Semiparametric Methods for the Generalized Linear Model

Jinsong Chen

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

George R. Terrell, Co-Chairman
Inyoung Kim, Co-chairman
Jeffrey B. Birch
Pang Du
Leanna L. House
Eric P. Smith

May 28, 2010
Blacksburg, Virginia

Keywords: Generalized linear model; Generalized linear mixed model; Penalized splines;
Single-Index model

Semiparametric Methods for the Generalized Linear Model

Jinsong Chen

(ABSTRACT)

The generalized linear model (GLM) is a popular model in many research areas. In the GLM, each outcome of the dependent variable is assumed to be generated from a particular distribution function in the exponential family. The mean of the distribution depends on the independent variables. The link function provides the relationship between the linear predictor and the mean of the distribution function. In this dissertation, two semiparametric extensions of the GLM will be developed. In the first part of this dissertation, we have proposed a new model, called a semiparametric generalized linear model with a log-concave random component (SGLM-L). In this model, the estimate of the distribution of the random component has a nonparametric form while the estimate of the systematic part has a parametric form. In the second part of this dissertation, we have proposed a model, called a generalized semiparametric single-index mixed model (GSSIMM). A nonparametric component with a single index is incorporated into the mean function in the generalized linear mixed model (GLMM) assuming that the random component is following a parametric distribution.

In the first part of this dissertation, since most of the literature on the GLM deals with the parametric random component, we relax the parametric distribution assumption for the random component of the GLM and impose a log-concave constraint on the distribution. An iterative numerical algorithm for computing the estimators in the SGLM-L is developed. We construct a log-likelihood ratio test for inference. In the second part of this dissertation, we use a single index model to generalize the GLMM to have a linear combination of covariates enter the model via a nonparametric mean function, because the linear model in the GLMM is not complex enough to capture the underlying relationship between the response and its associated covariates. The marginal likelihood is approximated using the Laplace method. A penalized quasi-likelihood approach is proposed to estimate the nonparametric function and parameters including single-index coefficients in the GSSIMM. We estimate variance components using marginal quasi-likelihood. Asymptotic properties of the estimators are developed using a similar idea by Yu (2008). A simulation example is carried out to compare the performance of the GSSIMM with that of the GLMM. We demonstrate the advantage of my approach using a study of the association between daily air pollutants and daily mortality adjusted for temperature and wind speed in various counties of North Carolina.

Acknowledgments

I have worked with a great number of people whose contribution in assorted ways to the research and the making of the dissertation deserve special mention. It is a pleasure to convey my gratitude to them all in my humble acknowledgment.

First, I thank my advisor George R. Terrell for his continuous support. He has been a tremendous support in our hours of meetings and in our communication using email when I am not on campus. He showed me different ways to approach a research problem and the need to be persistent to accomplish any goal. Above all and the most needed, he provided me unflinching encouragement and support in various ways.

A special thanks goes to my co-advisor, Inyoung Kim, who is responsible for helping me complete the writing of this dissertation as well as the challenging research that lies behind it. She helped me understand the process of writing academic papers. She was always there to meet and talk about my ideas, to proofread and mark up my papers and chapters.

Besides my advisors, I would like to thank the rest of my dissertation committee: Eric P. Smith, who gave me good suggestions in the study of the health effects of air pollution; Jeffrey B. Birch, for his friendship, encouragement, hard questions and help with administrative issues; Pang Du, who gave insightful comments; and Leanna House for reviewing my work on a very short notice.

I acknowledge the help of Mike Box with computer questions and for resolving computer problems. He always tried his best to help me with computational problems I faced. The simulations in this dissertation would have been impossible without his help.

Last, but not least, I thank my family: my parents, Qingen Chen, and Wenfang Zhu, for giving me life in the first place, for educating me with aspects from both arts and sciences, for unconditional support and encouragement to pursue my interests, even when the interests went beyond boundaries of language, field and geography; my sister, Huaming Chen, for her understanding and believing in me; my brother, Hesong Chen, who partially takes my responsibilities in my family, as a son for my mother and a brother for my sister.

—- Jinsong Chen

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Generalized Linear Model

The generalized linear model (GLM, see McCullagh and Nelder, 1989) is a popular model in many research areas. There are three components in the GLM: the random component, the systematic component, and the link function. In the GLM, each outcome of the dependent variable is assumed to be generated from a particular distribution function in the exponential family. The mean of the distribution depends on the independent variables. The link function provides the relationship between the linear predictor and the mean of the distribution function.

