random vector of predictors and the random vectors $(y_1, \mathbf{x}_1), \cdots, (y_n, \mathbf{x}_n)$ are independent and identically distributed. Model (2.3.1) is identical to model (2.1.1) if $\epsilon_i \sim N(0, \sigma^2)$. However this assumption will not be made in the model considered in this work. Assume, instead, that $\boldsymbol{\theta}_0$ is sparse, i.e, most components of $\boldsymbol{\theta}_0$ are exactly 0. An example of such $\boldsymbol{\theta}_0$ is $(1, 0, 0, 0, 0, 0.8, 0, 0, 2, 0, 0, 0, 0, 0, 0, 2.1)$. The purpose of variable selection is to simultaneously estimate $\boldsymbol{\theta}_0$ and identify the predictors in $(x_{i1}, \cdots, x_{ip})$ that are associated with the non-zero components of $\boldsymbol{\theta}_0$ as well as the ones associated with the zero components of $\boldsymbol{\theta}_0$. Traditional variable selection procedures use best-subset selection and its step-wise variants. However, best-subset selection is computationally prohibitive when the number of predictors, p, is large. In the attempt to address these fundamental issues of subset selection, penalized regression methods have been introduced. In particular, the Lasso method proposed by Tibshirani (1996) is very popular for simultaneous variable selection and estimation. We will discuss the Lasso and its variants below.

2.3.1 Lasso and Adaptive Lasso Regression

The Lasso (Tibshirani, 1996) estimator of $\boldsymbol{\theta}_0$ is obtained by minimizing the l_1 penalized least squares, i.e

$$\widehat{\boldsymbol{\theta}}(Lasso) = \underset{\Theta}{\text{Argmin}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^t \boldsymbol{\theta})^2 + \lambda \sum_{j=1}^{p} |\theta_j| \tag{2.3.2}$$

where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)$, and λ some regularization parameter to be specified. For some chosen λ, the l_1 penalty, $\sum_{j=1}^{p} |\theta_j|$ term enables the Lasso to simultaneously regularize the least squares fit and shrink some components of $\widehat{\boldsymbol{\theta}}(Lasso)$ to zero in order to recover the sparsity of $\boldsymbol{\theta}_0$.

Despite its popularity, the Lasso does have two serious deficiencies: instability in high-dimensional data and a non-ignorable bias, asymptotically, for estimating the nonzero coefficients (Fan and Li, 2001). Zou (2006) further showed that the Lasso could be inconsistent for variable selection unless the predictor matrix satisfies a rather strong regularity condition. In order to overcome this drawback, Zou (2006) proposed the following adaptive Lasso estimator

$$\widehat{\boldsymbol{\theta}}(AdaLasso) = \underset{\Theta}{\text{Argmin}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^t \boldsymbol{\theta})^2 + \lambda \sum_{j=1}^{p} \hat{w}_j |\theta_j| \ , \tag{2.3.3}$$

where $\{\hat{w}_j\}$ are some data-driven weights and can be computed by $\hat{w}_j = (|\widehat{\theta}_j^0|)^{-\gamma}$, where γ is a positive constant and $\widehat{\boldsymbol{\theta}}^0$ is an initial root-n consistent estimate of $\boldsymbol{\theta}_0$. Note that the weights $\hat{w}_j = (|\widehat{\theta}_j^0|)^{-\gamma}$ are adaptive in nature. That is, if the effect of a predictor on the response is strong (equivalently the corresponding component in $\boldsymbol{\theta}_0$ is nonzero), the corresponding coefficient is lightly penalized and vice-versa. In fact, Zou (2006) showed that with an appropriately chosen λ, the adaptive Lasso has the so-called *oracle property*, that is the adaptive Lasso works as well as if the correct submodel was known in advance.

2.3.2 Grouped Variable Selection

Both the Lasso and the Adaptive Lasso perform individual variables(predictors) selection. However, in some regression problems, predictors may present group structures. Grouping structures can arise for many reasons, and require different modeling strategies in variable selection. Common examples include the representation of multilevel categorical covariates in a regression model by a group of indicator variables, and the representation of

the effect of a continuous variable by a set of basis functions. The collinearity among predictors, usually encountered in high dimensional data regression problem, can also be a source of natural groupings among predictors, as is often the case in gene expression and genetic association studies. In the regression with categorical covariates problem, the interest lies in selecting important factors or groups. In some situations, like in genetic association studies, where the groups are naturally present in the data and often unknown to the investigator, selection of groups is just as important as selection of individual variables within a group. Building on the ideas of Lasso and adaptive Lasso, several adjustments to variable selection have been proposed to respond to these challenges. Yuan and Lin (2006) proposed the group Lasso. In order to improve the performance of the group Lasso, which suffers the same drawback as the Lasso, Wang and Leng (2008) improved the group Lasso into the adaptive group Lasso given by

$$\widehat{\boldsymbol{\theta}}(AdaGrpLasso) = \underset{\Theta}{\text{Argmin}} \sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{K} \mathbf{x}_{ik}^t \theta_k \right)^2 + n \sum_{k=1}^{K} \hat{w}_k \|\theta_k\| \,, \tag{2.3.4}$$

where $\theta_k = (\theta_{k1}, \cdots, \theta_{kp_k})$ and $\mathbf{x}_{ik} = (x_{ik1}, \cdots, x_{ikp_k})$ are, respectively, the regression coefficient and the vector of predictors associated with the kth group (factor).
Note that the adaptive group Lasso and many other related variable selection methods are suitable for the situations where the group structure is known to the investigator. A regression model with categorical variables (factors) as the only group variables is an example of such a situation.

Zou and Hastie (2005) proposed the Elastic-Net for simultaneous group structure identification and parameter estimation in high-dimensional linear models. The Elastic-Net identifies important and irrelevant groups in high dimensional linear models even when the grouping structure in the predictors is unknown to the investigator. This is particularly useful when the groupings in the data are due to collinearity among the predictors. The Elastic-Net was further improved by Zou and Zhang (2009) who proposed the following adaptive Elastic-Net

$$\widehat{\boldsymbol{\theta}}(AdaEnet) = (1+\frac{\lambda_2}{n})\Big\{ \underset{\Theta}{\text{Argmin}} \sum_{i=1}^{n} \Big(y_i - \mathbf{x}_i^t\boldsymbol{\theta}\Big)^2 + \lambda_2\|\boldsymbol{\theta}\|_2^2 + \lambda_1 \sum_{j=1}^{p} \hat{w}_j|\theta_j| \Big\} \ , \qquad (2.3.5)$$

where $\|\cdot\|_2$ denotes the L_2 norm, and λ_1 and λ_2 two parameters to be specified.

All the variable selection methods described so far and similar methods that are based on regularized least squares objective function as well as penalized maximum likelihood are known to have a poor performance when the data are contaminated with outliers or the error term ϵ_i, in (2.3.1) has a heavy tailed or skewed distribution (Johnson and Peng, 2008). In this dissertation, we propose a robust group variable selection method based on a rank objective function.

2.4 Robust Rank Based Regression

The regression model in (2.3.1) can be rewritten as

$$y_i = \alpha + \mathbf{x}_i^t\boldsymbol{\theta}_0 + \epsilon_i \ , \qquad (2.4.1)$$

where α is the intercept and $\mathbf{x}_i^t$ is a $p-1$ dimensional vector of predictors. It is convenient to have the location parameter α in the regression model for estimation in rank regression. A large body of literature exists of the estimation of parameters α and $\boldsymbol{\theta}_0$ as well as tests of linear hypotheses concerning them (Hettmansperger and McKean, 1998). However, we will limit our discussion to estimation theory.

2.4.1 Estimation

Consider the following operator

$$\|v\|_\varphi = \sum_{i=1}^{n} a(R(v_i))v_i, \qquad (2.4.2)$$

where $v = (v_1, \cdots, v_n) \in \mathbb{R}^n$, $R(v_i)$ is the rank of v_i among $v_1, \cdots, v_n$; $a(1) \leq a(2) \leq \cdots \leq a(n)$ is a set of scores generated as $a(i) = \varphi(i/(n+1))$ for some nondecreasing score function $\varphi(u)$ defined on the interval $(0, 1)$ and standardized such that $\int \varphi(u)du = 0$ and $\int \varphi^2(u)du = 1$.

The operator $\|.\|_\varphi$ is a **pseudo-norm** (see Hettmansperger and McKean, 1998); that is, it satisfies the following conditions:

1. $\|v\|_\varphi \geq 0$, for all $v \in \mathbb{R}^n$
2. $\|v\|_\varphi = 0$ if and only if $v_1 = \cdots = v_n$
3. $\|\lambda v\|_\varphi = |\lambda| \|v\|_\varphi$, for all $\lambda \in \mathbb{R}, v \in \mathbb{R}^n$
4. $||u + v||_\varphi \leq ||u||_\varphi + ||v||_\varphi$, for all $u, v \in \mathbb{R}^n$.

A rank estimator of $\boldsymbol{\theta}_0$ is a vector $\widehat{\boldsymbol{\theta}}_\varphi$ such that

$$\widehat{\boldsymbol{\theta}}_\varphi = \underset{\Theta}{\text{Argmin}} \, \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_\varphi \, . \tag{2.4.3}$$

The estimate of α can be obtained as the median of $y_1 - \mathbf{x}_1^t \widehat{\boldsymbol{\theta}}_\varphi, \cdots, y_n - \mathbf{x}_n^t \widehat{\boldsymbol{\theta}}_\varphi$.

The function $\varphi(u) = \sqrt{12}(u - 1/2)$ results in the so called *Wilcoxon* estimator of $\boldsymbol{\theta}_0$ and α.

Let $\mathbf{X}$ denote the $n \times p$ matrix whose ith row is $\mathbf{x}_i^t$, and $\Omega(\mathbf{X})$ the column space spanned by the columns of $\mathbf{X}$. Without loss of generality, assume that $\alpha = 0$. Geometrically, the rank estimator of $\boldsymbol{\theta}_0$ is a vector that minimizes the distance between $\mathbf{Y} = (y_1, \cdots, y_n)$ and $\Omega(X)$.

It is easy to see that the geometry of the rank estimation in the linear model is identical to the one of the least squares estimation (Hettmansperger and McKean, 1998). However, the rank estimator is robust to outliers in the response y, while the least squares estimator is not. The following section discusses the robustness properties of the rank estimation.

2.4.2 Robustness Properties

The *influence function*(IF) of an estimator (Hampel, 1974) is a measure of the sensitivity of the estimator to local changes. It provides an approximation of the behavior of the estimator when the sample contains a small fraction t of identical outliers. Let $\widehat{\boldsymbol{\theta}}$ denote the estimator of interest, and $F \equiv F(\mathbf{x}, y)$ a probability distribution function. The influence function of $\widehat{\boldsymbol{\theta}}$ at a point $(\mathbf{x}_0, y_0)$ is defined as

$$\mathrm{IF}(\mathbf{x}_0, y_0, \widehat{\boldsymbol{\theta}}, F) = \lim_{t \to 0} \frac{\widehat{\boldsymbol{\theta}}((1-t)F + t\Delta_{(\mathbf{x}_0, y_0)}) + \widehat{\boldsymbol{\theta}}(F)}{t}\,, \tag{2.4.4}$$

where $\Delta_{\mathbf{x}_0, y_0}$ denotes the point mass at $(\mathbf{x}_0, y_0)$.

An estimator is robust if its influence function IF is a bounded function of both $\mathbf{x}$ and y. If the IF is a bounded function of y (respectively $\mathbf{x}$) only, we say that the estimator $\widehat{\boldsymbol{\theta}}$ is robust in the y (respectively $\mathbf{x}$) direction. That is, infinitesimal contaminations in y (respectively in $\mathbf{x}$) do not affect the estimator.

In fact, for the model (2.3.1) if we can write

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Lambda_n(\mathbf{x}_i, y_i, \boldsymbol{\theta}_0) + o_p(1), \tag{2.4.5}$$

where Λ_n is a known function, then the influence of $\widehat{\boldsymbol{\theta}}$ is given by

$$\mathrm{IF}(\mathbf{x}, y, \widehat{\boldsymbol{\theta}}, F) = \lim_{n \to \infty} \Lambda_n(\mathbf{x}, y, \boldsymbol{\theta}_0). \tag{2.4.6}$$

(See Appendix 5.2 of Hettmansperger and McKean (1998)).

The following theorem (Corollary 3.5.7 of Hettmansperger and McKean (1998)) gives the influence function of the rank estimator in linear models.

Theorem 2.1. *Under some regularity conditions (see Hettmansperger and McKean (1998)) the estimator* $\widehat{\boldsymbol{\theta}}_\varphi$ *of* $\boldsymbol{\theta}_0$ *in model* (2.3.1) *has the following asymptotic representation*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_\varphi - \boldsymbol{\theta}_0) = \tau_\varphi (n^{-1}\mathbf{X}^t\mathbf{X})^{-1} n^{-1/2} \sum_{i=1}^{n} \varphi(F(y_i - \mathbf{x}_i^t\boldsymbol{\theta}_0))\mathbf{x}_i + o_p(1)$$

where $(\mathbf{x}_i, y_i)$ *are independent and identically distributed from* F *and* τ_φ *some constant depending on* φ *and the distribution of* ϵ_i; $i = 1, \cdots, n$.

Therefore, the influence function of $\widehat{\boldsymbol{\theta}}_\varphi$, given by

$$\mathrm{IF}(\mathbf{x}, y, \widehat{\boldsymbol{\theta}}_\varphi, F) = \lim_{n\to\infty} \tau_\varphi (n^{-1}\mathbf{X}^t\mathbf{X})^{-1} \varphi(F(y - \mathbf{x}^t\boldsymbol{\theta}_0))$$

is a bounded function of y, since the distribution function F is such that $0 < F < 1$. That is, the rank estimator is robust in the y direction.

In this dissertation we establish a similar result for the rank estimator of parameters in *generalized linear models* ((2.4.6)) as well as *longitudinal data analysis*.

Chapter 3

Rank Based Generalized Linear Models

3.1 Introduction

Let y be a response variable and $\mathbf{x}$ be a vector of predictors. Assume that both $\mathbf{x}$ and y are random and that we have a random sample $(\mathbf{x}_i, y_i),\ i = 1, \ldots, n$. We consider the generalized linear regression model

$$h[E(y_i|\mathbf{x}_i = \mathbf{x})] = \mathbf{x}^t\boldsymbol{\theta}_0, \tag{3.1.1}$$

where $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta} \subset \mathbb{R}^p$ is an unknown vector of parameters. We assume that $y_1, y_2, \ldots, y_n$ are independent absolutely continuous random variables with distribution in the exponential family of distributions, $\mathbf{x}_i \in \mathbb{X} \subset \mathbb{R}^p, \quad 1 \leq i \leq n$, are independent random vectors. The function h is a known function such that its inverse $g \equiv h^{-1}$ is a real valued function defined on the set $U = \{u \mid u = \mathbf{x}^t\boldsymbol{\theta} \text{ for } \boldsymbol{\theta} \in \boldsymbol{\Theta} \text{ and } \mathbf{x} \in \mathbb{X}\} \subset \mathbb{R}$, is monotone and three times continuously differentiable. We shall assume that, $\mathbb{X}$ is compact, $\boldsymbol{\Theta}$ is convex and compact, and $\boldsymbol{\theta}_0$ is an interior point of $\boldsymbol{\Theta}$.

Since the fundamental work of Nelder and Wedderburn (1972), there has been continued interest in the development of the theory and the methodology related to generalized linear models to estimate the parameter $\boldsymbol{\theta}_0$. Wedderburn (1974) proposed the least squares and the quasi-likelihood estimators. These estimators are asymptotically efficient in the sense that the limiting variance-covariance matrix attains a Cramer-Rao-type lower bound. We also refer to McCullagh and Nelder (1989) for a comprehensive account of the generalized linear models and quasi-likelihood based inference procedures. However, in the presence of outliers it is desirable to use a robust estimation procedure. Pregibon (1982), Stefanski

et al. (1986), and Künsch et al. (1989) considered the robust estimation of generalized linear models parameters with particular emphasis on logistic regression. Morgenthaler (1992) studied least absolute deviations fits for generalized linear models. Robust M-estimators in logistic regression model were proposed by Kordzakhia et al. (2001). An alternative way to develop robust estimators is to use rank based procedures. The rank-based approach has not been described in the literature and is the focus of this work. Our rank based method uses the so called Wilcoxon objective function that provides us with an initial estimator which is, afterwards, updated iteratively. The procedure results in estimators that are robust in the response space. Extensions of this method give us bounded influence and high-breakdown estimators.

Rank estimation for the linear regression model was proposed by Jurečková (1971) and Jaeckel (1972). Over the years, several extensions and refinements of the rank regression approach were proposed. A comprehensive treatise is given in Hettmansperger and McKean (1998). Particularly relevant to our discussion are the works of Jung and Ying (2003) and Abebe and McKean (2007). Jung and Ying (2003) studied the linear model with correlated and non-i.i.d errors. Although our proposed method is developed for independent errors, it allows for arbitrary link functions with minimal assumption on the error distribution. Abebe and McKean (2007) studied the Wilcoxon estimator of a general nonlinear regression function. The initial estimator we use in the iteratively defined rank estimator proposed in this chapter is a special case of the estimator developed in Abebe and McKean (2007).
The remainder of this chapter is organized as follows. After introducing the model and the estimators in Section 3.2, the consistency and the asymptotic normality of the rank version of the maximum quasi-likelihood estimator are studied in Section 3.3. We illustrate the robustness and the efficiency of the estimators in Section 3.4 via simulation studies and real world data examples. Section 3.5 provides the conclusion. Proofs and technical details are found in Section 3.6.

3.2 Iterative Rank Estimator

Take $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and define the Pearson residuals as $z_i(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = (y_i - g(\mathbf{x}_i^t\boldsymbol{\theta}))/\phi\sqrt{\nu(g(\mathbf{x}_i^t\boldsymbol{\theta}_0))}$, $1 \leq i \leq n$, where ν is a continuous function (McCullagh and Nelder, 1989, Page 30) related to the variance of y_i through $var(y_i) = \phi^2\nu(g(\mathbf{x}_i^t\boldsymbol{\theta}_0))$ and the dispersion parameter $\phi > 0$ will be assumed to be an unknown constant and the function ν will be assumed to be twice continuously differentiable.The assumptions on g and ν are the same as $M3$ in Chiou and Müller (1999). Consider the estimator of $\boldsymbol{\theta}$ defined as the minimizer of the rank dispersion function of Jaeckel (1972)

$$W_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{R(z_i(\boldsymbol{\theta}, \boldsymbol{\theta}_0))}{n+1} - \frac{1}{2}\right] z_i(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \ ,$$

where $R(z_i(\boldsymbol{\theta}, \boldsymbol{\theta}_0))$ is the rank of $z_i(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ among $z_1(\boldsymbol{\theta}, \boldsymbol{\theta}_0), \ldots, z_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$. Since $\boldsymbol{\theta}_0$ is not known, the dispersion function cannot be directly minimized. As a solution, we plug-in an initial estimator of $\boldsymbol{\theta}_0$ in W_n and minimize the resulting dispersion function with respect to $\boldsymbol{\theta}$. That is, given an initial estimator $\widehat{\boldsymbol{\theta}}_n^0$, $\widehat{\boldsymbol{\theta}}_n^k = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\text{Argmin}}\, W_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_n^{k-1})$ for $k = 1, 2, \ldots$. Note that, by Lemma 2 of Jennrich (1969), $\widehat{\boldsymbol{\theta}}_n^k$ exists because $W_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_n^{k-1})$ is continuous as a function of $\boldsymbol{\theta}$ on the compact space $\boldsymbol{\Theta}$ for k fixed. This process gives rise to a Fisher scoring scheme given by

$$\widehat{\boldsymbol{\theta}}_n^k = \widehat{\boldsymbol{\theta}}_n^{k-1} + \left[\dot{\Psi}_n^k(\widehat{\boldsymbol{\theta}}_n^{k-1})\right]^{-1}\Psi_n^k(\widehat{\boldsymbol{\theta}}_n^{k-1}) \qquad k = 1, 2, 3, \cdots \ , \tag{3.2.1}$$

where $\Psi_n^k(\boldsymbol{\theta})$ is the rank score function defined by

$$\Psi_n^k(\boldsymbol{\theta}) \equiv W_n'(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_n^{k-1}) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{r_i(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_n^{k-1})}{n+1} - \frac{1}{2}\right]\frac{g'(\mathbf{x}_i^t\boldsymbol{\theta})}{\sqrt{\nu(g(\mathbf{x}_i^t\widehat{\boldsymbol{\theta}}_n^{k-1}))}}\mathbf{x}_i \tag{3.2.2}$$

and $\dot{\Psi}_n^k(\boldsymbol{\theta}) = d\Psi_n^k(\boldsymbol{\theta})/d\boldsymbol{\theta}$. Here $r_i(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_n^{k-1})$ is the rank of $e_i(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_n^{k-1})$ among $e_1(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_n^{k-1}), \ldots, e_n(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_n^{k-1})$, where

$$e_i(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_n^{k-1}) = \frac{y_i - g(\mathbf{x}_i^t\boldsymbol{\theta})}{\sqrt{\nu(g(\mathbf{x}_i^t\widehat{\boldsymbol{\theta}}_n^{k-1}))}} \ , \quad i = 1, \ldots, n \ .$$

The quantity ϕ is removed from the denominator since it has no influence on the ranks and hence the kth step estimator $\widehat{\boldsymbol{\theta}}_n^k$.

As an initial estimator, we propose the *naïve* Wilcoxon estimator

$$\widehat{\boldsymbol{\theta}}_n^0 = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\text{Argmin}}\, W_n^0(\boldsymbol{\theta}) \ ,$$

where

$$W_n^0(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{R(e_i(\boldsymbol{\theta}))}{n+1} - \frac{1}{2}\Big)e_i(\boldsymbol{\theta}) \tag{3.2.3}$$

and $R(e_i(\boldsymbol{\theta}))$ is the rank of the ith raw residual $e_i(\boldsymbol{\theta}) = y_i - g(\mathbf{x}_i^t\boldsymbol{\theta})$ among $e_1(\boldsymbol{\theta}) = y_1 - g(\mathbf{x}_1^t\boldsymbol{\theta}), e_2(\boldsymbol{\theta}) = y_2 - g(\mathbf{x}_2^t\boldsymbol{\theta}), \cdots, \ e_n(\boldsymbol{\theta}) = y_n - g(\mathbf{x}_n^t\boldsymbol{\theta})$. Note that $\widehat{\boldsymbol{\theta}}_n^0$ is a zero of the function $\Psi_n^0(\boldsymbol{\theta})$ defined by

$$\Psi_n^0(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{R(e_i(\boldsymbol{\theta}))}{n+1} - \frac{1}{2}\Big)g'(\mathbf{x}_i\boldsymbol{\theta})\mathbf{x}_i \ . \tag{3.2.4}$$

Once again, because $W_n^0(\boldsymbol{\theta})$ is continuous and $\boldsymbol{\Theta}$ is compact, Lemma 2 of Jennrich (1969) implies the existence of a minimizer of $W_n^0(\boldsymbol{\theta})$.

For independent and identically distributed raw residuals $e_i(\boldsymbol{\theta}) = y_i - g(\mathbf{x}_i^t\boldsymbol{\theta})$, the initial estimator $\widehat{\boldsymbol{\theta}}_n^0$ is referred to as the Wilcoxon estimator of $\boldsymbol{\theta}_0$, in linear and nonlinear regression. It is usually preferred to least squares estimators in the presence of outliers for its robustness in the response space. The theory of the nonlinear Wilcoxon estimator, that includes $\widehat{\boldsymbol{\theta}}_n^0$ as a special case, has been studied by Abebe and McKean (2007).

For the case $g(\mathbf{x}_i^t\boldsymbol{\theta}) = \mathbf{x}_i^t\boldsymbol{\theta}$ and $\nu(.) \equiv 1$ with possibly dependent and non-identically distributed residuals $e_i(\boldsymbol{\theta}), \ 1 \le i \le n$, the method proposed in this chapter for obtaining the initial estimator $\widehat{\boldsymbol{\theta}}_n^0$ and the sequence $\{\widehat{\boldsymbol{\theta}}_n^k\}$ reduces to the method considered by Jung and Ying (2003). The theory presented in this and the next sections, though, is for a general

function g satisfying the conditions given in the introduction, and for independent but not necessarily identically distributed residuals $e_i(\boldsymbol{\theta})$.

In our case, since we are studying generalized linear models, the raw residuals are not as useful as they are in linear models. The Pearson residuals capture the mean-variance relationship that exists in generalized linear models. Although under certain conditions $\widehat{\boldsymbol{\theta}}_n^0$ is asymptotically unbiased, it converges very slowly to $\boldsymbol{\theta}_0$. The iterative scheme defined above for $\widehat{\boldsymbol{\theta}}_n^k$ for $k = 1, 2, 3, \cdots$, updates $\widehat{\boldsymbol{\theta}}_n^0$ to an estimator that converges much faster to $\boldsymbol{\theta}_0$ than $\widehat{\boldsymbol{\theta}}_n^0$.

The estimators $\widehat{\boldsymbol{\theta}}_n^k$, $k = 0, 1, \ldots$, are related to a quasi-likelihood type rank estimator $\widehat{\boldsymbol{\theta}}_n$ that solves

$$\Psi_n(\widehat{\boldsymbol{\theta}}_n) = 0 \ , \tag{3.2.5}$$

where

$$\Psi_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{r_i(\boldsymbol{\theta},\boldsymbol{\theta})}{n+1} - \frac{1}{2}\right]\frac{g'(\mathbf{x}_i^t\boldsymbol{\theta})}{\sqrt{\nu(g(\mathbf{x}_i^t\boldsymbol{\theta}))}}\mathbf{x}_i \ . \tag{3.2.6}$$

Note that, $\Psi_n(\boldsymbol{\theta})$ is a rank version of the quasi-score function given by Wedderburn (1974). Moreover, $\Psi_n(\boldsymbol{\theta}) \equiv \Psi_n^0(\boldsymbol{\theta})$ for $\nu(g(t)) \equiv 1$; that is, $\widehat{\boldsymbol{\theta}}_n \equiv \widehat{\boldsymbol{\theta}}_n^0$ for $\nu(g(t)) \equiv 1$. As Theorem 3.4 in the following section shows, $\widehat{\boldsymbol{\theta}}_n^k \to \widehat{\boldsymbol{\theta}}_n$ as $k \to \infty$ for n fixed; that is, our iterative estimator converges to the quasi-likelihood type estimator based on ranks. This is somewhat intuitive since the score function (3.2.2) used in our iteration scheme satisfies $\Psi_n^k(\widehat{\boldsymbol{\theta}}_n^{k-1}) = \Psi_n(\widehat{\boldsymbol{\theta}}_n^{k-1})$. For simplicity of notation, $r_i(\boldsymbol{\theta},\boldsymbol{\theta})$ will be denoted by $r_i(\boldsymbol{\theta})$.

3.3 Asymptotic Properties

3.3.1 Consistency

We will present sufficient conditions for strong consistency of the rank quasi-likelihood estimator as well as the initial estimator. Let $(\Omega, \mathscr{F}, P)$ be a probability space. Assume, for $i = 1, \cdots, n$, the random vectors $(y_i, \mathbf{x}_i)$, are independent and that y_i and $\mathbf{x}_i$ are each carried by $(\Omega, \mathscr{F}, P)$ for all $i = 1, \cdots, n$. Let $a.s$ convergence denote *almost sure* convergence; that

is, pointwise convergence everywhere on Ω except possibly for an event in $\mathscr{F}$ of probability 0.

Let $\mu(t) = \dfrac{g'(t)}{\sqrt{\nu(g(t))}}$ and $\mu'(t)$ its derivative. Moreover, let $\lambda_{min}(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^t)$ and $\lambda_{max}(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^t)$ denote the minimum and the maximum eigenvalues of $\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^t$, respectively. The following assumptions will be needed in establishing the strong consistency of $\widehat{\boldsymbol{\theta}}_n$:

C_1 : $\inf\limits_{i\in\mathbb{N}} \ \mu'(\mathbf{x}_i^t\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

C_2 : $\lambda_{min}(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^t) \to \infty \ a.s$ as $n \to \infty$, and

C_3 : there exist finite constants $n_0 > 0$ and $c > 0$ such that $\dfrac{\lambda_{max}(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^t)}{\lambda_{min}(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^t)} < c$ for all $n \geq n_0$.

Remark 3.1. *Assumption C_1 is equivalent to the assumption: $\dfrac{d\mu(t)}{dt} > 0$ made in Chen et al. (1999) and assumption C_2 is the same as assumption $C1$ of Chen et al. (1999). Assumption C_3 is equivalent to equation* (3.6) *of Fahrmeir and Kaufmann (1985).*

The following theorem gives the consistency of $\widehat{\boldsymbol{\theta}}_n$. The consistency of $\widehat{\boldsymbol{\theta}}_n^0$ follows as a special case by taking $\nu(g(t)) \equiv 1$ in the proof.

Theorem 3.1. *Under C_1, C_2, and C_3, $\widehat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$ a.s when $n \to \infty$. In fact, $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = O\Big(\lambda_{min}^{-1/2}\Big(\sum_{i=1}^{n-1} \mathbf{x}_i\mathbf{x}_i^t\Big)\Big)$.*

The proof of Theorem 3.1 is given in the appendix. The proof requires the following lemma, Lemma 8 of den Boer and Zwart (2012):

Lemma 3.1. *Let $(\mathbf{x}_i)_{i\in\mathbb{N}}$ be a sequence of vectors in $\mathbb{R}^p$ and $(\omega_i)_{i\in\mathbb{N}}$ a sequence of scalars with $\inf_{i\in\mathbb{N}} \omega_i > 0$. Then for all $n \in \mathbb{N}$*

$$\lambda_{min}\Big(\sum_{i=1}^{n} \omega_i\mathbf{x}_i\mathbf{x}_i^t\Big) \geq \lambda_{min}\Big(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^t\Big) \inf_{i\in\mathbb{N}} \omega_i$$

Remark 3.2. *Note that Lemma 3.1 guarantees the existence of* $\Big(\sum_{i=1}^{n}\omega_i\mathbf{x}_i\mathbf{x}_i^t\Big)^{-1}$ *whenever* $\lambda_{min}\Big(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^t\Big) > 0$.

3.3.2 Asymptotic Normality and Robustness

We will now give conditions needed for the asymptotic normality of $\widehat{\boldsymbol{\theta}}_n$, and $\widehat{\boldsymbol{\theta}}_n^0$ in particular. We will start by defining some quantities that will be used hereafter. Let $\tau \in \mathbb{R}^p$ and define

$$L_n(\boldsymbol{\theta},\tau) = \frac{1}{n}\sum_{i=1}^{n}\gamma_n(\mathbf{x}_i,\boldsymbol{\theta})\Big(\frac{r_i(\boldsymbol{\theta})}{n+1} - \frac{1}{2}\Big)$$
$$s_n^2(\boldsymbol{\theta}) = n^2E\Big(L_n^2(\boldsymbol{\theta},\tau)\Big)$$

where

$$\gamma_n(\mathbf{x}_i,\boldsymbol{\theta}) = \frac{\tau^t\mathbf{x}_i\mu'(\mathbf{x}_i^t\boldsymbol{\theta})}{max_{1\le i\le n}\|\tau^t\mathbf{x}_i\mu'(\boldsymbol{\theta}^t\mathbf{x}_i)\|}\ .$$

The notations L_n and s_n^2 are borrowed from Brunner and Denker (1994) in the particular case of $J(x) = x - 1/2$. As required in Brunner and Denker (1994), we have $\max_{1\le i\le n}|\gamma_n(\mathbf{x}_i,\boldsymbol{\theta})| = 1$. Conditioning on $\mathbf{x}_i$ gives deterministic regression coefficients $\gamma_n(\mathbf{x}_i,\boldsymbol{\theta})$.

Let $\mathbf{X}$ be the $n \times p$ matrix of regressors, with rows $\mathbf{x}_i$. In addition to C_1, C_2, and C_3 consider the following assumptions

$N_1:\ s_n^2(\boldsymbol{\theta}_0) \to \infty$ as $n \to \infty$ and

$N_2:\ \dot{\Psi}_n(\boldsymbol{\theta}_0) \xrightarrow{\mathcal{P}} H_0$ invertible.

The following lemma is a consequence of corollaries 3.4 and 3.6 of Brunner and Denker (1994), with $J(t) = t - \frac{1}{2}$, $m_i = 1$ for all i and $N = n$.

Lemma 3.2. *Under* N_1,

$$\sqrt{n}\Psi_n(\boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N(0,\Omega_0) \quad as\ n \to \infty$$

with $\Omega_0 = \lim_{n\to\infty} nE(\Psi_n(\boldsymbol{\theta}_0)\Psi_n^t(\boldsymbol{\theta}_0))$.

The following theorem gives the asymptotic normality of $\widehat{\boldsymbol{\theta}}_n$. Since taking $\nu(g(t)) \equiv 1$ gives $\widehat{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_n^0$, the asymptotic normality of $\widehat{\boldsymbol{\theta}}_n^0$ follows as a special case of the theorem.

Theorem 3.2. *Under* $C_1 - C_3$ *and* N_1, N_2,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N(\mathbf{0}, H_0^{-1}\Omega_0 H_0^{-1}) \quad as\ n \to \infty\ .$$

The asymptotic representations used in the proof of Theorem 3.2 may be used to obtain the influence function of $\widehat{\boldsymbol{\theta}}_n$ (Hettmansperger and McKean, 1998). As the theorem below shows, the influence function of $\widehat{\boldsymbol{\theta}}_n$ is bounded in y-space but unbounded in $\mathbf{X}$ space. This mirrors the behavior of the Wilcoxon estimator in the linear model.

Theorem 3.3. *Let* $e_i(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) \sim F_j$.*If* $F_j(t) > 0\ \forall j \in N$ *and* $t \in \mathbf{R}$, *then the influence function of the estimator* $\widehat{\boldsymbol{\theta}}_n$ *is*

$$IF(\widehat{\boldsymbol{\theta}}_n; \mathbf{x}, y) = \mathbf{x}\mu(\mathbf{x}^t\boldsymbol{\theta}_0)\Big(\bar{F}\Big(\frac{y - g(\mathbf{x}^t\boldsymbol{\theta}_0)}{\sqrt{\nu(g(\mathbf{x}^t\boldsymbol{\theta}_0))}}\Big) - .5\Big)H_0^{-1}$$

where $\bar{F} = \lim_{n\to\infty} \bar{F}_n$; $\bar{F}_n = \dfrac{1}{n}\sum_{j=1}^n F_j$;

Remark 3.3. *Note that the influence function of* $\widehat{\boldsymbol{\theta}}_n$ *depends on the response* y *only through the function* $\bar{F}$, *and that* $0 < \bar{F} < 1$. *Since* $\bar{F}$ *is a bounded function (therefore bounded in* y*), so is the influence function of* $\widehat{\boldsymbol{\theta}}_n$. *This in particular implies that the influence function of* $\widehat{\boldsymbol{\theta}}_n^0$ *is also bounded with respect to the response* y, *since it is a special case (* $\nu(g(t)) \equiv 1$*) of the influence function of* $\widehat{\boldsymbol{\theta}}_n$. *In fact, it is a well known fact (cf. Abebe and McKean, 2007; Hettmansperger and McKean, 1998) that the influence function of the initial Wilcoxon estimator* $\widehat{\boldsymbol{\theta}}_n^0$ *used in this work is bounded in the* y *direction. However, minimizing the function*

$$W_n(\boldsymbol{\theta}, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \left[\frac{R(z_i(\boldsymbol{\theta}, \boldsymbol{\theta}))}{n+1} - \frac{1}{2}\right] z_i(\boldsymbol{\theta}, \boldsymbol{\theta})\ ,$$

will result in an estimator that is not robust in either x or y direction, since the gradient of $W_n(\boldsymbol{\theta}, \boldsymbol{\theta})$ is unbounded in both x and y direction; giving rise to unbounded influence function. Hence the choice of $W_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$.

3.3.3 Iterations

The theorem below establishes the relationship between the quasi-likelihood rank type estimator, $\widehat{\boldsymbol{\theta}}_n$, and the sequence $\{\widehat{\boldsymbol{\theta}}_n^k\}_k$, as the number of iterations (k) increases while holding the sample size (n) constant. As the following theorem shows, increasing k leads to the quasi-likelihood rank type estimator.

Theorem 3.4. *For n fixed,*

$$\lim_{k\to\infty} \widehat{\boldsymbol{\theta}}_n^k = \widehat{\boldsymbol{\theta}}_n \ ,$$

where, again, $\Psi_n(\widehat{\boldsymbol{\theta}}_n) = 0$ with $\Psi_n(\boldsymbol{\theta})$ defined in (3.2.6).

3.4 Simulations and Example

3.4.1 Monte Carlo Simulations

We conducted a small simulation study in order to evaluate the finite sample performance of the proposed rank based k-step estimator. The behavior of the estimator was studied for different sample sizes and various covariate types. The boxplots, the mean and the mean squared error were produced from $s = 200$ estimates of the parameter corresponding to 200 simulated data sets. The process was repeated in the presence of outliers and compared to the corresponding quantities based on the maximum likelihood estimates of the parameter.

We considered two settings for the simulation study. In both settings, the true parameter vector was taken to be $\boldsymbol{\theta}_0 = (2, 1)$, and each response y_i was generated from a gamma distribution with shape parameter according to a model with the log link function and scale parameter fixed at 6. In the first setting, the predictor $\mathbf{x}_i = (x_{i1}, x_{i2})$ is such that

$x_{i1} \sim$ Bernoulli(0.5) and $x_{i2} \sim$ Bernoulli(0.4) whereas in the second setting we take $x_{i1} \sim$ Bernoulli(0.5) and $x_{i2} \sim N(0, 1)$.

Figure 1 presents the boxplots of the estimates and Figure 2 gives a comparison of the accompanying mean squared error for the first simulation setting. Figures 3 and 4 give the same summary for the second setting. The contamination of the response was done by replacing the maximum element of $y = (y_1, y_2, \cdots, y_n)$ by eight times its original value.

Figures 1 – 4 reveal that the proposed rank estimator (labelled as "Ours") is comparable to the maximum likelihood estimator (labelled as "MLE") when there is no outlier in the data. In addition, as Figures 2 and 4 show, the MSE goes to zero as the sample size increases for both MLE and rank. This illustrates our theoretical result that the proposed estimator converges to the true parameter in L_2 norm. It is also observed that, in the presence of a gross outlier, the MLE loses both its accuracy and its precision whereas the proposed rank estimator remains unaffected. This can well be viewed on the boxplots (Figures 1 and 3) and the right panels of Figures 2 and 4.

Figure 3.1: Boxplot of the estimates

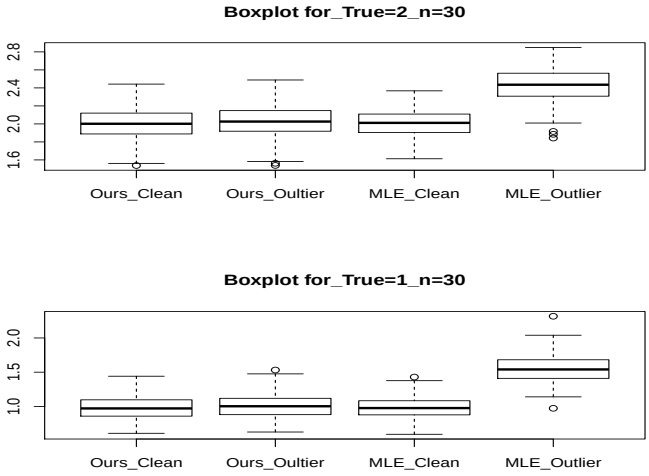


Figure 3.2: MSE comparison

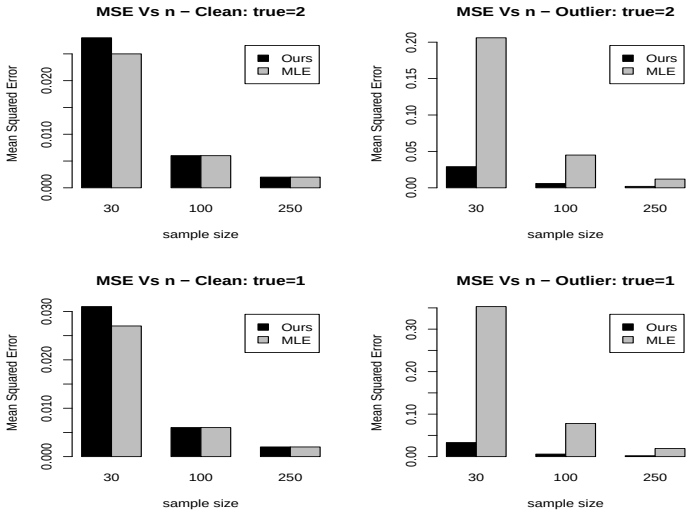


Figure 3.3: Boxplot of the estimates

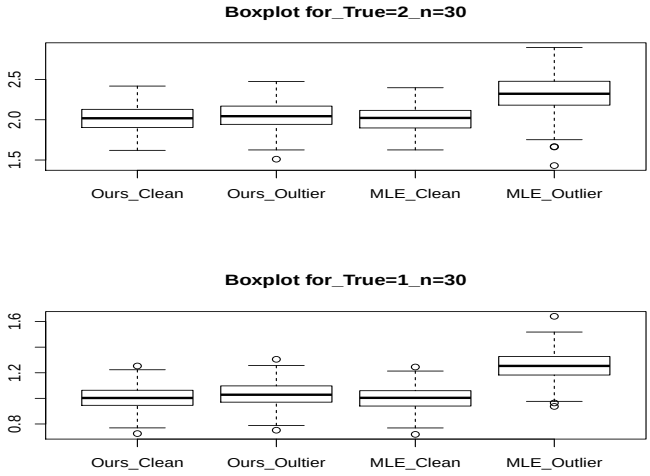
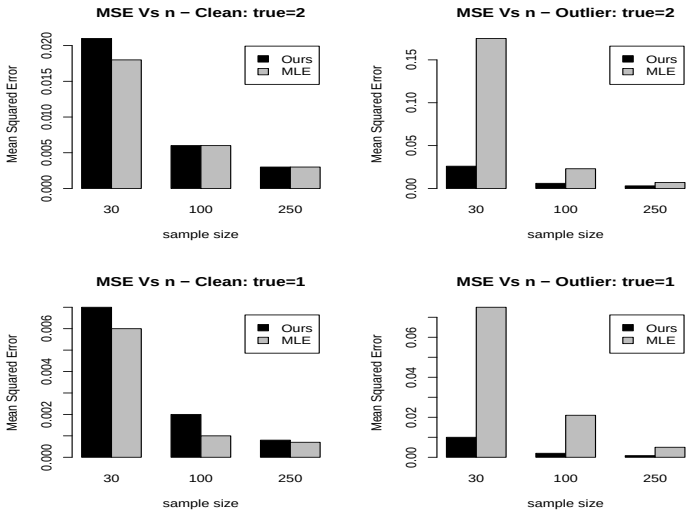


Figure 3.4: MSE comparison



3.4.2 Example 1: Swedish Third Party Motor Insurance

A real world data example was considered to compare the performance of the proposed estimator to that of the maximum likelihood estimator on real data. The findings reveal that our estimator is robust to local contamination of the response and is comparable to the maximum likelihood estimator when the data have no outliers.

The data were compiled by the Swedish Committee on the Analysis of Risk Premium in Motor Insurance in 1977. The Committee was asked to evaluate the real influence on claims of the risk arguments and to compare this structure with the actual tariff. Among other variables, the following variables were considered: the total amount of the claim, the number of claims, kilometers traveled by the automobile per year, and the make of the car. In this work, we study the relationship between the average amount of the claim and the make of the car as well as the kilometers traveled by year.The variable "make of the automobile" has 9 categories. Categories 1 to 8 represent eight different common car models and all other models are combined in class 9. The variable "kilometers traveled per year" was categorized into 5 groups based on the number of kilometers traveled per year. These categories were $[0, 1000), [1000, 15000), [15000, 20000), [20000, 25000)$, and $[25000, \infty)$. We used the data from the largest cities and their surroundings (zone 1) that resulted in a sample size of 295. The original complete study for this data set can be found in Hallin and Ingenbleek (1983). The data set for zone 1 is found in Andrews and Herzberg (1985). Traditionally the average claims amount is modeled using a gamma distribution. Therefore, we fitted a gamma generalized linear model with response the "average claim amount" and regressors "the make of the car" and "kilometers traveled per year". We created an outlier by replacing the average claim amount in position 70 of y by 100000. For our data, the average maximum payment was 31442. So, a payment of 100000 is not out of realm. One may think of a total-loss expensive luxury vehicle. The results of the study are summarized in Table 3.1 as well as Figures 5 and 6.