The GLM is a flexible generalization of ordinary least squares regression. In this section, we introduce each component of the GLM using the normal linear model, which is a special case of the GLM, then extend the normal linear model results to the GLM framework. A vector of observations $\mathbf{y}$ having n components is assumed to be a realization of a random variable $\mathbf{Y}$ whose components are independently distributed with mean $\boldsymbol{\mu}$. Let $\mathbf{X}$ be the n by p model matrix. The systematic part of the model is a specification for the vector $\boldsymbol{\mu}$ in terms of a small number of unknown parameters $\beta_1, \cdots, \beta_p$. In the case of ordinary linear models, this specification takes the form $\boldsymbol{\mu} = \sum_{j=1}^{p} \mathbf{x}_j \beta_j$. Here, $\mathbf{x}_j$ is the vector of size n associated with the jth covariate. In matrix notation we write $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the p by 1 vector of parameters.

The classical linear model may be summarized in the following form. The components of $\mathbf{Y}$ are independent normal variables with constant variance σ^2 and

$$\mathbf{E}(\mathbf{Y}) = \boldsymbol{\mu} \quad \text{where} \quad \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}. \tag{1.1}$$

To simplify the transition to the generalized linear model, we rearrange (1.1) to produce the following three-part specification:

1. The random component: the components of $\mathbf{Y}$ are independent, normally distributed random variables with $\mathbf{E}(\mathbf{Y}) = \boldsymbol{\mu}$ and constant variance σ^2;

2. The systematic component: covariates $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p$ produce a linear predictor $\boldsymbol{\zeta} = \sum_{j=1}^{p} \mathbf{x}_j \beta_j$;

3. The link between the random and systematic components is just the identity function: $\boldsymbol{\mu} = \boldsymbol{\zeta}$.

This generalization introduces a new symbol $\boldsymbol{\zeta}$ for the linear predictor and the third component then specifies that $\boldsymbol{\mu}$ and $\boldsymbol{\zeta}$ happen to be identical. If we write $\boldsymbol{\zeta}_i = g(\boldsymbol{\mu}_i)$, the function $g(\cdot)$ will be called the link function. The generalized linear model allows two extensions of the classical linear model; first the distribution of the random component must come from a distribution within the exponential family, and secondly the link function must be a monotonic differentiable function (Green and Silverman, 1994; McCullagh and Nelder, 1989).

In the generalized linear model, the conditional distribution of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ is a distribution in the exponential family, taking the form

$$f_{Y|\mathbf{X}}(y|\mathbf{x}) = \exp[\{y\theta(\mathbf{x}) - \psi\{\theta(\mathbf{x})\}\}/a(\phi) + C(y, \phi)] \tag{1.2}$$

for some specific functions $\psi(\cdot)$, $a(\cdot)$ and $C(\cdot)$. Here, θ is called the natural parameter, and ϕ is called the dispersion parameter. In this dissertation, we assume $a(\phi) = 1$ for all cases.

Let us denote $\psi'\{\theta(\mathbf{x})\}$ as the first derivative of $\psi\{\theta(\mathbf{x})\}$ with respect to $\theta(\mathbf{x})$. In the parametric generalized linear model, the unknown regression function $\mu(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x}) = \psi'\{\theta(\mathbf{x})\}$ is modeled linearly via a link function g by

$$g\{\mu(\mathbf{x})\} = \mathbf{x}^T\boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ is the vector of parameters. The link function relates the linear predictor ζ to the expected value μ of a datum y. If $g = \theta$, where θ is the canonical parameter as defined in (1.2), then g is the canonical link function. For example, the logit link, $\log\{\mu/(1-\mu)\}$, is the canonical link for a binomial distribution, and the log link, $\log(\mu)$, is the canonical link for a Poisson distribution. Although canonical links lead to desirable statistical properties of the model, the form of $g(\cdot)$ is not restricted to the canonical form. For an example of these properties, the canonical links allow for $\mathbf{X}^T\mathbf{Y}$ to be a sufficient statistic for $\boldsymbol{\beta}$ (McCullagh and Nelder, 1989). It implies that $\mathbf{X}^T\mathbf{Y}$ is a statistic which captures all the information about $\boldsymbol{\beta}$ contained in the sample. The GLM finds extensive use in the applied literature, but recent emphasis has been on the generalized linear mixed models (Breslow and Clayton, 1993)