Table 3.1: Coefficient (Coef) and Standard Error (SE) estimates for Swedish Third Party Motor Insurance

	Coef - Rank		Coef - MLE		SE - Rank		SE - MLE	
	Clean	Outlier	Clean	Outlier	Clean	Outlier	Clean	Outlier
intercept	8.369	8.355	8.397	8.000	0.077	0.081	0.057	0.445
x2	0.112	0.135	0.089	1.523	0.059	0.066	0.055	1.455
x3	0.082	0.082	0.064	0.098	0.061	0.062	0.056	0.112
x4	0.076	0.079	0.082	0.125	0.075	0.078	0.068	0.126
x5	0.116	0.117	0.111	0.124	0.082	0.083	0.080	0.088
z2	0.080	0.081	0.072	0.197	0.079	0.082	0.071	0.267
z3	0.112	0.111	0.143	0.190	0.145	0.148	0.114	0.249
z4	-0.122	-0.119	-0.093	0.079	0.126	0.129	0.110	0.307
z5	-0.112	-0.112	-0.103	-0.096	0.098	0.097	0.094	0.264
z6	-0.027	-0.031	0.021	0.086	0.098	0.100	0.092	0.299
z7	-0.131	-0.131	-0.106	-0.113	0.135	0.133	0.114	0.297
z8	0.196	0.169	0.328	0.354	0.275	0.274	0.176	0.282
z9	-0.025	-0.010	-0.031	0.470	0.058	0.066	0.045	0.545

Table 3.1 shows the rank estimates of the coefficients are comparable with the maximum likelihood estimates. The maximum likelihood estimates of the coefficients are sensitive to the outlier whereas the rank estimates remain relatively unchanged by the introduction of a gross outlier. The standard errors of the coefficients based on the rank procedure are quite comparable with the standard error estimates based on the maximum likelihood procedure for the clean data. Similar to the coefficients, rank standard errors remain relatively unchanged when the outlier is included. However, the maximum likelihood based estimates of the standard error are inflated by the outlier. The iterative scheme converged for $k = 3$ and $k = 4$, respectively, for the clean data and contaminated data.

The Pearson and deviance residual plots (Figures 5 and 6, respectively) show that the outlier inflated both the Pearson and deviance residuals of the MLE. We can also observe that the outlier induced a clustering effect on the MLE estimated responses. This can result in unusually high or unusually low estimates of insurance premiums. This is not true for the rank estimator as the outlier has no apparent effect on the residual plots.

Figure 3.5: Pearson residuals plot for clean and contaminated response

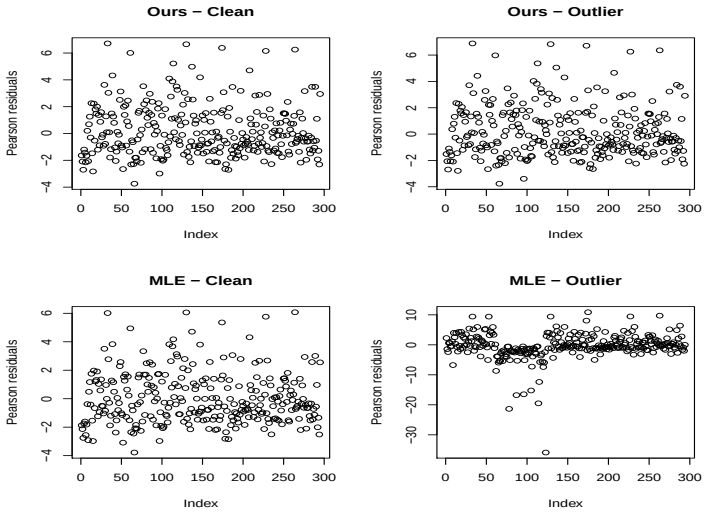
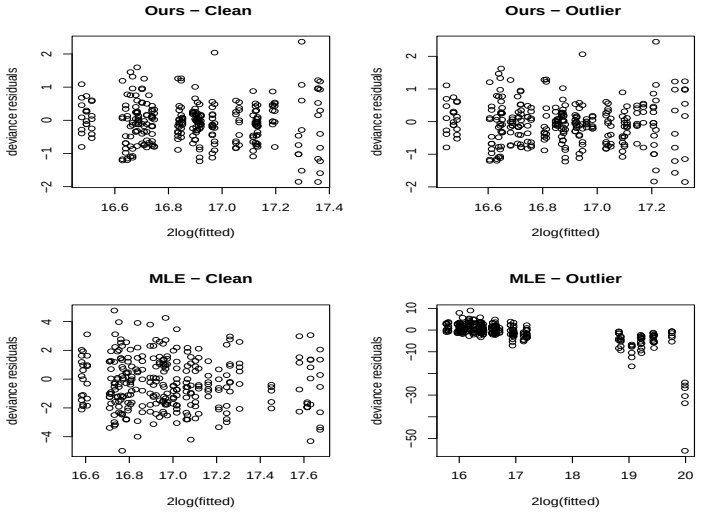


Figure 3.6: Deviance residuals vs Fitted values for clean and contaminated response



3.4.3 Example 2: Universities presidents' Compensation

A second real world data example was considered to further compare the performance of the proposed estimator to that of the maximum likelihood estimator. As in the previous example, the findings reveal that our estimator is robust to local contamination of the response and is comparable to the maximum likelihood estimator in the absence of the potential outlier.

The data is about the compensation received in the 2010-2011 fiscal year by 199 chief executives at 190 public universities and systems in the United States. *The Chronicle* surveyed institutions to collect compensation data.It includes public colleges and their affiliated systems that were classified as research universities by the Carnegie Foundation for the Advancement of Teaching in 2010. The four-year institutions included here comprise universities with total fall enrollments of at least 10,000 and universities with smaller enrollments that are state flagships. At some colleges, more than one president served during the year 2010-2011. All people who served in the capacity of chief executive were used in the study, including interim leaders if they served for at least six months.

In this work, we use the data from 49 states flagship universities included in the study. We divided the 49 states flagship universities into four geographic region, mainly the South, the West, the Northeast, and the Midwest. The purpose of the analysis conducted in this work is to estimate and compare the compensation of states' flagship universities by geographic region. However, the president of the Ohio State University had a compensation of 1,893,911, where as the rest of the presidents had their compensation ranging between 126,340 to 944,697. That is, the data contained a potential outlier. Such a data point when included in the study requires the use of a robust method like the one developed in this work. Traditionally, income data are modeled with a gamma distribution and the same distribution is adopted in this work. We fit a generalized linear model with a gamma response and *log* link function. The predictor is the *geographic region* variable with four levels: the South,

the West, the Northeast, and the Midwest. This gives rise to

$$\log(E[\mathbf{y}_i]) = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3, \qquad i = 1, 2, \cdots, 49 \ ,$$

where

1. y = President's Compensation (US dollars)
2. $x_1 = 1$ if "region = South", $x_1 = 0$ "otherwise"
3. $x_2 = 1$ if "region = West", $x_2 = 0$ "otherwise"
4. $x_3 = 1$ if "region = Midwest", $x_3 = 0$ "otherwise".

We computed 1000 bootstrap estimates of the coefficients with both the MLE and the proposed method, with and without the potential outlier. The (componentwise) mean and standard deviation of these 1000 estimates are reported in table 3.2. In addition we report the sum of the squared deviance residuals (sqr.Dev) and the sum of the absolute deviance residuals (abs.Dev). We also perform a test of hypotheses to determine whether the compensations of the universities's presidents in the Midwest differ from the ones of Northeastern universities's presidents. This can be expressed as

$$H_0 : \beta_3 = 0 \quad versus \quad H_a : \beta_3 \neq 0.$$

Table 3.3 provides the p-values of both the proposed method and the MLE. Although both the MLE and the proposed rank procedure fail to reject the null hypothesis, the MLE P-values are more affected by the outlier than the rank based ones.

The Pearson residual plots and the deviance residuals plots are respectively given by figures 3.7 and 3.8. We observe that in the presence of the outlier, the proposed rank method has a smaller sum of absolute deviance residuals than the MLE. Whereas this sum of absolute deviance residuals is nearly the same for both methods when this potential outlier

Table 3.2: Coefficient (Coef) and Standard Error (SE) estimates for universities presidents's compensation

	Coef - Rank		Coef - MLE		SE - Rank		SE - MLE	
	Clean	Original	Clean	Original	Clean	Original	Clean	Original
inte	12.857	12.844	12.953	12.952	0.099	0.143	0.152	0.150
x2	0.249	0.252	0.173	0.176	0.163	0.154	0.173	0.164
x3	0.081	0.083	0.106	0.105	0.174	0.177	0.190	0.191
x4	-0.016	0.045	-0.032	0.201	0.189	0.203	0.204	0.282
sqr. Dev					6.973	13.412	5.689	12.846
abs. Dev					13.991	17.770	13.222	19.064

Table 3.3: P-values Comparison

	Rank	MLE
Clean	pvalue=0.9325	pvalue=0.8753
Original	pvalue=0.8245	pvalue=0.4759

is removed from the data. The MLE has a smaller sum of squared deviance residuals either with or without the potential outlier in the data. This is expected as the MLE minimizes the sum of squared deviance residuals. These observations are illustrated in the residuals plots. Although the effect of the potential outlier is not remarkably significant, the outlier does affect the MLE estimation as revealed by the deviance.

Figure 3.7: Pearson residuals plot for clean and contaminated response

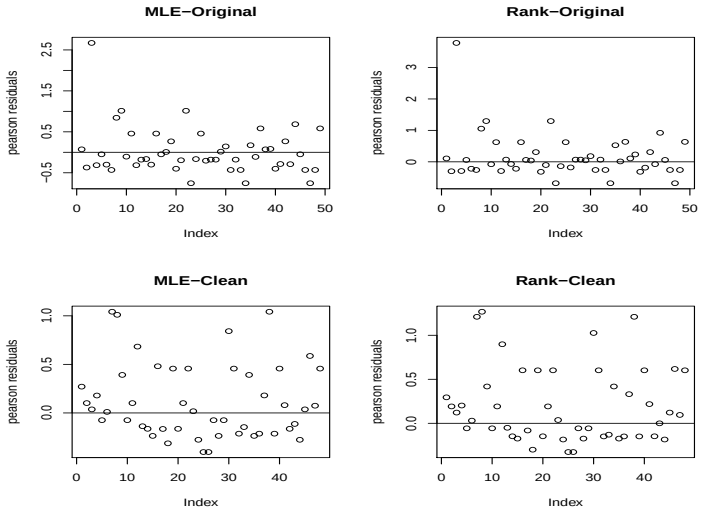
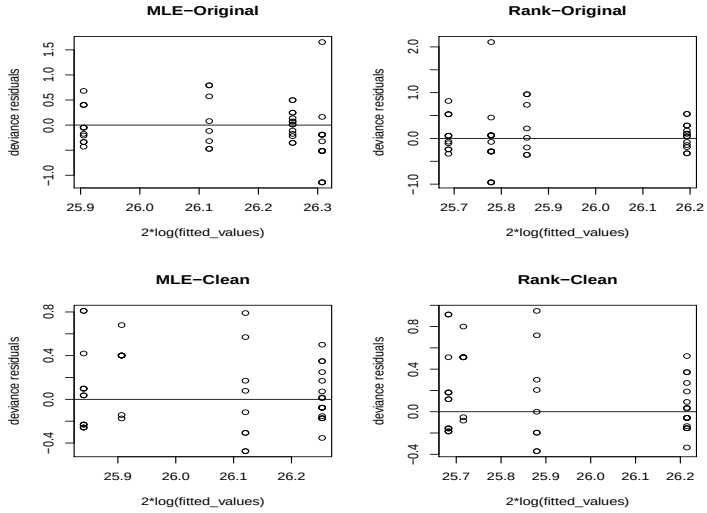


Figure 3.8: Deviance residuals vs Fitted values for clean and contaminated response



3.5 Conclusion

A k-step rank based estimator for generalized linear model has been developed, with its performance evaluated in comparison with the MLE for both simulated and real data. Our estimation procedure produces its initial estimator and iteratively updates it through minimization of a rank based objective function, yielding estimators that have bounded influence in the response space. As such, the method developed in this work is ideal for data from designed experiments where the x's are controlled. There is no guarantee that our procedure results in robust estimates for uncontrolled studies. It is interesting to extend our procedure to the discrete response case.

3.6 Proofs

Proof of Theorem 3.1. Let $\omega_i(\boldsymbol{\theta}) = \Big(\frac{r_i(\boldsymbol{\theta})}{n+1} - \frac{1}{2}\Big)\mu'(\mathbf{x}_i^t\boldsymbol{\theta})$. Observe that by C_1, $\inf_{i\in\mathbb{N}} \mu'(\mathbf{x}_i^t\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. On the other hand, $\delta_i(\boldsymbol{\theta}) = \Big(\frac{r_i(\boldsymbol{\theta})}{n+1} - \frac{1}{2}\Big)\mu(\mathbf{x}_i^t\boldsymbol{\theta})$, is such that $|\delta_i(\boldsymbol{\theta})| \leq K$ for all $i \in \mathbb{N}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, for some constant K since $\mu(.)$ is continuous and $U = \{u \mid u = \mathbf{x}^t\boldsymbol{\theta}$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\mathbf{x} \in \mathbb{X}\}$ is compact as a consequence of both $\boldsymbol{\Theta}$ and $\mathbb{X}$ being compact and also $\Big|\frac{r_i(\boldsymbol{\theta})}{n+1} - \frac{1}{2}\Big| \leq \frac{1}{2}$.
Now, by definition of $\widehat{\boldsymbol{\theta}}_n$

$$\sum_{i=1}^{n} \delta_i(\widehat{\boldsymbol{\theta}}_n)\mathbf{x}_i = 0.$$

By the mean value theorem, there exists a random vector, $\tilde{\boldsymbol{\theta}}_n$, on the line segment between $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\theta}}_n$ such that

$$0 = \sum_{i=1}^{n} \delta_i(\boldsymbol{\theta}_0)\mathbf{x}_i + \sum_{i=1}^{n} \omega_i(\tilde{\boldsymbol{\theta}}_n)\mathbf{x}_i\mathbf{x}_i^t(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0).$$

So, for large n

$$(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \Big(\sum_{i=1}^{n} \omega_i(\tilde{\boldsymbol{\theta}}_n)\mathbf{x}_i\mathbf{x}_i^t\Big)^{-1} \sum_{i=1}^{n} \delta_i(\boldsymbol{\theta}_0)\mathbf{x}_i.$$

Note that, in addition to, the fact that $\Big|\frac{r_i(\boldsymbol{\theta})}{n+1} - \frac{1}{2}\Big| \leq \frac{1}{2}$ $\forall\ i \in \{1, 2, \cdots, n\}$ and $n \in \mathbb{N}$; $\frac{r_i(\boldsymbol{\theta})}{n+1} - \frac{1}{2}$ is equal to zero, only at a single point.That is, for large n, the random variable $\frac{r_i(\boldsymbol{\theta})}{n+1} - \frac{1}{2}$ is only equal to zero on a set of measure zero. Therefore,

$$(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = O(1)\Big(\sum_{i=1}^{n} \mu'(\mathbf{x}_i^t\tilde{\boldsymbol{\theta}}_n)\mathbf{x}_i\mathbf{x}_i^t\Big)^{-1} \sum_{i=1}^{n} \delta_i(\boldsymbol{\theta}_0)\mathbf{x}_i \quad a.s.$$

which in turn implies

$$\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \leq K \left\|\Big(\sum_{i=1}^{n} \mu'(\mathbf{x}_i^t\tilde{\boldsymbol{\theta}}_n)\mathbf{x}_i\mathbf{x}_i^t\Big)^{-1}\right\| \sum_{i=1}^{n} \|\mathbf{x}_i\| \quad a.s$$

and so

$$\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \leq K\lambda_{min}^{-1}\Big(\sum_{i=1}^{n}\mu'(\mathbf{x}_i^t\tilde{\boldsymbol{\theta}}_n)\mathbf{x}_i\mathbf{x}_i^t\Big)\sum_{i=1}^{n}\|\mathbf{x}_i\| \quad a.s\ .$$

So by Lemma 3.1, we have

$$\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \leq K\Big(\inf_{i\in\mathbb{N}}\mu'(\mathbf{x}_i^t\tilde{\boldsymbol{\theta}}_n)\Big)^{-1}\lambda_{min}^{-1}\Big(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^t\Big)\sum_{i=1}^{n}\|\mathbf{x}_i\| \quad a.s.$$

Note that $\sum_{i=1}^{n-1}\|\mathbf{x}_i\| = O\Big(\lambda_{max}^{1/2}\Big(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^t\Big)\Big)$ since all norms are equivalent on finite dimensional vector spaces. In addition, $\Big(\inf_{i\in\mathbb{N}}\mu'(\mathbf{x}_i^t\tilde{\boldsymbol{\theta}}_n)\Big)^{-1} < \infty$ since $\Big(\inf_{i\in\mathbb{N}}\mu'(\mathbf{x}_i^t\tilde{\boldsymbol{\theta}}_n)\Big) > 0$. Therefore,

$$\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \leq O(1)\lambda_{min}^{-1}\Big(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^t\Big)\lambda_{max}^{1/2}\Big(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^t\Big) \quad a.s;$$

and by C_3

$$\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \leq O(1)\lambda_{min}^{-1/2}\Big(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^t\Big) \quad a.s;$$

which converges to 0 a.s by C_2, and the proof is complete.

□

Proof of Lemma 3.2. First, observe that

$$\begin{aligned}
E\Big(\gamma_n(\mathbf{x}_i,\boldsymbol{\theta}_0)\Big(\frac{r_i(\boldsymbol{\theta}_0)}{n+1}-\frac{1}{2}\Big)\Big) &= E\Big(E\Big(\gamma_n(\mathbf{x}_i,\boldsymbol{\theta}_0)\Big(\frac{r_i(\boldsymbol{\theta}_0)}{n+1}-\frac{1}{2}\Big)|\mathbf{x}_i\Big)\Big) && (3.6.1)\\
&= E\Big(\gamma_n(\mathbf{x}_i,\boldsymbol{\theta}_0)E\Big(\Big(\frac{r_i(\boldsymbol{\theta}_0)}{n+1}-\frac{1}{2}\Big)\Big)\Big) && (3.6.2)\\
&= E(0) && (3.6.3)\\
&= 0 && (3.6.4)
\end{aligned}$$

which implies that

$$E\Big(L_n(\boldsymbol{\theta}_0,\tau)\Big) = 0$$

So, by corollaries 3.4 and 3.6 of Brunner and Denker (1994)

$$\sqrt{n}L_n(\boldsymbol{\theta}_0,\tau)=\frac{\tau^t\sqrt{n}\Psi_n(\boldsymbol{\theta}_0)}{max_{1\le i\le n}\|\tau^t\mathbf{x}_i\mu(\boldsymbol{\theta}_0^t\mathbf{x}_i)\|}\qquad \text{has an asymptotic } N\Big(0,\frac{s_n^2(\boldsymbol{\theta}_0)}{n}\Big) \text{ distribution.}$$

That is

$$\frac{\tau^t\sqrt{n}\Psi_n(\boldsymbol{\theta}_0)}{max_{1\le i\le n}\|\tau^t\mathbf{x}_i\mu(\boldsymbol{\theta}_0^t\mathbf{x}_i)\|}=N\Big(0\ ,\ \frac{n\tau^tE(\Psi_n(\boldsymbol{\theta}_0)\Psi_n^t(\boldsymbol{\theta}_0))\tau}{(max_{1\le i\le n}\|\tau^t\mathbf{x}_ig'(\boldsymbol{\theta}_0^t\mathbf{x}_i)\|)^2}\Big)+o_p(1)\ .$$

Note that for each $\boldsymbol{\theta}\in\boldsymbol{\Theta}$, $max_{1\le i\le n}\|\tau^t\mathbf{x}_i\mu(\mathbf{x}_i^t\boldsymbol{\theta})\|\neq 0$; otherwise $\Psi_n(\boldsymbol{\theta})=0$ on $\boldsymbol{\Theta}$. Now, there is some finite constant K such that $\|\tau^t\mathbf{x}_i\mu(\boldsymbol{\theta}_0^t\mathbf{x}_i)\|\le K$ for all i since μ is continuous and $U=\{u\mid u=\mathbf{x}^t\boldsymbol{\theta}\ for\ \boldsymbol{\theta}\in\boldsymbol{\Theta}\ and\ \mathbf{x}\in\mathbb{X}^p\}$ is compact as a consequence of both $\boldsymbol{\Theta}$ and $\mathbb{X}$ being compact. So,

$$\tau^t\sqrt{n}\Psi_n(\boldsymbol{\theta}_0)=N\left(0\ ,\ n\tau^tE(\Psi_n(\boldsymbol{\theta}_0)\Psi_n^t(\boldsymbol{\theta}_0))\tau\right)+o_p(1)\ .$$

Applying the Cramér-Wold device and taking $\lim_{n\to\infty}$ we get

$$\sqrt{n}\Psi_n(\boldsymbol{\theta}_0)\xrightarrow{\mathcal{D}}N(0,\Omega_0)\ .$$

□

Proof of Theorem 3.2. The technique used in this proof, Taylor expansion of $\Psi_n(\boldsymbol{\theta})$ around $\boldsymbol{\theta}_0$, is similar to the one used in establishing the asymptotic normality of M-estimators. Note that $\Psi_n(\boldsymbol{\theta})$ and all its subsequent derivatives are defined on the interior of $\boldsymbol{\Theta}$ which is open and convex. Since $\widehat{\boldsymbol{\theta}}_n$ is consistent for $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_0$ is in the interior of $\boldsymbol{\Theta}$, this implies that the sequence $\widehat{\boldsymbol{\theta}}_n$ will eventually be in the interior of $\boldsymbol{\Theta}$ as well.

Let $\alpha=(\alpha_1,\cdots,\alpha_p)\in\mathbb{N}_0^n$, $\mathbb{N}_0=\mathbb{N}\cup\{0\}$, be a multi-index. We define the differential operator

$$D_{\boldsymbol{\theta}}^{\alpha}=\frac{\partial^{|\alpha|}}{\partial\boldsymbol{\theta}_1^{\alpha_1}\cdots\partial\boldsymbol{\theta}_p^{\alpha_p}},$$

where $|\alpha| = \sum_{i=1}^{p} \alpha_i$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_p)$. For any function $\psi : \boldsymbol{\Theta} \to \mathbb{R}$, we will take

$$\dot{\psi}(\boldsymbol{\theta}) = D^{\alpha}_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}) \quad \text{for all } \alpha \text{ such that } |\alpha| = 1.$$

By Taylor's theorem, there exists a random vector $\tilde{\boldsymbol{\theta}}_n$ on the line segment between $\widehat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_0$ for which

$$0 = \Psi_n(\widehat{\boldsymbol{\theta}}_n) = \Psi_n(\boldsymbol{\theta}_0) + \dot{\Psi}_n(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^t \left\{ \frac{1}{\alpha!} D^{\alpha}_{\boldsymbol{\theta}} \Psi_n(\tilde{\boldsymbol{\theta}}_n)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^{\delta} \right\} (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$$

that can be rearranged as

$$-\left\{ (\dot{\Psi}_n(\boldsymbol{\theta}_0) + (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^t \Big(\frac{1}{\alpha!} D^{\alpha}_{\boldsymbol{\theta}} \Psi_n(\tilde{\boldsymbol{\theta}}_n)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^{\delta} \Big) \right\} \sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \sqrt{n}\Psi_n(\boldsymbol{\theta}_0) \ , \quad (3.6.5)$$

where $\alpha! = \Pi_{i=1}^{p} \alpha_i!$ for all α such that $|\alpha| = 2$ and $(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^{\delta} = \Pi_{i=1}^{p} (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)_i^{\delta^i}$ with $\delta = (\delta^i, \cdots, \delta^n)$ and $\sum_{i=1}^{n} \delta^i = 1$.

Now, for each value of α with $|\alpha| = 2$, we have

$$\left\| \frac{1}{\alpha!} D^{\alpha}_{\boldsymbol{\theta}} \Psi_n(\tilde{\boldsymbol{\theta}}_n) \right\| \leq \frac{L}{n} \sum_{i=1}^{n} \Big| \frac{r_i(\boldsymbol{\theta})}{n+1} - \frac{1}{2} \Big| \quad \text{for some } L \text{ (independent of } n), \text{ with } 0 \leq L < \infty \ ,$$

since μ'' is continuous and $U = \{u \mid u = \mathbf{x}^t\boldsymbol{\theta} \text{ for } \boldsymbol{\theta} \in \boldsymbol{\Theta} \text{ and } \mathbf{x} \in \mathbb{X}\}$ is compact as a consequence of both $\boldsymbol{\Theta}$ and $\mathbb{X}$ being compact. Moreover, we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Big| \frac{r_i(\boldsymbol{\theta})}{n+1} - \frac{1}{2} \Big| = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Big| \frac{i}{n+1} - \frac{1}{2} \Big| = \int_0^1 \Big| u - \frac{1}{2} \Big| du$$

that is bounded (see Hettmansperger, 1984, page 307 Definition A4). Now since $\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 = o_p(1)$, equation (3.6.5) can be written as

$$-\Big(\dot{\Psi}_n(\boldsymbol{\theta}_0) + o_p(1) \Big) \sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \sqrt{n}\Psi_n(\boldsymbol{\theta}_0) \quad (3.6.6)$$

which by Lemma 3.2 implies

$$\dot{\Psi}_n(\boldsymbol{\theta}_0)\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N(0, \Omega_0)$$

Since $\dot{\Psi}_n(\boldsymbol{\theta}_0) = H_0 + o_p(1)$, where H_0 is invertible by N_2, we have

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N(0, H_0^{-1}\Omega_0 H_0^{-1}) \ .$$

□

Proof of Theorem 3.3. The consistency of $\widehat{\boldsymbol{\theta}}_n$ together with equation (3.6.6) imply that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \sqrt{n}\Big(\dot{\Psi}_n(\boldsymbol{\theta}_0)\Big)^{-1}\Psi_n(\boldsymbol{\theta}_0) + o_p(1)$$

that is

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i\mu(\mathbf{x}_i^t\boldsymbol{\theta}_0)\Big(\frac{r_i(\boldsymbol{\theta}_0)}{n+1} - \frac{1}{2}\Big)\Big(\dot{\Psi}_n(\boldsymbol{\theta}_0)\Big)^{-1} + o_p(1). \tag{3.6.7}$$

Note that, by definition $\dfrac{r_i(\boldsymbol{\theta}_0)}{n+1} = \dfrac{1}{n+1}\sum_{j=1}^{n} I(e_j(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) \le e_i(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0))$; where I is the indicator function. By the uniform law of large numbers

$$\frac{1}{n+1}\Big|\sum_{j=1}^{n} I(e_j(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) \le e_i(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)) - \sum_{j=1}^{n} F_j(e_i(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0))\Big| \xrightarrow{\mathcal{P}} 0$$

where $e_j(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) \sim F_j$.

Therefore, for large n, equation 3.6.7 can be rewritten as

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i\mu(\mathbf{x}_i^t\boldsymbol{\theta}_0)\Big(\bar{F}_n(e_i(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)) - \frac{1}{2}\Big)\Big(\dot{\Psi}_n(\boldsymbol{\theta}_0)\Big)^{-1} + o_p(1). \tag{3.6.8}$$

with $\bar{F}_n = \dfrac{1}{n}\sum_{j=1}^{n} F_j$.

From this representation, the influence function of $\widehat{\boldsymbol{\theta}}_n$ is the limit as $n \to \infty$ of the function under the summation (Hettmansperger and McKean, 1998, Corollary 3.5.7). This function can be written as

$$\mathrm{IF}_n(\mathbf{x}, y) = \mathbf{x}\mu(\mathbf{x}^t\boldsymbol{\theta}_0)\Big(\bar{F}_n(e_i(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)) - .5\Big)H_0^{-1}\ ,$$

Therefore the influence function is

$$\mathrm{IF}(\widehat{\boldsymbol{\theta}}_n; \mathbf{x}, y) = \mathbf{x}\mu(\mathbf{x}^t\boldsymbol{\theta}_0)\Big(\bar{F}\Big(\frac{y - g(\mathbf{x}^t\boldsymbol{\theta}_0)}{\sqrt{\nu(g(\mathbf{x}^t\boldsymbol{\theta}_0))}}\Big) - .5\Big)H_0^{-1}$$

where $\bar{F} = \lim_{n\to\infty}\bar{F}_n$. Note that the assumption $F_j(t) > 0\ \forall j \in \mathrm{N}$ and $t \in \mathbf{R}$, guarantees that for each $t \in \mathbf{R}$, $\bar{F}_n(t)$ is a Cauchy sequence of real numbers and therefore converges.

Observe that $0 < \bar{F} < 1$ since each F_j is such that $0 < F_j < 1$. As a consequence, the influence function of $\widehat{\boldsymbol{\theta}}_n$, $\mathrm{IF}(\widehat{\boldsymbol{\theta}}_n; \mathbf{x}, y)$,is a bounded function of y. From this argument, it follows that the influence function of $\widehat{\boldsymbol{\theta}}_n^0$, $\mathrm{IF}(\widehat{\boldsymbol{\theta}}_n^0; \mathbf{x}, y)$, is bounded in the y direction; by setting $\nu(g(t)) \equiv 1$, since $\widehat{\boldsymbol{\theta}}_n \equiv \widehat{\boldsymbol{\theta}}_n^0$ in this case.

□

Proof of Theorem 3.4. Recall that $\Psi_n(\widehat{\boldsymbol{\theta}}_n^k) = \Psi_n^{k+1}(\widehat{\boldsymbol{\theta}}_n^k)$ and $\Psi_n^{k+1}(\widehat{\boldsymbol{\theta}}_n^{k+1}) = \mathbf{0}$. Now expanding $\Psi_n^{k+1}(\widehat{\boldsymbol{\theta}}_n^{k+1})$ about $\widehat{\boldsymbol{\theta}}_n^k$ gives

$$\widehat{\boldsymbol{\theta}}_n^{k+1} = \widehat{\boldsymbol{\theta}}_n^k + \Big[\Gamma_n(\widehat{\boldsymbol{\theta}}_n^k)\Big]^{-1}\Psi_n(\widehat{\boldsymbol{\theta}}_n^k) \qquad k = 1, 2, 3, \cdots$$

where Γ_n is the gradient of Ψ_n^{k+1} which is a function of $\widehat{\boldsymbol{\theta}}_n^k$ alone. Define the function $T : \boldsymbol{\Theta} \to \boldsymbol{\Theta}$ by $T(\boldsymbol{\theta}) = \boldsymbol{\theta} + [\Gamma_n(\boldsymbol{\theta})]^{-1}\Psi_n(\boldsymbol{\theta})$. Note that T is continuous. Since $\boldsymbol{\Theta}$ is compact

and convex, by the Schauder fixed point theorem, T has a fixed point. That is, $\exists\ \boldsymbol{\theta}_n^* \in \boldsymbol{\Theta}$ such that $T(\boldsymbol{\theta}_n^*) = \boldsymbol{\theta}_n^*$.

Therefore, $\widehat{\boldsymbol{\theta}}_n^{k+1} = T(\widehat{\boldsymbol{\theta}}_n^k)$ implies that

$$\lim_{k\to\infty} \widehat{\boldsymbol{\theta}}_n^k = \boldsymbol{\theta}_n^* \quad \text{for each } n\ . \tag{3.6.9}$$

If $[\Gamma_n(\boldsymbol{\theta})]^{-1} = \mathbf{0}$, then $T(\boldsymbol{\theta}) = \boldsymbol{\theta}$, in which case $\widehat{\boldsymbol{\theta}}_n^k = \widehat{\boldsymbol{\theta}}_n^0$ for all k. So, the sequence trivially converges. So, assume $[\Gamma_n(\boldsymbol{\theta})]^{-1} \neq \mathbf{0}$. Then, by the definition of T, $T(\boldsymbol{\theta}_n^*) = \boldsymbol{\theta}_n^* \ \Rightarrow\ \Psi_n(\boldsymbol{\theta}_n^*) = \mathbf{0}$. □

Chapter 4

Rank Based Estimation for Longitudinal Data Analysis

We extend the method developed in the previous chapter to longitudinal data. The same notation will be used for corresponding expressions. In addition, identical arguments or proof of theorems will be omitted in this chapter. We will, instead, focus on the specificities pertaining to longitudinal data and how the method developed in the previous chapter are extended to this type of data. All assumptions made in the previous chapter apply here.

4.1 Introduction

Longitudinal studies describe the relationship between a response variable and some covariates when the observations made on the response are repeated over a certain period of time, or in space. Such studies commonly arise in various fields of science, including medicine, psychology, sociology and economics. Even though the measurements for different subjects can be considered independent, this is not the case for repeated measurements on the same subjects, and this within-cluster correlation must be taken into account. Over the years, extensive research has been conducted on statistical methods for inferences in longitudinal data analysis.Liang and Zeger (1986) developed the approach of generalized estimating equations (GEE), which involves a "working"correlation matrix to improve estimation efficiency. This approach only requires specification of marginal mean and covariance functions.The theory stems from constructing optimal linear combinations of pearson residuals for parameters estimation.

However, this approach and many other related methods are vulnerable to outliers in the data. Recently, several authors have considered robust methods for longitudinal data analysis. For example,He et al. (2002) proposed M-estimators in partly linear models. Qaqish

and Preisser (1999) developed a resistant version of the GEE method by down-weighting influential data points.Huggins (1993) and Gill(2000) also applied the robust approach to the repeated measures based on multivariate normal distributions. Welsh and Richardson (1997) investigated multivariate t-distributions and truncated normal distributions.Linear transformation of the Pearson residuals can result in uncorrelated residuals so that traditional M-estimation may be used. However for these approaches. symmetry of the joint distribution , a rather strong assumption, is required. Hu and Lachin (2001)proposed to robustify the GEE approach by applying the Huber function to the standardized residuals. This approach is appropriate only when error distributions are symmetric. This assumption is also required by many others, including Schrader and Hettmansperger (1980) and Gill (2000). Cantoni (2004) also developed a robust approach to longitudinal data analysis based on weighted quasi-likelihood functions. Jung and Ying (2003) explored rank methods for repeated measures in linear models. Their approach assumes that the pairwise differences of errors have symmetric distributions. In this chapter we use rank procedures in the context of longitudinal data analysis described in Liang and Zeger (1986). Unlike Jung and Ying (2003), we do not assume a linear relationship between the response and the predictor. Instead, we model the relationship between the mean response and the covariates through a more general link function. Jung and Ying (2003) studied the linear model with correlated and non-i.i.d errors. The proposed method is developed for correlated and non-i.i.d errors and allows for arbitrary link functions with minimal assumption on the error distribution.

The remainder of this chapter is organized as follows. After introducing the model and the estimators in Section 4.2, the consistency and the asymptotic normality of the rank version of the maximum quasi-likelihood estimator are studied in Section 4.3. We illustrate the robustness and the efficiency of the estimators in Section 4.4 via simulation studies and a real world data example. Section 4.5 provides the conclusion.

4.2 The Model and Estimator

Our notations follows that of Liang and Zeger (1986). Consider a longitudinal set of observations over n subjects. Let y_{ij} denote the jth response for the ith subject for $j = 1, 2, \cdots, m_i$ and $i = 1, 2, \cdots, n$. Assume that $\mathbf{x}_{ij}$ is a p by 1 vector of corresponding covariates. Let $N = \sum_{i=1}^{n} m_i$ denote the total sample size. Assume that the random vectors $(y_{i1}, \cdots, y_{im_i})$ $1 \leq i \leq n$ are independent, the marginal distribution of y_{ij} is absolutely continuous and from the exponential class of distribution and that the mean of y_{ij} is related to $\mathbf{x}_{ij}$ by

$$h[E(y_{ij}|\mathbf{x}_{ij} = \mathbf{x})] = \mathbf{x}^t\boldsymbol{\theta}_0, \tag{4.2.1}$$

where $\boldsymbol{\theta}_0 \in \Theta \subset \mathbb{R}^p$ is a vector of parameters. We assume, like in the previous chapter, that $\mathbf{x}_{ij} \in \mathbb{X} \subset \mathbb{R}^p, 1 \leq i \leq n$, are independent random vectors for each fixed $j : 1 \leq j \leq m_i$. The function h is such that its inverse $g \equiv h^{-1}$ is a real valued function defined on the set $U = \{u \mid u = \mathbf{x}^t\boldsymbol{\theta}$ for $\boldsymbol{\theta} \in \Theta$ and $\mathbf{x} \in \mathbb{X}\} \subset \mathbb{R}$, monotone and three times continuously differentiable. We shall assume that, $\mathbb{X}$ is compact, Θ is convex and compact, and $\boldsymbol{\theta}_0$ is an interior point of Θ. We will also assume, like in Denker (1994) and Jung and Ying (2003), that $max_{1\leq i\leq n} m_i \leq m < \infty$, where m is independent of n. This condition, in part implies that, $N \to \infty$ if and only if $n \to \infty$.

We consider the rank quasi-likelihood for longitudinal data ,as the estimator of $\boldsymbol{\theta}_0$. That is the estimator of $\boldsymbol{\theta}_0$, $\widehat{\boldsymbol{\theta}}_n$ is such that

$$\Psi_n(\widehat{\boldsymbol{\theta}}_n) = 0 \ , \tag{4.2.2}$$

where

$$\Psi_n(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left[\frac{r_{ij}(\boldsymbol{\theta}, \boldsymbol{\theta})}{N+1} - \frac{1}{2}\right]\frac{g'(\mathbf{x}_{ij}^t\boldsymbol{\theta})}{\sqrt{\nu(g(\mathbf{x}_{ij}^t\boldsymbol{\theta}))}}\mathbf{x}_{ij} \ . \tag{4.2.3}$$

where $r_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} I(e_{i'j'}(u) \leq e_{ij}(u))$, as defined in Jung and Ying (2003), is the rank of $e_{ij}(\boldsymbol{\theta},\boldsymbol{\theta})$ among $e_{11}(\boldsymbol{\theta},\boldsymbol{\theta}),\ldots,e_{1m_1}(\boldsymbol{\theta},\boldsymbol{\theta})\ldots,e_{n1}(\boldsymbol{\theta},\boldsymbol{\theta}),\ldots,e_{nm_n}(\boldsymbol{\theta},\boldsymbol{\theta})$, with

$$e_{ij}(\boldsymbol{\theta},\boldsymbol{\theta}) = \frac{y_{ij} - g(\mathbf{x}_{ij}^t\boldsymbol{\theta})}{\sqrt{\nu(g(\mathbf{x}_{ij}^t\boldsymbol{\theta}))}}\ , \quad j = 1,\ldots,m_i \quad \textit{and} \quad i = 1,\ldots,n\ .$$

Note that the above defined rank quasi-likelihood estimator $\widehat{\boldsymbol{\theta}}_n$ reduces to the one defined in the previous chapter in the special case of $m_i = 1\ \forall\ i \in 1,\cdots,n$.

As shown in the previous chapter, for each n, the estimator $\widehat{\boldsymbol{\theta}}_n$ is the limit of the sequence $\widehat{\boldsymbol{\theta}}_n^k$ defined as:

$$\widehat{\boldsymbol{\theta}}_n^k = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\text{Argmin}}\, W_n(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_n^{k-1})$$

where

$$W_n(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_n^k) = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left[\frac{r_{ij}(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_n^k)}{N+1} - \frac{1}{2}\right] e_{ij}(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_n^k)\ .$$

The initial estimator $\widehat{\boldsymbol{\theta}}_n^0$ is chosen similarly to the one in the previous chapter.

Once again, the existence of $\widehat{\boldsymbol{\theta}}_n^k$ and $\widehat{\boldsymbol{\theta}}_n$ are justified by Lemma 2 of Jennrich (1969).The argument of the convergence of $\widehat{\boldsymbol{\theta}}_n^k$ to $\widehat{\boldsymbol{\theta}}_n$ as $k \to \infty$ and n fixed will not be included in this chapter as it is identical to the one developed in the previous chapter.

For simplicity of notation, $r_{ij}(\boldsymbol{\theta},\boldsymbol{\theta})$ will be denoted by $r_{ij}(\boldsymbol{\theta})$.

Jung and Ying (2003) studied the linear model in the case of $\{e_{ij} \quad i = 1,\ldots,n \quad ; \quad j = 1,\ldots,m_i\}$ that are such that $\{e_{ij}, \quad j = 1,\ldots,m_i\}$ are dependent for each i and $\{e_{1j}, \quad j = 1,\ldots,m_1\}\ \ldots\{e_{nj}, \quad j = 1,\ldots,m_n\}$ are independent random processes. Therefore, for the case $g(\mathbf{x}_{ij}^t\boldsymbol{\theta}) = \mathbf{x}_{ij}^t\boldsymbol{\theta}$ and $\nu(.) \equiv 1$, the estimator proposed in this work is identical to the one defined by Jung and Ying (2003).

4.3 Asymptotic Properties

The expressions used in this sections are defined similarly to the corresponding quantities in chapter 3. Our intent will be on the ingredients necessary to extend the proof of the

previous chapter to longitudinal data. Note that most of the assumptions made below are similar to the ones in Chapter 3.

4.3.1 Consistency

In this subsection we will present sufficient conditions for strong consistency of the initial estimator. Denote $Y_i = (y_{i1}, \cdots, y_{im_i})$,$\mathbf{X}_i = (\mathbf{x}_{i1}^t, \cdots, \mathbf{x}_{im_i}^t)$ and let (Ω, F, P) be a probability space. Assume, for $i = 1, \cdots, n$, the random vectors $(Y_i, \mathbf{X}_i)$, are independent and that Y_i and $\mathbf{X}_i$ are each carried by (Ω, F, P) for all $i = 1, \cdots, n$.

The first ingredient necessary to extend the proofs of Chapter 3 to longitudinal data is to introduce the following notation. The jth observation of the ith subject can be rewritten as:

$$\mathbf{x}_{ij} \equiv \mathbf{x}_q; \quad q = j + \sum_{l=0}^{i-1} m_l \quad for \quad 1 \leq j \leq m_i; \quad with\ m_0 = 0 \quad and\ 1 \leq i \leq n.$$

The response y_q, the ranks r_q and the residuals e_q are defined via similar vectorizations. Observe that the resulting predictors $\mathbf{x}_q$, responses y_q, and residuals e_q are not independent. Instead they are clusters of independent random processes (vectors).

Similarly to the previous chapter,to establish the strong consistency of $\widehat{\boldsymbol{\theta}}_n$, assumptions C_1, C_2, and C_3 will be needed. Mainly

C_1 : $\inf\limits_{q \in \mathbb{N}} \ \mu'(\mathbf{x}_q^t \boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

C_2 : $\lambda_{min}(\sum_{q=1}^{N} \mathbf{x}_q \mathbf{x}_q^t) \to \infty \ a.s$ as $n \to \infty$, and

C_3 : there exist finite constants $n_0 > 0$ and $c > 0$ such that $\dfrac{\lambda_{max}(\sum_{q=1}^{N} \mathbf{x}_q \mathbf{x}_q^t)}{\lambda_{min}(\sum_{q=1}^{N} \mathbf{x}_q \mathbf{x}_q^t)} < c$ for all $n \geq n_0$.

The following theorem gives the consistency of $\widehat{\boldsymbol{\theta}}_n$. The consistency of $\widehat{\boldsymbol{\theta}}_n^0$ follows as a special case by taking $\nu(g(t)) \equiv 1$ in the proof.

Theorem 4.1. *Under* C_1 , C_2, *and* C_3, $\widehat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$ *a.s when* $n \to \infty$.
In fact, $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = O\Big(\lambda_{min}^{-1/2}\Big(\sum_{q=1}^{N} \mathbf{x}_q \mathbf{x}_q^t\Big)\Big)$.

Under the assumption $max_{1\leq i\leq n} m_i \leq m < \infty$, the proof of Theorem 4.1 is identical to the one in Chapter 3 and will, therefore, not be discussed here.

4.3.2 Asymptotic Normality and Robustness

We now examine the conditions under which the result on the asymptotic normality of $\widehat{\boldsymbol{\theta}}_n$, and $\widehat{\boldsymbol{\theta}}_n^0$ in particular continues to hold here. We will start by reminding the quantities used to establish this result, in the context of longitudinal data. Let $\tau \in \mathbb{R}^p$ and define

$$L_n(\boldsymbol{\theta}, \tau) = \frac{1}{N} \sum_{q=1}^{N} \gamma_n(\mathbf{x}_q, \boldsymbol{\theta}) \Big(\frac{r_q(\boldsymbol{\theta})}{N+1} - \frac{1}{2} \Big)$$
$$s_n^2(\boldsymbol{\theta}) = N^2 E\Big(L_n^2(\boldsymbol{\theta}, \tau) \Big).$$

Once again, $\Psi_n(\boldsymbol{\theta}_0)$ and $\dot{\Psi}_n(\boldsymbol{\theta}_0)$ respectively denote the gradient and the Hessian of $W_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ evaluated at $\boldsymbol{\theta}_0$;
and

$$\gamma_n(\mathbf{x}_q, \boldsymbol{\theta}) = \frac{\tau^t \mathbf{x}_q \mu'(\mathbf{x}_q^t \boldsymbol{\theta})}{max_{1\leq k\leq N} \|\tau^t \mathbf{x}_q \mu'(\boldsymbol{\theta}^t \mathbf{x}_q)\|} \ .$$

Let $\mathbf{X}$ be the $N \times p$ matrix of regressors, with rows $\mathbf{x}_q$. Similarly to chapter 3, the following assumptions are made

$N_1 :$ $s_n^2(\boldsymbol{\theta}_0) \to \infty$ as $n \to \infty$ and

$N_2 :$ $\dot{\Psi}_n(\boldsymbol{\theta}_0) \xrightarrow{\mathcal{P}} H_0$ invertible

in addition to C_1, C_2, and C_3. Recall the lemma below.

Lemma 4.1. *Under* A_1,

$$\sqrt{N} \Psi_n(\boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N(0, \Omega_0) \quad as \ n \to \infty$$

with $\Omega_0 = \lim_{n\to\infty} NE(\Psi_n(\boldsymbol{\theta}_0)\Psi_n^t(\boldsymbol{\theta}_0))$.

This lemma continues to hold for longitudinal data since, corollaries 3.4 and 3.6 of Brunner and Denker (1994) are still valid for independent random processes. This fact combined with the the assumption $max_{1\leq i\leq n} m_i \leq m < \infty$ are the main ingredients necessary for the result below, the asymptotic normality of $\widehat{\boldsymbol{\theta}}_n$, to continue to hold for longitudinal data. The assumption $max_{1\leq i\leq n} m_i \leq m < \infty$ guarantees that $n \to \infty$ if and only if $N \to \infty$. The proofs of the lemma above and the theorem below remain unchanged and we refer the reader to the previous chapter for details.

We now state the result on the asymptotic normality of $\widehat{\boldsymbol{\theta}}_n$. Once gain, Since taking $\nu(g(t)) \equiv 1$ gives $\widehat{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_n^0$, the asymptotic normality of $\widehat{\boldsymbol{\theta}}_n^0$ follows as a special case of the theorem.

Theorem 4.2. *Under* $C_1 - C_3$ *and* $N_1 - N_2$,

$$\sqrt{N}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N(\mathbf{0}, H_0^{-1}\Omega_0 H_0^{-1}) \quad as\ n \to \infty\ .$$

The result below on the influence function also persists in the case of longitudinal data. The ingredient that allows this result to hold for longitudinal data is the uniform law of large numbers for independent random processes (Pollard (1990)). This in fact implies that

$$\frac{1}{N+1}\Big| r_{ij}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) - \sum_{t=1}^{n}\sum_{l=1}^{m_t} F_{tl}(e_{ij}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0))\Big| \xrightarrow{\mathcal{P}} 0$$

where $e_{ij}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) \sim F_{ij}$.

We recall the statement, of the result on the influence function.

Theorem 4.3. *The influence function of the estimator* $\widehat{\boldsymbol{\theta}}_n$ *is*

$$IF(\widehat{\boldsymbol{\theta}}_n; \mathbf{x}, y) = \mathbf{x}\mu(\mathbf{x}^t\boldsymbol{\theta}_0)\Big(\bar{F}\Big(\frac{y - g(\mathbf{x}^t\boldsymbol{\theta}_0)}{\sqrt{\nu(g(\mathbf{x}^t\boldsymbol{\theta}_0))}}\Big) - .5\Big)H_0^{-1}$$

where $\bar{F} = \lim_{n\to\infty} \bar{F}_n$*;* $\bar{F}_n = \dfrac{1}{n}\sum_{q=1}^{n} F_q$*; with* $e_{ij}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) \sim F_{ij}$.

Once again this influence function is bounded in the response space y, as discussed in Chapter 3.

4.4 Simulations and Examples

4.4.1 Monte Carlo Simulations

To evaluate the performance of the proposed estimator,simulation studies were conducted. Clustered gamma distributed variables were generated as follows:

1. the parameter $\boldsymbol{\theta}_0 = (1/2, 2, 1)$;
2. n=30, 100, 250; $1 \leq i \leq n$
3. the covariates $x = (x_1, x_2, x_3)$ with $x_1 \sim N(0,1)$, $x_2 \sim N(0,1)$, and $x_3 \sim$ Bernoulli$(1/2)$;
4. $g(t) = e^t$, $z_j \sim$ Gamma$[(2/3)g(x^t\boldsymbol{\theta}_0), 3/2]$ $\quad y_q \sim \sum_{j=1}^{l} z_j/k \quad 1 \leq k, j \leq 5$

The vector $(y_1, y_2, \cdots, y_5)$ is therefore a cluster of dependent gamma random variables.

We generated 250 data sets of the kind described above and computed estimates for each and every data set using the proposed method. The process was repeated in the presence an outlier and compared to the corresponding estimates based on the GEE with both the auto-regressive correlation and user defined correlation structure estimation method. The outlier was introduced by replacing the maximum coordinates of the response value Y by its original value multiplied by 20. The boxplots below give a summary of the results.

We observe that our estimator is robust in the response space and perform better compared to the GEE in the presence of the outlier. In the case of the data with no outlier the proposed rank based method is found to be comparable to the GEE. In addition to the robustness, the proposed method does not require estimation of the variance-covariance matrix prior to the estimation of the model parameters, unlike the GEE.

Figure 4.1: Boxplot of the estimates

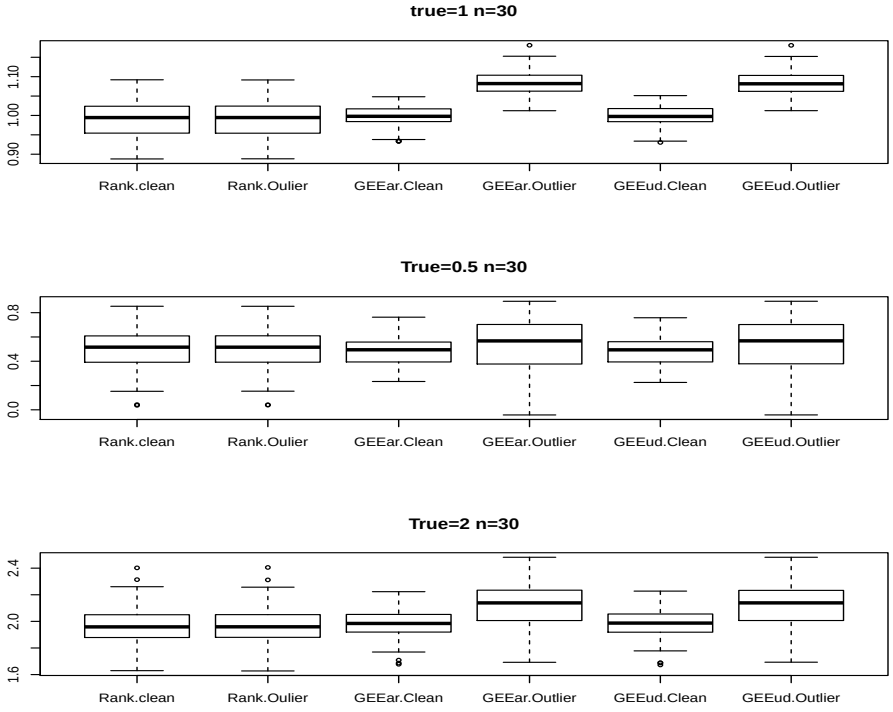
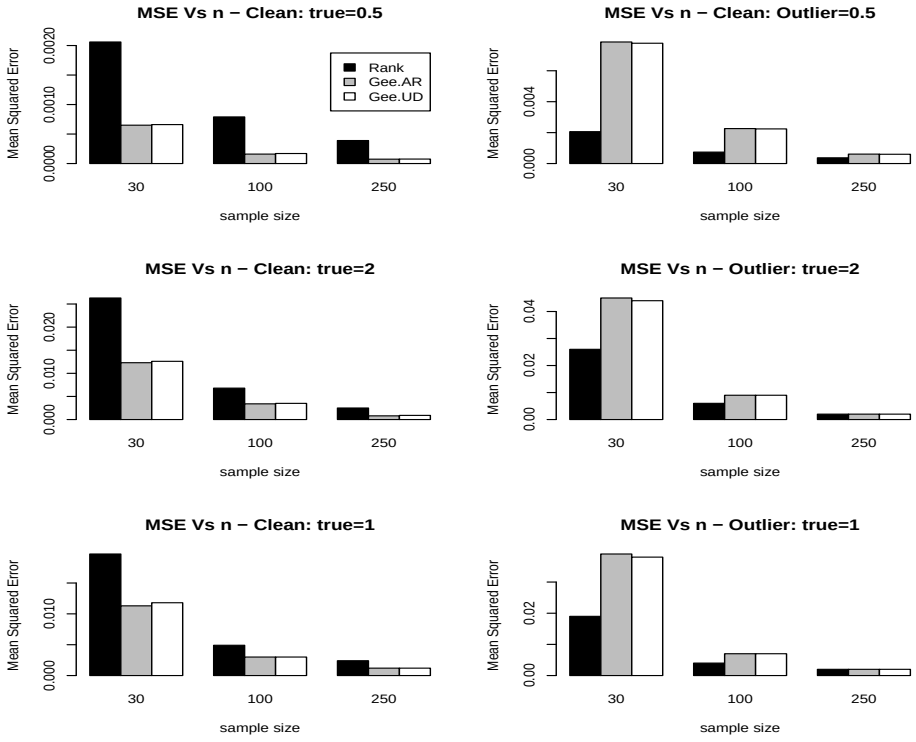


Figure 4.2: MSE comparison



4.4.2 Examples

As an example, we have selected part of the study by Plaisance et al. (2007) concerning the effect of a single session of high intensity aerobic exercise on inflammatory markers of subjects taken over time. One purpose of the study was to see if these markers differed depending on the fitness level of the subject. Subjects were placed into one of the two groups (High Fitness and Moderate Fitness) depending on the level of their peak oxygen uptake. The response we consider here is C-reactive protein (CRP). Elevated CRP levels are a marker of low-grade chronic inflammation and may predict a higher risk for cardiovascular disease (Ridker et al., 2002).

Of the 21 subjects in the study, three were removed due to noncompliance or incomplete information. Thus, we consider the remaining 18 individuals, 9 in each group.CRP level was obtained 24 hours and immediately prior to the acute bout of exercise and subsequently 24, 72, and 120 hours following exercise giving 90 data points in all. Let $\mathbf{y}_i$ and t_i denote respectively the 5×1 vectors of observations and times of measurements for subjects i and let x_i denote his/her indicator variable for group, i.e., its components are either 0 (for Moderate Fitness) or 1 (for High Fitness). We fit an interactive gamma regression model with the log link function. That is,

$$\log(E[\mathbf{y}_i]) = \alpha \underline{1}_5 + \beta x_i + \gamma t_i + \mu x_i t_i, \qquad i = 1, 2, \cdots, 18,$$

The question of interest is to determine whether "High Fitness" is associated with lower levels of CRP. Hence, the one-sided hypothesis.

$$H_0 : \beta = 0 \;\; versus \;\; H_a : \beta < 0.$$

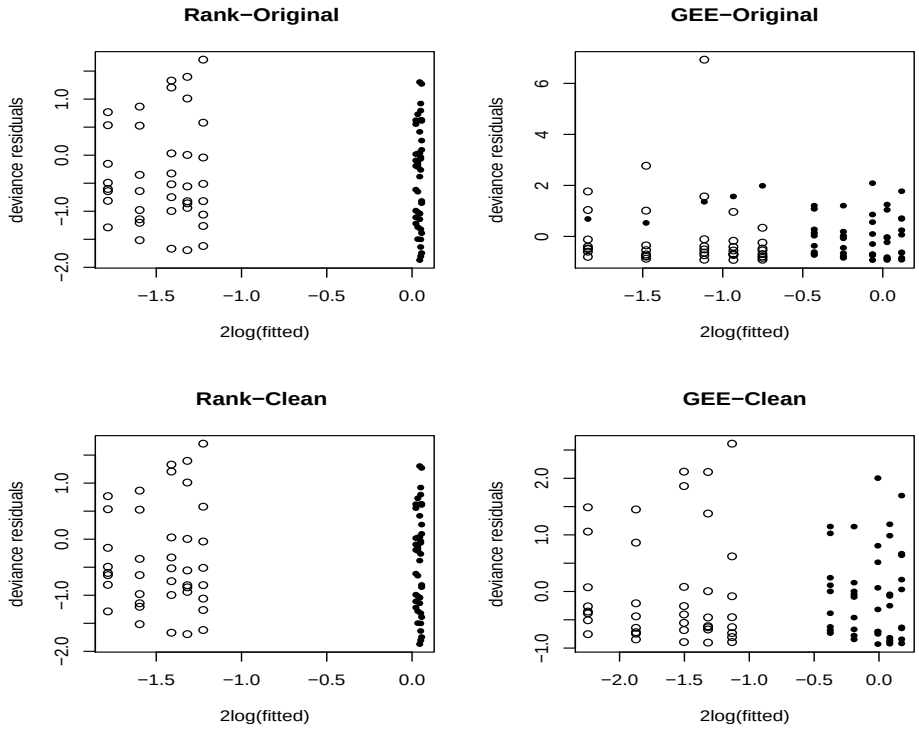
We present the results for the proposed method in comparison with the findings from the GEE estimation procedure with autoregressive correlation structure, using all the 90 data

points. We remove the suspected outlier from the data and repeat the computations. This reduces the data point to 85 points. The Table below and the residual plots give a summary of the findings.

	Coef - Rank		Coef- GEE		SE - Rank		SE - GEE		Pvalue - Rank		Pvalue - GEE	
	Orig	Clean	Orig	Clean	Orig	Clean	Orig	Clean	Orig	Clean	Orig	Clean
α	0.01709	0.0253	0.0134	0.0414	0.2024	0.2462	0.1136	0.1134	0.5335	0.5409	0.5470	0.6425
β	-0.6026	-0.6841	-0.4802	-0.7004	0.3420	0.3101	0.3155	0.1543	0.0391	0.0137	0.0640	0.0000
γ	-0.0022	-0.0085	-0.1365	-0.1375	0.2216	0.2081	0.1234	0.1372	0.4960	0.4838	0.1343	0.1581
μ	-0.0341	-0.1313	-0.1362	-0.1403	0.3371	0.2512	0.3378	0.1806	0.4598	0.3006	0.3434	0.2186

At 5% significance level, with and without the potential outlier, both the proposed method and the GEE procedure reveal that the interaction between time and group is not significant. The main effect of time is not significant either. For the group factor, the proposed method detects a significant difference, with and without the suspected outlier. However, the GEE fails to detect the difference in the group factor for the data containing the potential outlier. So the performance of the GEE procedure is influenced by this potential outlier as it reveals that the group factor is highly significant for the data without the suspected outlier. This clearly shows that the proposed method is robust to local contamination in the response space.

Figure 4.3: Deviance Vs fitted



Deviance versus fit plot gives a pictorial representation of these observations. In the case of the rank estimation, with and without the outlier, the residuals plot display a clustering of subject into two groups (high fitness and moderate fitness). However, for the GEE, this clustering is destroyed in the presence of the outlier. So, the rank based deviance residuals plot can be used for clustering (unsupervised learning) purposes in contaminated data.

4.5 Conclusion

The estimation technique developed in the previous chapter has been extended to longitudinal data. The resulting procedure reduces to the one of Chapter 3 when each and every cluster in the data is of size 1. Unlike the GEE, this procedure does not require estimation of the correlation structure prior to the estimation of the parameters. In addition, the estimator inherits the robustness properties of the estimator in Chapter 3. Its performance evaluated in comparison with the GEE for both simulated and real data, confirms the theoretical results. Like the rank based estimator for generalized linear model, the estimator in this chapter is not protected against outliers in the design space "x". As such, it should only be employed when there is knowledge that x comes from some bounded space (eg. designed experiments).

Chapter 5

Rank Based Group Variable Selection

5.1 Introduction

Group structures in linear models arise for several reasons. For example, in ANOVA, a factor may have several levels and can be expressed via several binary variables. The binary variables corresponding to the same factor form a natural group. Similarly, in additive models, each original prediction variable may be expanded into different order polynomials or a set of basis functions. These polynomials corresponding to the same original prediction variable form a natural group. Another example is the one encountered in gene expression analysis, where genes belonging to the same biological pathway can be considered a group whereas in genetic association studies, genetic markers from the same gene can be considered a group. It is desirable to take into account the grouping structure in the analysis of such data. Several statistical methods have been developed for variable selection that respect the grouping structure. Yuan and Lin (2006) and Zhao et al. (2009) studied the Lasso model for group variable selection.Yuan and Lin (2006) used a penalty function based on the L_2 norm of the coefficients within each group to achieve a group selection. Zhao et al. (2009) on the other hand employed the L_∞ based penalty. In order to achieve the oracle property, Wang and Leng (2008) extended the group Lasso to the adaptive group Lasso. Antoniadis and Fan (2001) studied a class of block-wise shrinkage approaches for regularized wavelets estimation in nonparametric regression problems.

In many applications, however, the data are contaminated with outliers, or even worse have a noise distribution that is heavy tailed. Variable selection methods based on least-squares objective function or maximum likelihood estimation, like the adaptive group Lasso and many other related methods, are not guaranteed to be protected against the adverse

effect of outliers and heavy tailed noise distributions. In linear models, influential outliers are often associated with the explosion of parameter vector estimates. Therefore variable selection in the presence of outliers and heavily asymmetric noise distribution can result in recruiting irrelevant variables or failing to detect important predictors. The problems become much more severe in group variable selection or high dimensional regression where the aim is to both reduce the dimension and estimate the model parameters. To mitigate the adverse effects of such issues, we propose a rank based group variable selection. Rank based variable selection methods have been studied by Wang and Li (2009) and Johnson and Peng (2008). Both address individual variable selection. In addition, the method proposed by Wang and Li (2009) is robust both in the predictor and response space, whereas Johnson and Peng (2008) was concerned with robustness in the response space. In this work we penalize a weighted rank based objective function, identical to the one in Wang and Li (2009) except for our use of a group adaptive Lasso type penalty function. We achieve robustness in both the response and the predictor space while simultaneously performing variable selection that respects the group structure in the data.

The remainder of this chapter is organized as follows: Section 5.2 introduces the proposed rank based group variable selection. The asymptotic distribution and the oracle property of the proposed estimator are developed in Section 5.3. In Section 5.4 we give simulation studies to investigate the theoretical results established in the previous section for finite samples. We conclude the present work in Section 5.5. Proofs are presented in Section 5.6.

5.2 Rank Estimator

5.2.1 Model and Notation

Let $(\mathbf{x}_1, y_1), \cdots (\mathbf{x}_n, y_n)$ be n independent and identically distributed random vectors, where $y_i \in \mathbb{R}$ is the response of interest and $\mathbf{x}_i \in \mathbb{R}^p$ is the associated p-dimensional predictor. Furthermore, it is assumed that $\mathbf{x}_i$ can be grouped into K groups as $\mathbf{x}_i = (\mathbf{x}_1^t, \cdots, \mathbf{x}_K^t)^t$, where $\mathbf{x}_{ik} = (x_{ik1}, \cdots, x_{ikp_k})^t \in \mathbb{R}^{p_k}$ is a group of p_k variables. To model the relationship

between the responses y_i and the predictors $\mathbf{x}_i$, we consider the linear regression model

$$y_i = \sum_{k=1}^{K} \mathbf{x}_{ik}^t \boldsymbol{\theta}_k + \epsilon_i = \sum_{k=1}^{K} \sum_{j=1}^{p_k} \mathbf{x}_{ikj} \theta_{kj} + \epsilon_i.$$

Where $\boldsymbol{\theta}_k = (\theta_{k1}, \cdots, \theta_{kp_k})^t \in \mathbb{R}^{p_k}$ is the regression coefficient vector associated with the kth group, $\boldsymbol{\theta}$ is defined as $\boldsymbol{\theta} = (\theta_1^t, \cdots, \theta_K^t)^t$, and $(\epsilon_1, \cdots, \epsilon_n)$ are independent and identically distributed errors with absolutely continuous density f.

In such models, the interest lies in identifying important groups or factors instead of individual variables (cf. Yuan and Lin, 2006). The terms group and factor are used interchangeably to indicate grouping of variables. These grouped variables can be encountered in many statistical models. For example, in ANOVA a factor may have several levels and can be expressed via several dummy variables, then the dummy variables corresponding to the same factor form a natural group. Similarly, in additive models, each original predictor may be expanded into different order polynomials, then these polynomials corresponding to the same original predictor form a natural group.

Multiple group variable selection methods based on a Penalized Least Square objective function have been proposed, see for example Wang and Leng (2008). However in the presence of outliers or error terms ϵ_i from a heavy tailed distribution, the Penalized Least Square Estimators may perform poorly.Johnson and Peng (2008) as well as Wang and Li (2009) have proposed penalized rank based procedures for variable selection, in linear models, as a remedy to this problem. We extend these rank based techniques to linear models with grouped variables. Hence the proposed objective function is

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i<j} b_{ij} |\epsilon_i - \epsilon_j| + n \sum_{k=1}^{K} \sum_{j=1}^{p_k} \lambda_{kj} |\theta_{kj}| \ ,$$

where $\lambda_{kj} = \lambda_n / (\|\tilde{\boldsymbol{\theta}}_k\|^2 |\tilde{\theta}_{kj}|)$, $\tilde{\boldsymbol{\theta}}_k$ and $\tilde{\theta}_{kj}$ the unpenalized estimators of $\boldsymbol{\theta}_k$ and θ_{kj}, respectively, λ_n some data driven regularization parameter, and b_{ij} some positive and symmetric

weights used to downweight high leverage points. Note that λ_{kj} is the penalty term of the jth variable of the kth group. This choice of the penalty function combining two norms (one at the group level and the other at the individual variable level) is in part motivated by the elastic net (Zou and Hastie, 2005). Unlike the elastic net, here both norms are L_1 norm and are combined in a multiplicative way. Note that the penalty term $\|\tilde{\boldsymbol{\theta}}_k\|^{-2}$ is the common penalty imposed on the members of the kth group, and within the kth group we regularize the jth member by $|\tilde{\theta}_{kj}|^{-1}$ in order to reduce the bias introduced by the group penalty on individual members of the group. A more general penalty term, $\lambda_n/\|\tilde{\boldsymbol{\theta}}_k\|^{\gamma_1}|\tilde{\theta}_{kj}|^{\gamma_2}$, can be considered; where the choice of $\gamma_1 \geq 0$, and $\gamma_2 \geq 0$ are driven by whether the interest lies in individual or group selection. Observe that the weights λ_{kj} are adaptive in nature. That is, if the effect of a variable is strong, the corresponding coefficient is lightly penalized and vice-versa, while respecting the grouping structure among the variables.

The corresponding estimator is defined by

$$\widehat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta}}{\operatorname{Argmin}}\, Q_n(\boldsymbol{\theta}) \ .$$

In the presence of outliers in either y or x direction, $\widehat{\boldsymbol{\theta}}_n$ defined above remains relatively unaffected as will be shown both theoretically and in simulations. In contrast, the least squares based group variable selection methods like adaptive hierarchical Lasso of Zhou and Zhu (2010) are vulnerable to outliers in either x or y direction.

Observe that the objective function Q_n reduces to the ww-scad by Wang and Li (2009) for the particular choice of the scad penalty function in the special case of $\boldsymbol{\theta}_k$ with dimension 1 for each k, that is no grouped variables are present in the model or all factors have at most two levels. If in addition, $b_{ij} \equiv 1$, $Q_n(\boldsymbol{\theta})$ is identical to the penalized rank dispersion function proposed by Johnson and Peng (2008). While the method of Wang and Li (2009) may not be appropriate to identify important factors or remove irrelevant groups in linear models, the penalized estimator given by Johnson and Peng (2008), in addition, is not protected

against high leverage points. A consequence of variable selection procedures that are not robust to high leverage points is that not only does it affect the estimation, it also affects the selection procedure. The method proposed in this work, however, yields estimators that are robust in both x and y direction while taking into account the grouping structures among the covariates.

5.3 Main Results

In this section, we study the asymptotic properties of the proposed rank based group variable selection estimator. We show that the penalized rank based group variable selection estimator has the oracle property under some regularity conditions.

Without loss of generality, we assume that only the first $k_0 \leq K$ groups are important. That is, we assume that $\|\boldsymbol{\theta}_k\| \neq 0$ for $k \leq k_0$ and $\|\boldsymbol{\theta}_k\| = 0$ for $k > k_0$. Denote $\boldsymbol{\theta}_0$ the true parameter, $\boldsymbol{\theta}_a = (\boldsymbol{\theta}_1^t, \cdots, \boldsymbol{\theta}_{k_0}^t)^t$ the vector containing all relevant groups and $\boldsymbol{\theta}_b = (\boldsymbol{\theta}_{1+k_0}^t, \cdots, \boldsymbol{\theta}_K^t)^t$ the vector made of all the irrelevant groups. Furthermore, let $\widehat{\boldsymbol{\theta}}_a$ and $\widehat{\boldsymbol{\theta}}_b$ be their corresponding penalized rank estimator.

Similarly to Wang and Li (2009) and following their notation, we will use the GR weights (Sievers (1983)), given by $b_{ij} = b(\mathbf{x}_i, \mathbf{x}_j) = h(\mathbf{x}_i)h(\mathbf{x}_j)$, to downweight high leverage points. $h(\mathbf{x}_i)$ is defined as:

$$h(\mathbf{x}_i) = min\left[1, \frac{b}{(\mathbf{x}_i - \hat{\mu})' S^{-1} (\mathbf{x}_i - \hat{\mu})}\right]$$

with $(\hat{\mu}, S)$ being the robust minimum volume ellipsoid estimators of the location and scatter and b the 95th percentile of $\chi^2(p)$

The following assumptions will be made.

A1. The errors's ϵ_i density function f has a finite Fisher information. That is,

$$I(f) = \int_{-\infty}^{\infty} \left[\frac{f'(e)}{f(e)}\right]^2 f(e)de < \infty$$

A2. The matrices $\mathbf{X}$ and $W\mathbf{X}$ both satisfy the Huber's condition.

A3. $n^{-1}\mathbf{X}'\mathbf{W}\mathbf{X}\xrightarrow{\mathcal{P}}\mathbf{C}$, $n^{-1}\mathbf{X}'\mathbf{W}^2\mathbf{X}\xrightarrow{\mathcal{P}}\mathbf{V}$, and $n^{-1}\mathbf{X}'\mathbf{X}\xrightarrow{\mathcal{P}}\mathbf{\Sigma}$,where $\mathbf{C}$, $\mathbf{V}$,and $\mathbf{\Sigma}$ are positive definite matrices.

given by

$$\mathbf{C} = \frac{1}{2}\int\int(\mathbf{x}_2-\mathbf{x}_1)(\mathbf{x}_2-\mathbf{x}_1)'b(\mathbf{x}_1,\mathbf{x}_2)dM(\mathbf{x}_2)dM(\mathbf{x}_1)$$
$$\mathbf{V} = \int\{(\mathbf{x}_2-\mathbf{x}_1)b(\mathbf{x}_1,\mathbf{x}_2)dM(\mathbf{x}_2)\}\{(\mathbf{x}_2-\mathbf{x}_1)b(\mathbf{x}_1,\mathbf{x}_2)dM(\mathbf{x}_2)\}'dM(\mathbf{x}_1)$$
$$\mathbf{\Sigma} = \frac{1}{2}\int\int(\mathbf{x}_2-\mathbf{x}_1)(\mathbf{x}_2-\mathbf{x}_1)'dM(\mathbf{x}_2)dM(\mathbf{x}_1)$$

and $M(\mathbf{x})$ denotes the CDF of $\mathbf{x}$, $\mathbf{X}$ is a matrix whose rows are $\mathbf{x}_i$; and the entries ω_{ij} of the matrix $\mathbf{W}$ are defined, like in Naranjo and Hettmansperger (1994), by

$$\omega_{ij} = \begin{cases} n^{-1}b_{ij} & \text{if } i \neq j \\ n^{-1}\sum_{k\neq i} b_{ij} & \text{if } i = j \end{cases}$$

Remark 5.1. *Assumptions A1 to A3 are identical to the ones in Wang and Li (2009). As noted in their paper, these assumptions guarantee the $\sqrt{n}$-consistency and the asymptotic normality of the unpenalized estimator through the asymptotic quadraticity of the unpenalized objective function and the asymptotic linearity of the corresponding score function.*
All results will be conditional on the matrix $\mathbf{X}$. That is the matrix $\mathbf{X}$ is treated as fixed.

We are now ready to state the Theorem that gives the estimation consistency, the selection consistency and the oracle property of the proposed estimator. Following the notation in Wang and Leng (2008), define

$$a_n = max\{\lambda_{kj} : 1 \leq j \leq p_k \ ; \ k \leq k_0\} \quad and \quad b_n = min\{\lambda_{kj} : 1 \leq j \leq p_k \ ; \ k > k_0\} \ .$$

Theorem 5.1. *Let* $(y_1, \mathbf{x}_1), \cdots, (y_n, \mathbf{x}_n)$ *be independent and identically distributed from* $H(x, y)$*. Assume the regularity conditions* $A1 - A3$*.*

a. If $\sqrt{n}a_n \xrightarrow{\mathcal{P}} 0$ *then* $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = O_p(n^{-1/2})$

b. If $\sqrt{n}a_n \xrightarrow{\mathcal{P}} 0$ *and* $\sqrt{n}b_n \xrightarrow{\mathcal{P}} \infty$ *then* $\widehat{\boldsymbol{\theta}}_b \xrightarrow{\mathcal{P}} \mathbf{0}$

c. Under local shrinking contamination, $H_n^*(\mathbf{x}, y)$, $\sqrt{n}(\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_a) \xrightarrow{\mathcal{D}} N(\eta, \tau^2 C_{11}^{-1} V_{11} C_{11}^{-1})$

where $\tau^2 = [\sqrt{12}\int f^2(u)du]^{-1}$*,* C_{11} *is the* $k_0 \times k_0$ *submatrix in the upper-left corner of* C*,* V_{11} *the* $k_0 \times k_0$ *submatrix in the upper-left corner of* V*,* $H_n^*(\mathbf{x}, y) = \left(1 - \frac{\delta}{\sqrt{n}}\right) H(\mathbf{x}, y) + \frac{\delta}{\sqrt{n}} \Delta_{(\mathbf{x}^*, y^*)}$*,* $\eta = \delta[2F(y^* - \mathbf{x}^*\boldsymbol{\theta}_0) - 1] \int b(\mathbf{x}^*, \mathbf{x})(\mathbf{x}^* - \mathbf{x}) dM(\mathbf{x})$*, with* $\Delta_{(\mathbf{x}^*, y^*)}$ *representing a point mass at* $(\mathbf{x}^*, y^*)$ *and* δ *some constant.*

The assumptions $\sqrt{n}a_n \xrightarrow{\mathcal{P}} 0$ and $\sqrt{n}b_n \xrightarrow{\mathcal{P}} \infty$ are identical to the ones in Wang and Leng (2008) for the special case $\lambda_{kj} = \lambda_n / \|\tilde{\boldsymbol{\theta}}_k\| \ \forall j$. So with probability tending to 1, the proposed estimation technique correctly identifies relevant groups, removes irrelevant ones, and estimate the corresponding coefficients as if the true model was known in advance.

Remark 5.2. *As noted by Wang and Li (2009), the asymptotic bias* η *is bounded in* y^* *and also bounded in* x^* *with the proper choice of the weights* b_{ij}*, such as the GR weights introduced above. In addition, in the absence of local contamination, the asymptotic bias* $\eta = 0$*. The proof of part (c) of Theorem 5.1 is identical to the proof of Theorem 2 of Wang and Li (2009) and will therefore be omitted in this work.*

5.4 Monte Carlo Simulations

Two set of simulation studies were conducted to evaluate the performance of the rank based group variable selection, for finite sample sizes, in comparison with the rank based variable selection (Johnson and Peng, 2008), the ww-scad (Wang and Li, 2009), and the adaptive hierarchical Lasso (Zhou and Zhu, 2010). Mainly, In example 1 we compare the proposed method to the rank based variable selection and the adaptive hierarchical Lasso. In

the second example we compare it to all the three methods. For simplicity of presentation, we label the adaptive hierarchical Lasso by "hLasso", the proposed method by "grvs", the rank based variable selection by "rvs", and we keep "ww-scad" for the ww-scad method.

5.4.1 Example 1:

We borrow the model and simulation settings used in Zhou and Zhu (2010). We consider a model which has both categorical and continuous predictors. We first generate seventeen independent standard normal random variables $z_1, z_2, \cdots, z_{16}$ and w. The predictors are then defined as $x_j = (z_j + w)/\sqrt{2}$. Each of the predictors $x_1, x_2, \cdots, x_8$ is expanded through a fourth-order polynomial. Subsequently, the last eight variables $x_9, \cdots, x_{16}$ are all discretized to 0,1,2, and 3 according to whether they are smaller than $\Phi^{-1}(1/4)$; between $\Phi^{-1}(1/4)$ and $\Phi^{-1}(1/2)$; between $\Phi^{-1}(1/2)$ and $\Phi^{-1}(3/4)$ or greater than $\Phi^{-1}(3/4)$. This results in eight continuous groups of size four each and eight categorical groups with four levels each corresponding to three binary variables per category. We consider the following model

$$y = [x_3 + 0.5x_3^2 + 0.1x_3^3 + 0.1x_3^4] + [x_6 - 0.5x_6^2 + 0.15x_6^3 + 0.1x_6^4] + [I(x_9 = 0) + I(x_9 = 1) + I(x_9 = 2)] + \epsilon \,, \tag{5.4.1}$$

where $I()$ is the indicator function. We consider the error term from the following distribution: standard normal, t with 3 degrees of freedom, standard Laplace, the standard Cauchy, and a standard normal contaminated with a normal distribution with mean zero and standard deviation 2. The proportions of contamination considered are 0.1, 0.2, 0.3, 0.4, and 0.5.

The regularization parameter λ_n is chosen such that the corresponding estimator, $\widehat{\boldsymbol{\theta}}_{\lambda_n}$, minimizes the generalized cross validation (equivalently the AIC type criterion) as both the rank based variable selection and the adaptive hierarchical Lasso estimators were originally developed with λ_n tuned with the generalized cross validation. Each of the sub-models was run 200 times. The results, for the sample size $n = 400$, are summarized in Table 5.1. For

each case, we report the average proportion of coefficients correctly identified as zero, the average proportion of non-zero coefficients correctly identified, and the median model error, i.e the median of the quantity $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^t(\mathbf{x}^t\mathbf{x})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ over the 200 runs. Overall, We observe that under the standard normal distribution, the adaptive HLasso performs better than the proposed rank based group variable selection and the proposed method in turn performs better than the rank based variable selection. When the error is from the t distribution with 3 degrees of freedom the proposed estimation is the best of the three and the rank variable selection is comparable to the adaptive hierarchical Lasso. For error distributions from the contaminated normal distribution and the standard Cauchy distribution, the rank based group variable selection dominates the performance of the rank variable selection which in turn dominates the performance of the adaptive hierarchical Lasso. Figures 5.1, 5.2, and 5.3 provide a graphical representation of these findings.

Figure 5.1: Variable Selection Performance Comparison 1

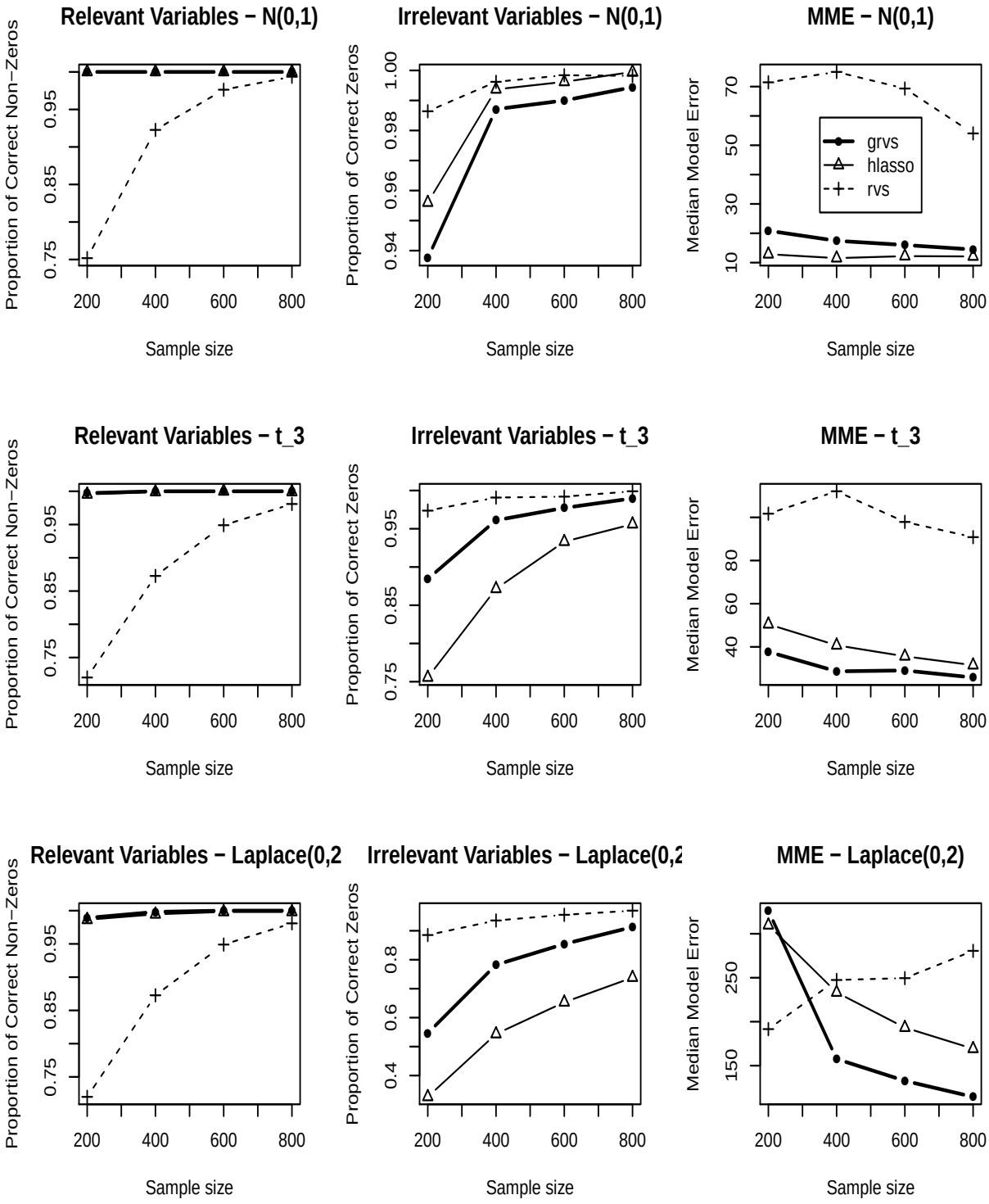


Figure 5.2: Variable Selection Performance Comparison 2

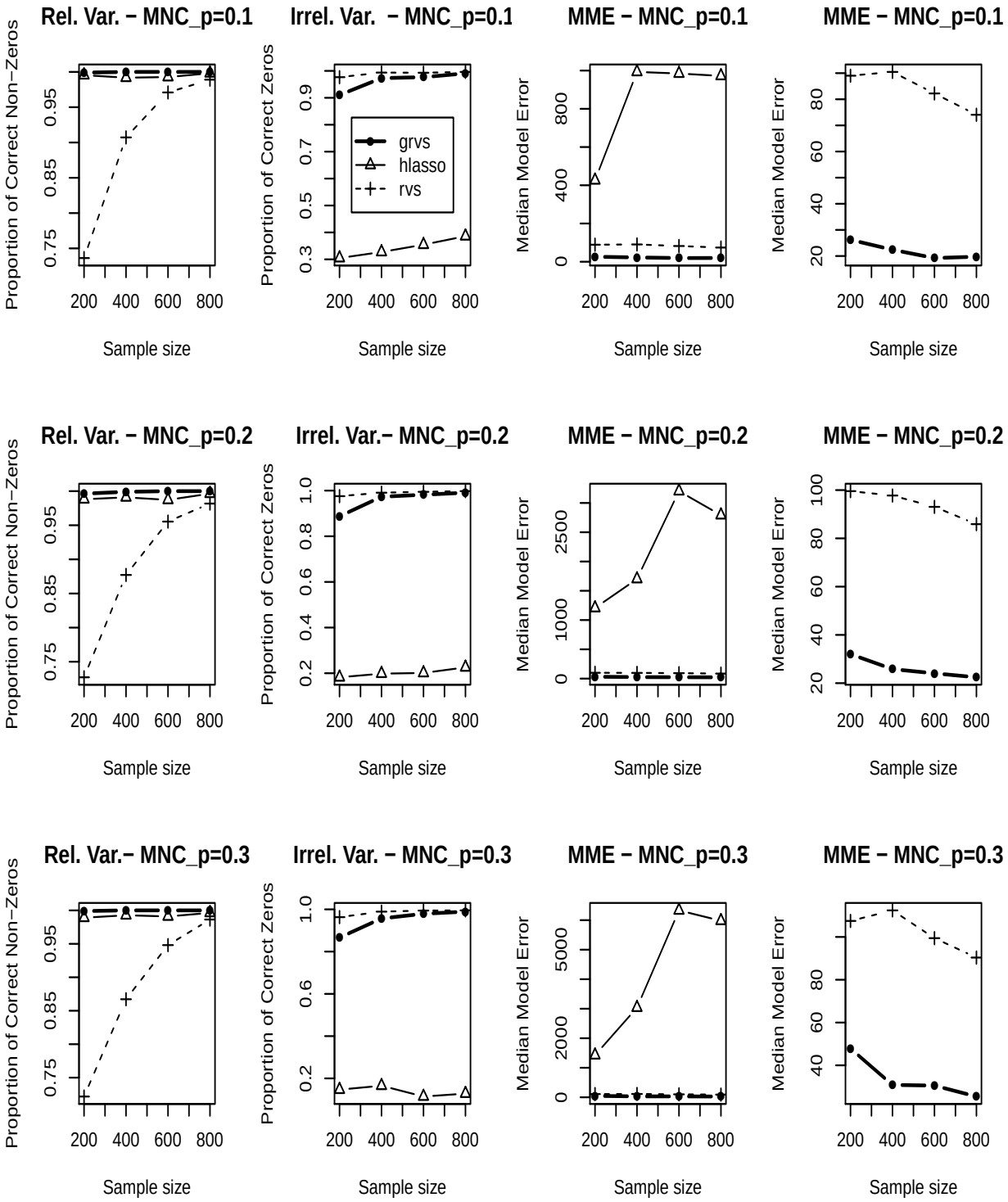
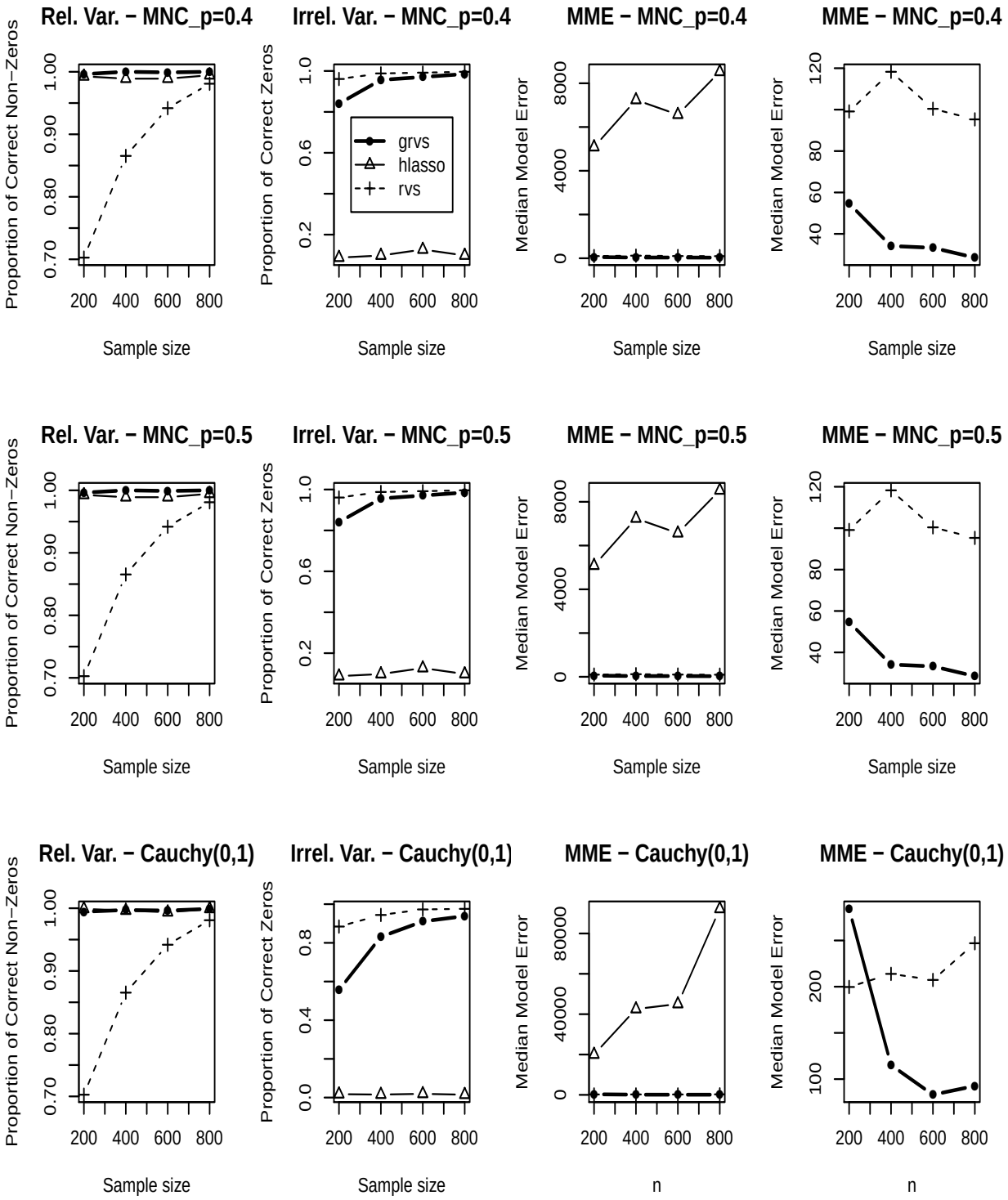


Figure 5.3: Variable Selection Performance Comparison 3



5.4.2 Example 2:

In this example, we also consider a model which has both categorical and continuous predictors. However, we generate ten latent variables $x_1, x_2, \cdots, x_{10}$ from a multivariate normal distribution with mean zero and covariance between x_i and x_j given by $0.5^{|i-j|}$. Each of $x_1, x_2, \cdots, x_5$ is expanded through a third-order polynomial. Subsequently, the last five latent variables $x_6, \cdots, x_{10}$ are all discretized to 0, 1, 2, and 3 according to whether they are smaller than $\Phi^{-1}(1/4)$; between $\Phi^{-1}(1/4)$ and $\Phi^{-1}(1/2)$; between $\Phi^{-1}(1/2)$ and $\Phi^{-1}(3/4)$ or greater than $\Phi^{-1}(3/4)$. This results in five continuous groups of size three each and five categorical groups with four levels each corresponding to three binary variables per category. We consider the following model:

$$y = [2.3x_3 + 2x_3^2 + 1.8x_3^3] + [3.2x_5 - 1.5x_5^2 + 2.5x_5^3] + [3I(x_8 = 0) + 2.4I(x_8 = 1) + 3.1I(x_8 = 2)] + \epsilon \, . \tag{5.4.2}$$

We first consider the error term, ϵ, from the normal distribution with mean zero and different standard deviations 1, 2, 4, and 8. Next to evaluate the estimators in the presence of heavy tails, we generate the random error terms from a t-distribution with 2 degrees of freedom and a Cauchy distribution with parameters 0 and 1. We also investigate the effect of outliers in the $\mathbf{x}$ direction on model selection. For this purpose, we consider models with normally distributed error with mean 0 each and standard deviations 1 and 4, respectively, where in both models we replace a random 1% of the elements of the design matrix $\mathbf{X}$ by observations from an exponential distribution with parameter 0.1.

Here the regularization parameter λ_n is chosen such that the corresponding estimator, $\widehat{\boldsymbol{\theta}}(\lambda_n)$, minimizes the BIC criterion, since this is what was used originally by Wang and Li (2009) in their simulations. Each of the five sub-models was run 250 times. The results are summarized in Table 5.2 for the sample size $n = 200$. For each case, we report the

average proportion of coefficients correctly identified as zero, the average proportion of non-zero coefficients correctly identified, and the mean model error. We observe that under the normally distributed errors model, the HLasso has the smallest model error and its performance improves when the signal to noise ration σ/n increases. This observation was made by Johnson and Peng (2008) for the Lasso and elastic net method.However the ww-scad, the rank based variable selection and the method proposed in this work perform better for small values of σ/n. The proposed method, the rank based group variable selection, has smaller model error and an overall better performance than the ww-scad and the rank based variable selection in all the four normally distributed error models.

When the error distribution is from the standard Cauchy distribution, or t-distribution with 2 degrees of freedom, rank based group variable selection performs better than the other three methods. The ww-scad and the rank variable selection have both smaller model error than the hierarchical Lasso. The rank based group variable selection continues to be the best of the four methods in the case of the normally distributed error with contaminated design matrix $\mathbf{X}$. Figures 5.4, 5.5, and 5.6 give us a pictorial description of these results.

Figure 5.4: Variable Selection Performance Comparison 4

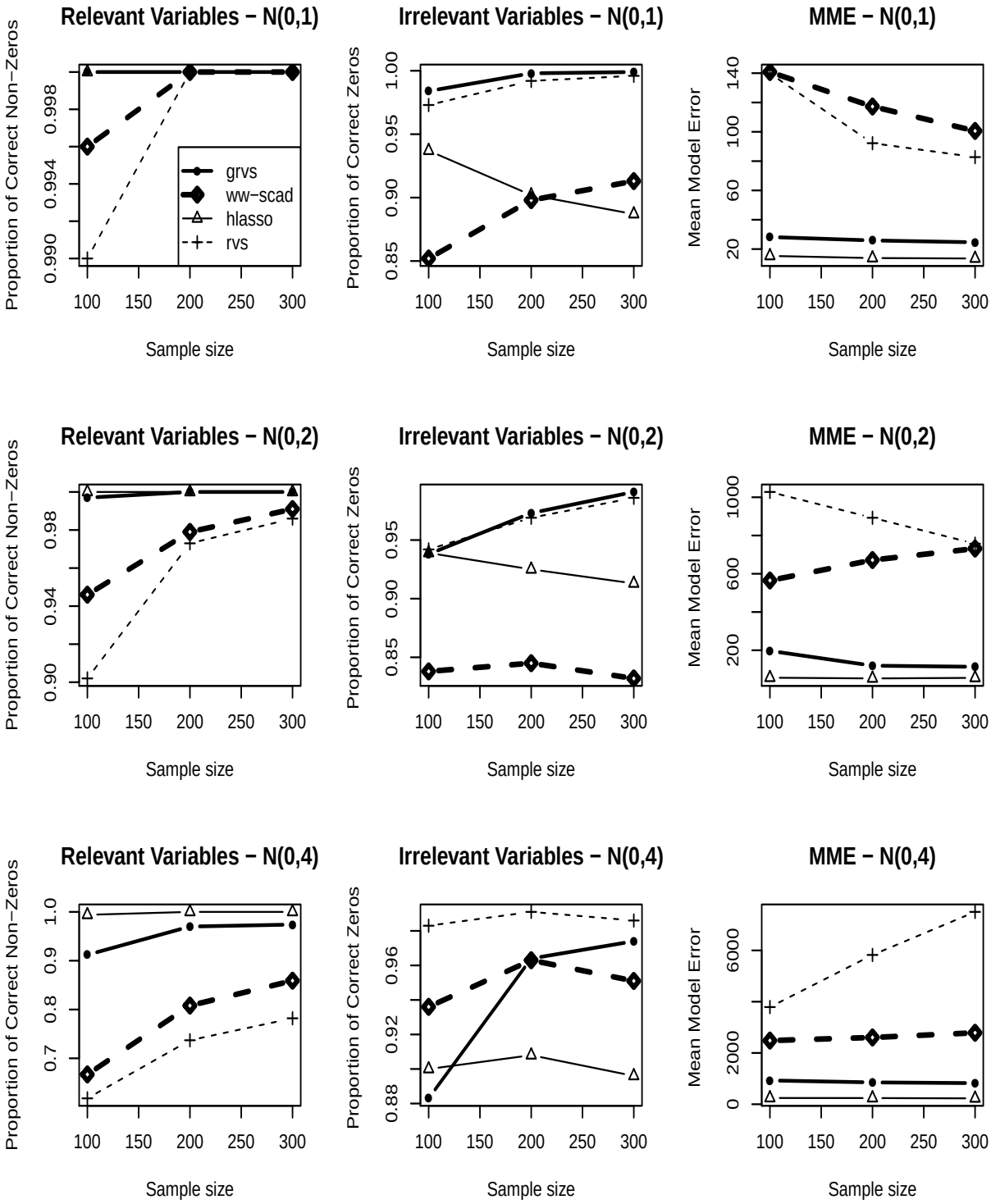


Figure 5.5: Variable Selection Performance Comparison 5

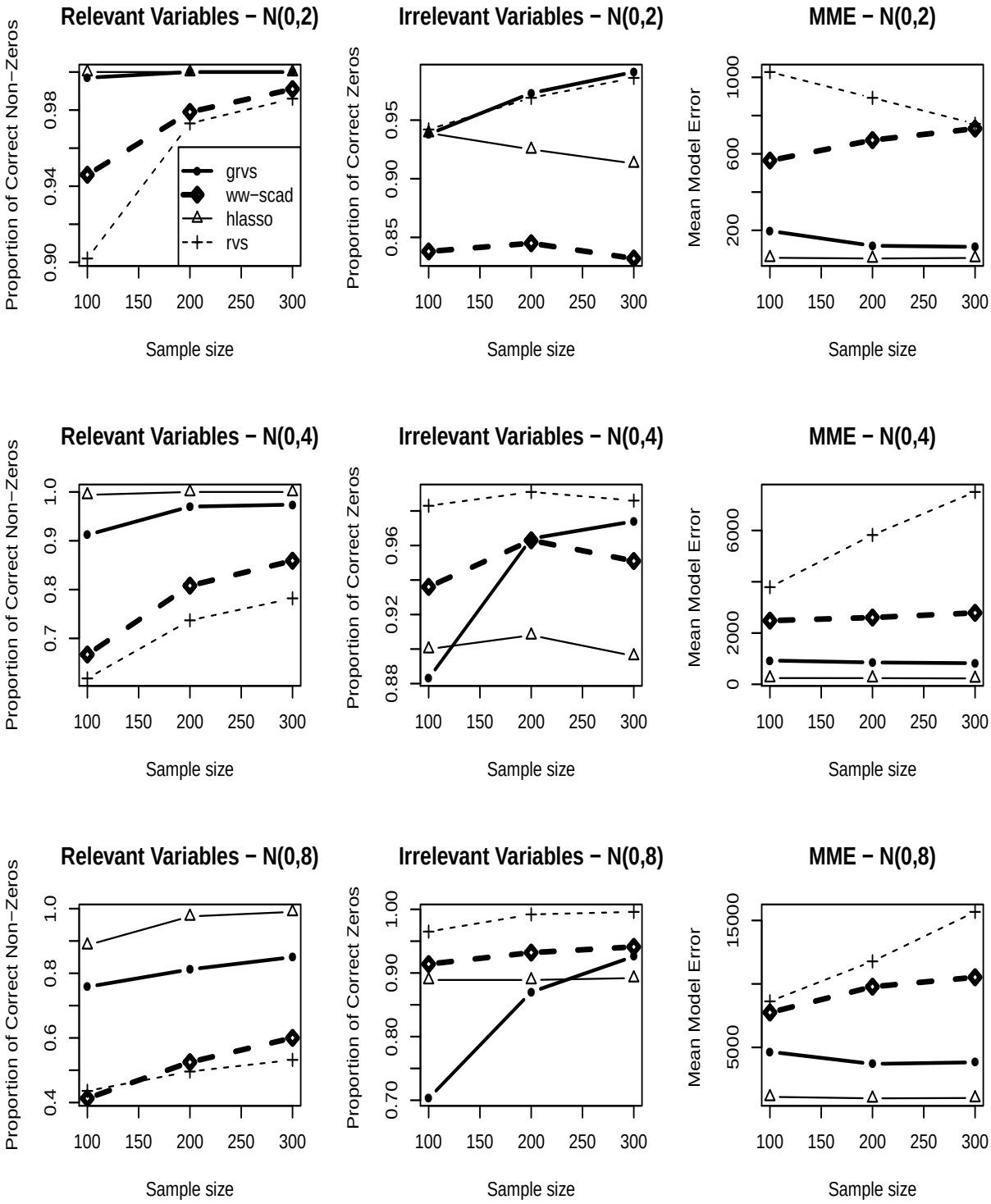


Figure 5.6: Variable Selection Performance Comparison 6

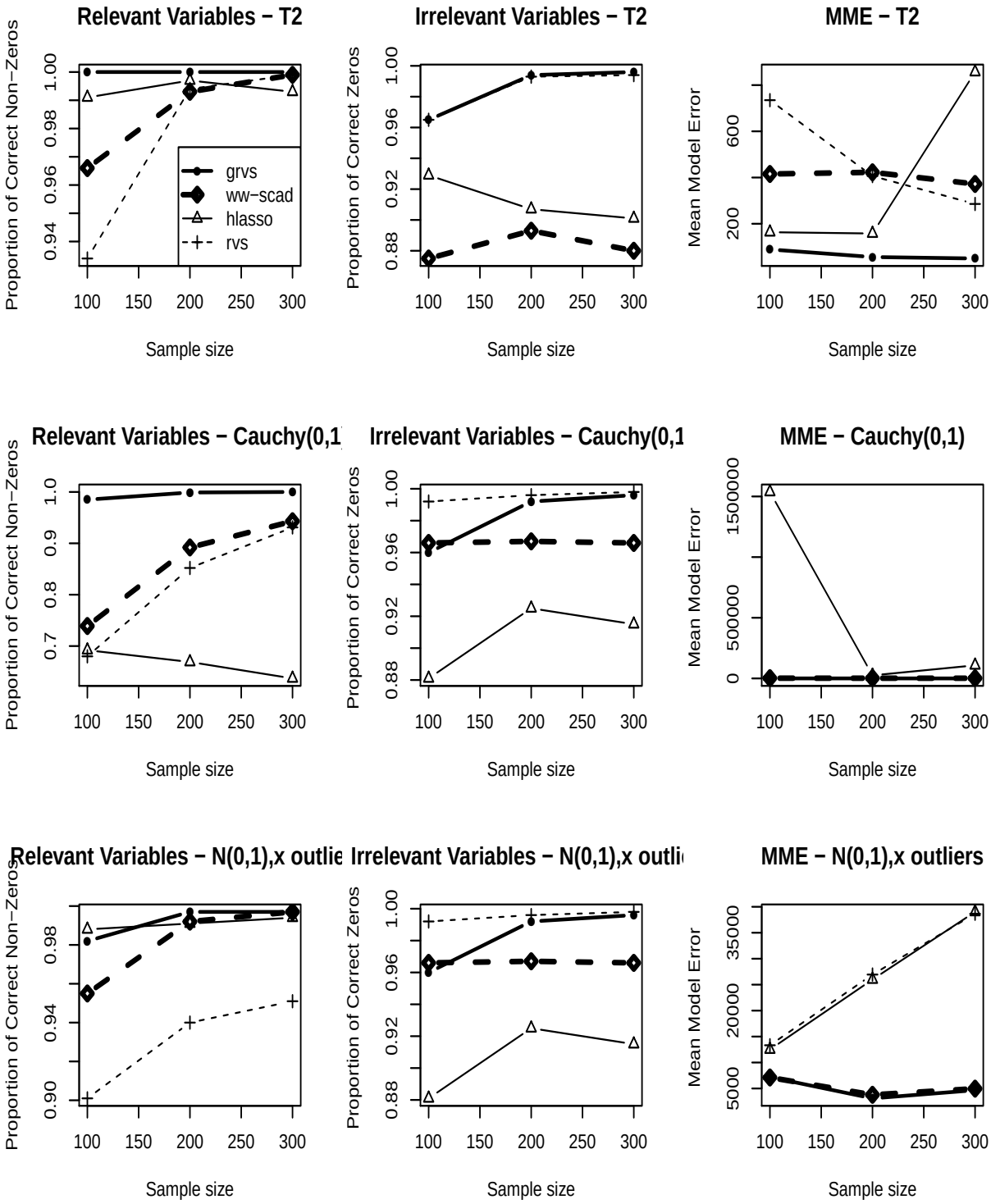


Figure 5.7: Variable Selection Performance Comparison 7

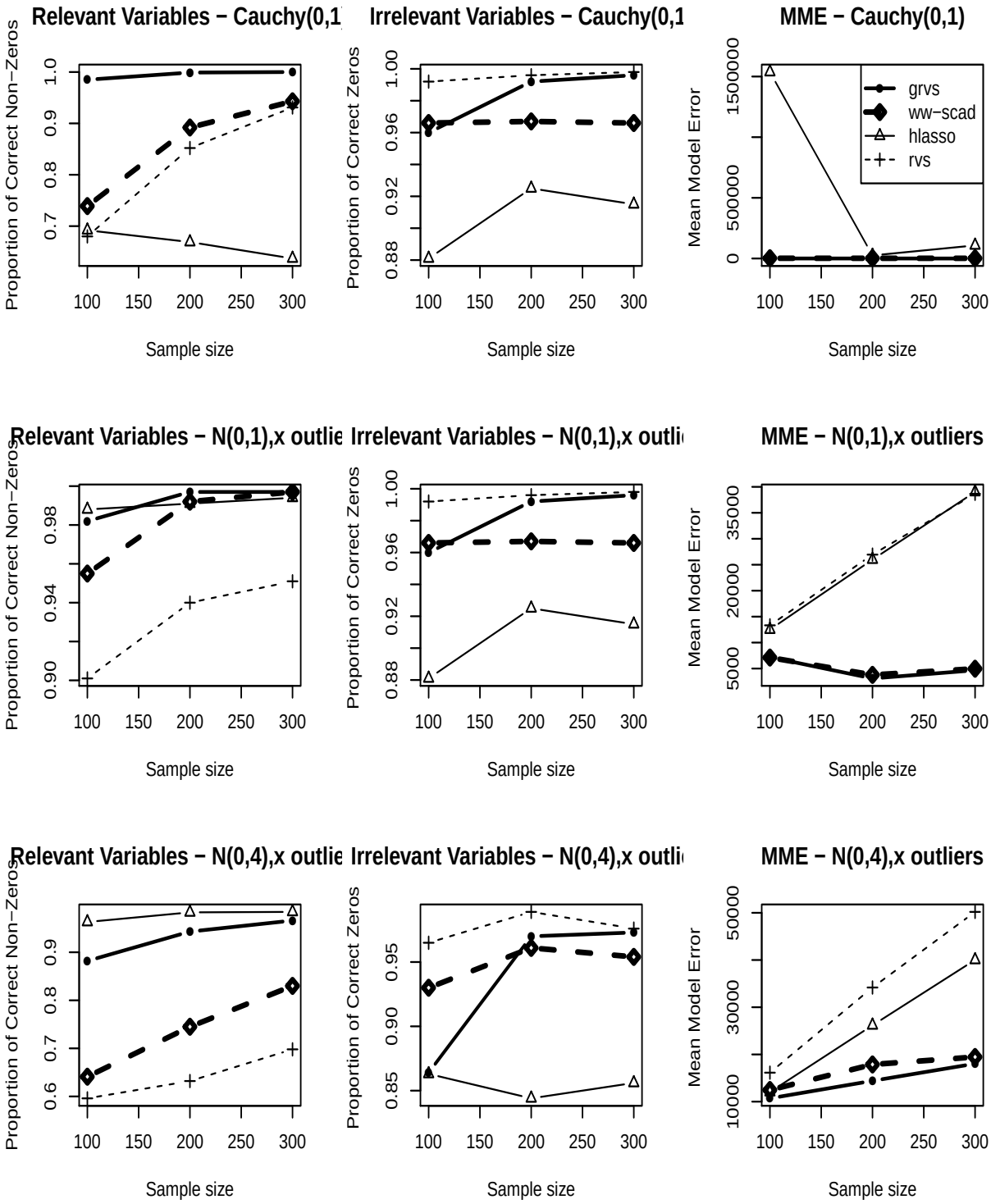


Figure 5.8: Variable Selection Performance Comparison 8

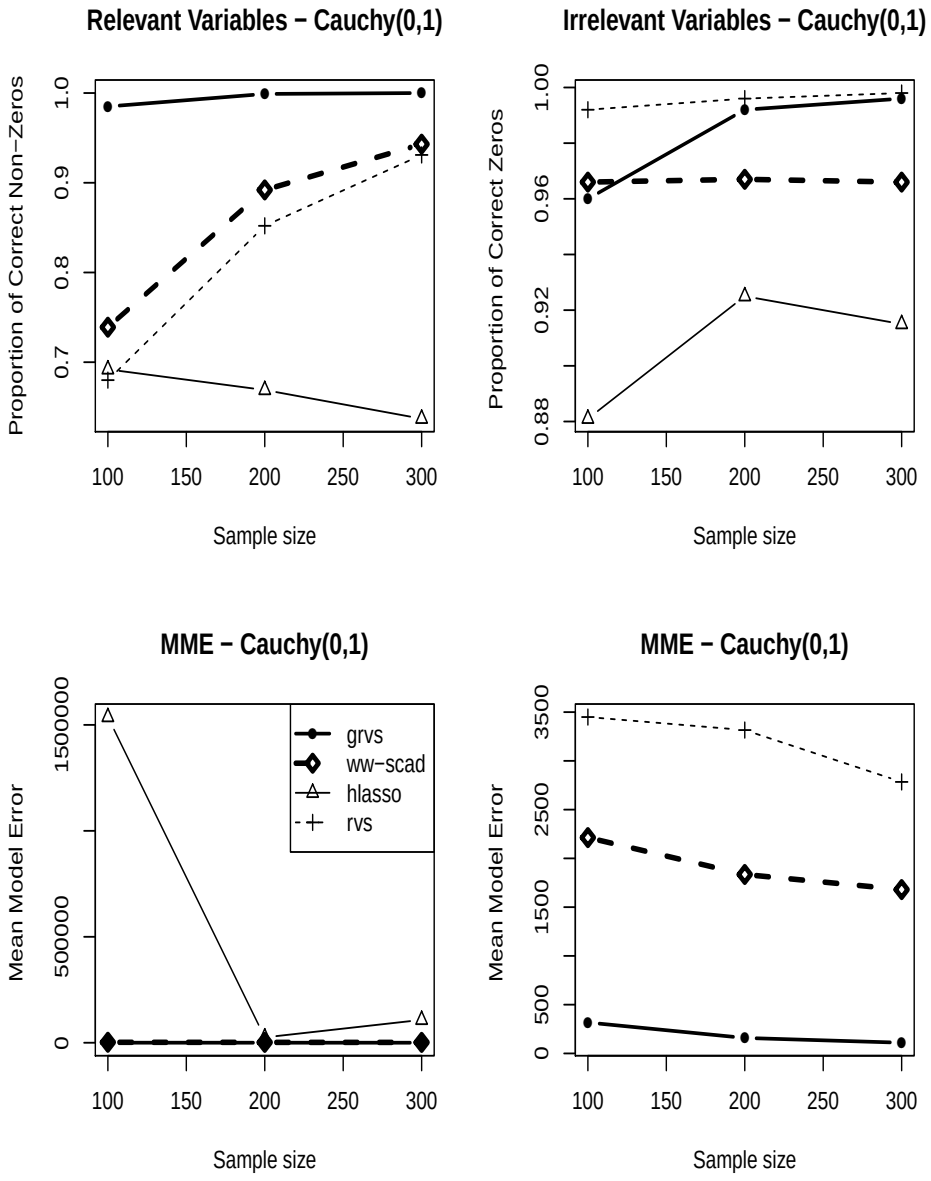


Table 5.1: Variable Selection Performance Comparison - Model 1

		grvs	hLasso	rvs
Normal(0,1)	relevant variables	1	1	0.922
	irrelevant variables	0.987	0.993	0.996
	median model error	17.44	11.53	74.99
t_3	relevant variables	1	0.999	0.872
	irrelevant variables	0.961	0.871	0.990
	median model error	28.66	40.73	112.08
Laplace(0,2)	relevant variables	0.998	0.994	0.748
	irrelevant variables	0.782	0.543	0.935
	median model error	157.55	233.25	247.41
0.7N(0,1)+0.3C(0,2)	relevant variables	1	0.992	0.867
	irrelevant variables	0.956	0.165	0.989
	median model error	30.85	3036.46	112.42
Cauchy(0,1)	relevant variables	0.997	0.996	0.767
	irrelevant variables	0.831	0.0155	0.944
	median model error	115.17	42541.18	213.86

Table 5.2: Variable Selection Performance Comparison - Model 2

		ww-scad	grvs	hLasso	rvs
Normal(0,2)	relevant variables	0.979	1	1	0.973
	irrelevant variables	0.845	0.973	0.925	0.969
	mean model error	671.70	119.47	52.83	892.68
t_2	relevant variables	0.993	1	0.997	0.994
	irrelevant variables	0.893	0.994	0.907	0.993
	mean model error	422.608	56.01	157.80	406.46
Cauchy(0,1)	relevant variables	0.892	0.999	0.669	0.852
	irrelevant variables	0.967	0.992	0.925	0.996
	mean model error	1835	158.94	25257.01	3315.77
Normal(0,1)-	relevant variables	0.992	0.997	0.991	0.940
x outliers	irrelevant variables	0.967	0.992	0.925	0.996
	mean model error	3781	3100	25989.70	26946.32

5.5 Conclusion

A robust group variable selection method for linear regression has been developed in this work. The proposed objective function is a weighted Wilcoxon objective function penalized with a group penalty function. The resulting estimator is robust both in the design and the response space. It provides robust simultaneous variable selection and estimation in linear models with grouped variables. An extension of this work to high dimensional linear models ($p \gg n$) will be considered in future work.

5.6 Proofs

We adopt the following expressions defined in Wang and Li (2009) .

$$
\begin{aligned}
Q_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i<j} b_{ij}|\epsilon_i - \epsilon_j| + n\sum_{k=1}^{K}\sum_{j=1}^{p_k}\lambda_{kj}|\theta_{kj}| \\
D_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i<j} b_{ij}|\epsilon_i - \epsilon_j| \\
S_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i<j} b_{ij}(\mathbf{x}_i - \mathbf{x}_j)sgn((y_i - y_j) - (\mathbf{x}_i - \mathbf{x}_j)'\boldsymbol{\theta}) \\
A_n(\boldsymbol{\theta}) &= (2\sqrt{3\tau})^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\mathbf{X}'\mathbf{W}\mathbf{X}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - (\boldsymbol{\theta} - \boldsymbol{\theta}_0)'S_n(\boldsymbol{\theta}_0) + D_n(\boldsymbol{\theta}_0)
\end{aligned}
$$

The following lemma establishes the asymptotic quadracity of $D_n(\boldsymbol{\theta})$ as well as the asymptotic normality of $S_n(\boldsymbol{\theta}_0)$. Its proof can be found in Wang and Li (2009) , and will therefore be omitted here.

Lemma 5.1. *Under assumptions* $A1 - A4$.

i. $\forall\epsilon > 0$, $\forall c > 0$,

$$\Big[\sup_{\sqrt{n}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le c}|D_n(\boldsymbol{\theta}) - A_n(\boldsymbol{\theta})| \ge \epsilon\Big] \xrightarrow{\mathcal{P}} 0$$

under either H *or* H_n^*.

ii. $n^{-1/2}S_n(\boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N(0, \mathbf{V}/3)$ *under* H

iii. $n^{-1/2}S_n(\boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N(\eta, \mathbf{V}/3)$ *under* H_n^*

We are now ready to give the proof of Theorem 5.1.

Proof of Theorem 5.1. To prove part (a.), it is sufficient to show that $\forall\epsilon > 0$, there exists a large constant C such that

$$P\Big(\inf_{\|\mathbf{u}\|=C} Q_n(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{u}) > Q_n(\boldsymbol{\theta}_0)\Big) \ge 1 - \epsilon,$$

where $\mathbf{u}$ is a vector of dimension p. Since $Q_n(\boldsymbol{\theta})$ is convex in $\boldsymbol{\theta}$, this implies that with probability at least $1-\epsilon$ the penalized estimator lies in the ball $\{\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{u} : \|\mathbf{u}\| \leq C\}$. Let $G_n(\mathbf{u}) = Q_n(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{u}) - Q_n(\boldsymbol{\theta}_0)$. Denote by u_{kj} the component of $\mathbf{u}$ corresponding to θ_{kj}. By lemma 5.1

$$\begin{aligned}
G_n(\mathbf{u}) &= (2\sqrt{3})^{-1}\mathbf{u}'[n^{-1}\mathbf{X}'\mathbf{W}\mathbf{X}]\mathbf{u} - \mathbf{u}'n^{-1/2}S_n(\boldsymbol{\theta}_0) + n\sum_{k=1}^{K}\sum_{j=1}^{p_k}\lambda_{kj}(|\theta_{kj} + n^{-1/2}u_{kj}| - |\theta_{kj}|) + o_p(1) \\
&\geq (2\sqrt{3})^{-1}\mathbf{u}'[n^{-1}\mathbf{X}'\mathbf{W}\mathbf{X}]\mathbf{u} - \mathbf{u}'n^{-1/2}S_n(\boldsymbol{\theta}_0) - \sqrt{n}\sum_{k=1}^{k_0}\sum_{j=1}^{p_k}\lambda_{kj}|u_{kj}| + o_p(1) \\
&= (2\sqrt{3})^{-1}\mathbf{u}'[n^{-1}\mathbf{X}'\mathbf{W}\mathbf{X}]\mathbf{u} - \mathbf{u}'O_p(1) - \sqrt{n}\sum_{k=1}^{k_0}\sum_{j=1}^{p_k}\lambda_{kj}|u_{kj}| + o_p(1) \\
&\geq (2\sqrt{3})^{-1}\mathbf{u}'[n^{-1}\mathbf{X}'\mathbf{W}\mathbf{X}]\mathbf{u} - \mathbf{u}'O_p(1) - k_0\sqrt{n}a_n(\|\mathbf{u}\|) + o_p(1)
\end{aligned}$$

Note that $n^{-1}\mathbf{X}'\mathbf{W}\mathbf{X}\xrightarrow{\mathcal{P}}\mathbf{C}$, a positive definite matrix, and $\sqrt{n}a_n\xrightarrow{\mathcal{P}}0$. Therefore, for n sufficiently large, the first term on the right-hand side of the inequation above dominates. $G_n(\mathbf{u})$ can be made positive when C is chosen to be sufficiently large.

We now prove part (b). Suppose that $\widehat{\boldsymbol{\theta}}_b \neq 0 \ \forall \ n \in \mathbb{N}$. Let k be such that $k_0 < k \leq K$ and

$\widehat{\theta}_{kj} \neq 0$ for some j such that $1 \leq j \leq p_k$. Since $Q_n(\boldsymbol{\theta})$ is differentiable at any point, except the origin, $\widehat{\theta}_{kj}$ must be solution of the equation

$$0 = n^{-3/2}\sum_{i<j} b_{ij}(\mathbf{x}_{ik} - \mathbf{x}_{jk})sgn((y_i - y_j) - (\mathbf{x}_i - \mathbf{x}_j)'\boldsymbol{\theta}) + \sqrt{n}\lambda_{kj}sgn(\theta_{kj}).$$

Now, by the consistency of $\widehat{\boldsymbol{\theta}}_n$ and part (ii.) of lemma 5.1, the first term of the right hand side of the equation above is $O_p(1)$. In addition, $\sqrt{n}b_n\xrightarrow{\mathcal{P}}\infty$ implies that $\sqrt{n}\lambda_{kj}\xrightarrow{\mathcal{P}}\infty$. So the equation does not hold for large values of n, as we assume that $\widehat{\theta}_{kj} \neq 0$. Therefore, $\widehat{\boldsymbol{\theta}}_b\xrightarrow{\mathcal{P}}0$.

The proof of part (c) is identical to the one of Wang and Li (2009) and will therefore be omitted here.

□

Bibliography

Abarin, T. and Wang, L. (2012). Instrumental variable approach to covariate measurement error in generalized linear models. *Ann. Inst. Statist. Math.*, 64(3):475–493.

Abebe, A. and McKean, J. W. (2007). Highly efficient nonlinear regression based on the Wilcoxon norm. In Umbach, D., editor, *Festschrift in Honor of Mir Masoom Ali*, pages 340–357.

Adichie, J. N. (1967). Estimates of regression parameters based on rank tests. *Ann. Math. Statist.*, 38:894–904.

Andrews, D. and Herzberg, A. (1985). *Data: a collection of problems from many fields for the student and research worker*. Springer series in statistics. Springer.

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations. *J. Amer. Statist. Assoc.*, 96(455):939–967. With discussion and a rejoinder by the authors.

Augustin, N. H., Sauleau, E.-A., and Wood, S. N. (2012). On quantile quantile plots for generalized linear models. *Comput. Statist. Data Anal.*, 56(8):2404–2409.

Bandyopadhyay, S., Ganguli, B., and Chatterjee, A. (2011). A review of multivariate longitudinal data analysis. *Stat. Methods Med. Res.*, 20(4):299–330.

Bindele, H. F. and Abebe, A. (2012). Bounded influence nonlinear signed-rank regression. *Canad. J. Statist.*, 40(1):172–189.

Brunner, E. and Denker, M. (1994). Rank statistics under dependent observations and applications to factorial designs. *J. Stat. Plann. Inference*, 42(3):353–378.

Cantoni, E. (2004). A robust approach to longitudinal data analysis. *Canad. J. Statist.*, 32(2):169–180.

Chen, K., Hu, I., and Ying, Z. (1999). Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *Ann. Statist.*, 27(4):1155–1163.

Chen, Y. and Hero, A. O. (2012). Recursive $\ell_{1,\infty}$ group lasso. *IEEE Trans. Signal Process.*, 60(8):3978–3987.

Cheng, G., Yu, Z., and Huang, J. Z. (2013). The cluster bootstrap consistency in generalized estimating equations. *J. Multivariate Anal.*, 115:33–47.

Chiou, J.-M. and Müller, H.-G. (1999). Nonparametric quasi-likelihood. *Ann. Statist.*, 27(1):36–64.

Collins, J. R. and Portnoy, S. L. (1981). Maximizing the variance of M-estimators using the generalized method of moment spaces. *Ann. Statist.*, 9(3):567–577.

Copas, A. J. and Seaman, S. R. (2010). Bias from the use of generalized estimating equations to analyze incomplete longitudinal binary data. *J. Appl. Stat.*, 37(5-6):911–922.

den Boer, A. and Zwart, B. (2012). Mean square convergence rates for maximum quasi-likelihood estimators. `http://homepages.cwi.nl/~boer/statpaper28Sep2012.pdf`. Unpublished manuscript, Accessed: Oct 10, 2012.

Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.*, 13(1):342–368.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360.

Gao, Q.-B., Lin, J.-G., Zhu, C.-H., and Wu, Y.-H. (2012). Asymptotic properties of maximum quasi-likelihood estimators in generalized linear models with adaptive designs. *Statistics*, 46(6):833–846.

Hallin, M. and Ingenbleek, J.-F. (1983). The swedish automobile portfolio in 1977: a statistical study. *Scandinavian Actuarial Journal*, 83:49 – 64.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, 69:383–393.

Hardin, J. W. and Hilbe, J. M. (2012). *Generalized linear models and extensions*. Stata Press, College Station, TX, third edition.

He, X., Zhu, Z.-Y., and Fung, W.-K. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika*, 89(3):579–590.

Hettmansperger, T. P. (1984). *Statistical inference based on ranks*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York.

Hettmansperger, T. P. and McKean, J. W. (1998). *Robust nonparametric statistical methods*, volume 5 of *Kendall's Library of Statistics*. Edward Arnold, London.

Hin, L.-Y., Carey, V. J., and Wang, Y.-G. (2007). Criteria for working-correlation-structure selection in GEE: assessment via simulation. *Amer. Statist.*, 61(4):360–364.

Hirose, K. and Konishi, S. (2012). Variable selection via the weighted group lasso for factor analysis models. *Canad. J. Statist.*, 40(2):345–361.

Huber, P. J. (1981). *Robust statistics*. John Wiley & Sons Inc., New York. Wiley Series in Probability and Mathematical Statistics.

Huggins, R. M. (1993). A robust approach to the analysis of repeated measures. *Biometrics*, 49(3):715–720.

Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *Ann. Math. Statist.*, 43:1449–1458.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.*, 40:633–643.

Johnson, B. A. and Peng, L. (2008). Rank-based variable selection. *J. Nonparametr. Stat.*, 20(3):241–252.

Jung, S.-H. and Ying, Z. (2003). Rank-based regression with repeated measurements data. *Biometrika*, 90(3):732–740.

Jurečková, J. (1971). Nonparametric estimate of regression coefficients. *The Annals of Mathematical Statistics*, 42(4):pp. 1328–1338.

Klar, B. and Meintanis, S. G. (2012). Specification tests for the response distribution in generalized linear models. *Comput. Statist.*, 27(2):251–267.

Klein, R. and Yohai, V. J. (1981). Asymptotic behavior of iterative M-estimators for the linear model. *Comm. Statist. A—Theory Methods*, 10(23):2373–2388.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.*, 28(5):1356–1378.

Kordzakhia, N., Mishra, G. D., and Reiersølmoen, L. (2001). Robust estimation in the logistic regression model. *J. Statist. Plann. Inference*, 98(1-2):211–223.

Künsch, H. R., Stefanski, L. A., and Carroll, R. J. (1989). Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *J. Amer. Statist. Assoc.*, 84(406):460–466.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Liu, T. and Yuan, X. (2012). Combining quasi and empirical likelihoods in generalized linear models with missing responses. *J. Multivariate Anal.*, 111:39–58.

Mbachu, H. I., Nduka, E. C., and Nja, M. E. (2012). Designing a pseudo R-squared goodness-of-fit measure in generalized linear models. *J. Math. Res.*, 4(2):148–154.

McCullagh, P. A. and Nelder, J. A. (1989). *Generalized linear models.* Chapman & Hall, Boca Raton Fl., 2nd. ed. edition.

Morgenthaler, S. (1992). Least-absolute-deviations fits for generalized linear models. *Biometrika*, 79(4):pp. 747–754.

Nakai, M. and Ke, W. (2011). Review of the methods for handling missing data in longitudinal data analysis. *Int. J. Math. Anal. (Ruse)*, 5(1-4):1–13.

Naranjo, J. D. and Hettmansperger, T. P. (1994). Bounded influence rank regression. *J. Roy. Statist. Soc. Ser. B*, 56(1):209–220.

Nardi, Y. and Rinaldo, A. (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electron. J. Stat.*, 2:605–633.

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

Plaisance, E. P., Taylor, J. K., Alhassan, S., Abebe, A., Mestek, M. L., and Grandjean, P. W. (2007). Cardiovascular fitness and vascular inflammatory markers after acute aerobic exercise. *International journal of sport nutrition and exercise metabolism*, 17(2):152.

Pollard, D. (1990). *Empirical processes: theory and applications.* NSF-CBMS Regional Conference Series in Probability and Statistics, 2. Institute of Mathematical Statistics, Hayward, CA.

Prakasa Rao, B. L. S. (1981). Asymptotic behavior of M-estimators for the linear model with dependent errors. *Bull. Inst. Math. Acad. Sinica*, 9(3):367–375.

Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrics*, 38(2):pp. 485–498.

Qaqish, B. F. and Preisser, J. S. (1999). Resistant fits for regression with correlated outcomes: an estimating equations approach. *J. Statist. Plann. Inference*, 75(2):415–431. The Seventh Eugene Lukacs Conference (Bowling Green, OH, 1997).

Ridker, P. M., Rifai, N., Rose, L., Buring, J. E., and Cook, N. R. (2002). Comparison of c-reactive protein and low-density lipoprotein cholesterol levels in the prediction of first cardiovascular events. *New England Journal of Medicine*, 347(20):1557–1565.

Schrader, R. M. and Hettmansperger, T. P. (1980). Robust analysis of variance based upon a likelihood ratio criterion. *Biometrika*, 67(1):93–101.

She, Y. (2012). An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Comput. Statist. Data Anal.*, 56(10):2976–2990.

Sievers, G. L. (1983). A weighted dispersion function for estimation in linear models. *Comm. Statist. A—Theory Methods*, 12(10):1161–1179.

Simon, N. and Tibshirani, R. (2012). Standardization and the group Lasso penalty. *Statist. Sinica*, 22(3):983–1001.

Song, L., Hu, H., and Cheng, X. (2012). Hypothesis testing in generalized linear models with functional coefficient autoregressive processes. *Math. Probl. Eng.*, pages Art. ID 862398, 19.

Stefanski, L. A., Carroll, R. J., and Ruppert, D. (1986). Optimally bounded score functions for generalized linear models with applications to logistic regression. *Biometrika*, 73(2):pp. 413–424.

Tang, C. Y. and Leng, C. (2011). Empirical likelihood and quantile regression in longitudinal data analysis. *Biometrika*, 98(4):1001–1006.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288.

Tsai, M.-Y., Wang, J.-F., and Wu, J.-L. (2011). Generalized estimating equations with model selection for comparing dependent categorical agreement data. *Comput. Statist. Data Anal.*, 55(7):2354–2362.

Wang, H. and Leng, C. (2008). A note on adaptive group lasso. *Comput. Statist. Data Anal.*, 52(12):5277–5286.

Wang, L. and Li, R. (2009). Weighted Wilcoxon-type smoothly clipped absolute deviation method. *Biometrics*, 65(2):564–571.

Wang, Y.-G. and Carey, V. J. (2004). Unbiased estimating equations from working correlation models for irregularly timed repeated measures. *J. Amer. Statist. Assoc.*, 99(467):845–853.

Wang, Y.-G. and Hin, L.-Y. (2010). Modeling strategies in longitudinal data analysis: covariate, variance function and correlation structure selection. *Comput. Statist. Data Anal.*, 54(12):3359–3370.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61(3):pp. 439–447.

Welsh, A. H. and Richardson, A. M. (1997). Approaches to the robust estimation of mixed models. In *Robust inference*, volume 15 of *Handbook of Statist.*, pages 343–384. North-Holland, Amsterdam.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67.

Zhang, X. (2011). *Generalized Estimating Equations and Gaussian Estimation in Longitudinal Data Analysis*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–University of Windsor (Canada).

Zhao, P., Rocha, G., and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.*, 37(6A):3468–3497.

Zhou, N. and Zhu, J. (2010). Group variable selection via a hierarchical lasso and its oracle property. *Stat. Interface*, 3(4):557–574.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320.

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist.*, 37(4):1733–1751.