1.2 Generalized Linear Mixed Models

Generalized linear mixed models (GLMMs) extend generalized linear models by allowing for the incorporation of random effects. In the GLMMs framework, a response Y is to be predicted by covariates $(\mathbf{X}, \mathbf{Z})$, where $\mathbf{X}$ is a vector-valued covariate associated with fixed effects, and $\mathbf{Z}$ is a vector-valued covariate associated with random effects. Assuming $a(\phi) = 1$ in equation (1.2), the exponential family model in section 1.1 can be extended in the following expression

$$f_{Y|\mathbf{X},\mathbf{Z}}(y|\mathbf{x}, \mathbf{z}) = \exp[y\theta(\mathbf{x}, \mathbf{z}) - \psi\{\theta(\mathbf{x}, \mathbf{z})\} + C(y)], \tag{1.3}$$

where $\psi(\cdot)$ and $C(\cdot)$ are known functions, and θ is the natural parameter. In the parametric generalized linear mixed models, the unknown regression function $\mu(\mathbf{x}, \mathbf{z}) = E(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = \psi'\{\theta(\mathbf{x}, \mathbf{z})\}$ is modeled linearly via a link function g such that

$$g\{\mu(\mathbf{x}, \mathbf{z})\} = \mathbf{x}^T\boldsymbol{\beta} + \mathbf{z}^T\boldsymbol{b}$$

where $\boldsymbol{\beta}$ is an unknown vector of parameters, and $\boldsymbol{b}$ is the random effect. It is often a reasonable approximation and certainly traditional to assume that the random effects have a multivariate normal distribution $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$ whose variance components, contained in $\boldsymbol{\Sigma}$, are to be estimated from the data.

The generalized linear mixed models are useful for accommodating the overdispersion often observed among outcomes; or for modeling the dependence among outcome variables inherent in longitudinal or repeated measure data; or for analyzing clustered data when subjects are observed nested within a larger unit (Diggle *et al.*, 2002; Jiang, 2007). Use of the GLMMs can be found in a broad variety of applications. For an example of the GLMMs, we consider the daily effect of $PM_{2.5}$, the fine Particulate Matter with a diameter smaller than 2.5 microns, and ozone concentration on mortality using data from twelve counties in North Carolina in the years 2004 and 2005. We let the response observation y_{ij} be the jth daily mortality count with major cause of death as cardiovascular disease in the ith county. Define $\mathbf{x}_{ij}$ as the vector of the jth daily observation for independent variables in the ith county, including $PM_{2.5}$, ozone, average temperature, and average wind speed. The random effect for the ith county is denoted as $\mathbf{z}_i^T\mathbf{b}$, with random effect variable $\mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{b})$, and a vector, $\mathbf{z}_i$, associated with random effect for the ith county. Denote m as the total number of counties in the study. $\mathbf{z}_i$ is $m \times 1$ vector with the ith element being 1 and all other element being 0. It is a classical way to fit a log-linear mixed model as

$$\log\{\mu(\mathbf{x}, \mathbf{z})\} = \mathbf{x}^T\boldsymbol{\beta} + \mathbf{z}^T\mathbf{b}.$$

The details of the GLMM analysis will be presented as well as an analysis by the new method that we introduce in chapter 3.

In the GLMMs, when the outcomes come in the form of proportions or counts, a full maximum likelihood analysis based on their joint marginal distribution (section 3.3) requires numerical integration techniques for calculation of the log-likelihood, score equations, and

information matrix. This method has been implemented successfully in relatively simple problems involving binomial (Crouch and Spiegelman, 1990) and Poisson (Hinde, 1982) mixtures with a high degree of independence among the observations. In this dissertation, approximate inference based on quasi-likelihood (section 3.3) is used for estimation.

1.3 Single-Index Model

Single-index models (Stoker, 1986; Hardle *et al.*, 1993) generalize linear regression. A linear regression for the dependence of a scalar variable Y and p-vector $\mathbf{x}$ has the form $Y = \mathbf{x}^T\boldsymbol{\alpha}_0 + \varepsilon$, where $\boldsymbol{\alpha}_0$ is the p-vector of unknown parameters and ε is a random variable with zero mean conditional on $\mathbf{x}$. More generally, we might define $Y = \eta(\mathbf{x}^T\boldsymbol{\alpha}_0) + \varepsilon$, where $\eta(\cdot)$ is an unknown univariate function. This is called a single-index model. The scale of $\mathbf{x}^T\boldsymbol{\alpha}_0$ in $\eta(\mathbf{x}^T\boldsymbol{\alpha}_0)$ may be determined arbitrarily, and so we may replace $\boldsymbol{\alpha}_0$ by the unit vector $\boldsymbol{\beta} = \boldsymbol{\alpha}_0||\boldsymbol{\alpha}_0||^{-1}$, where $||\cdot||$ denotes the Euclidean metric. The aim is to estimate both η and $\boldsymbol{\beta}$ in the equivalent model

$$Y = \eta(\mathbf{x}^T\boldsymbol{\beta}) + \varepsilon.$$

Single-index models are an important tool in multivariate nonparametric regression. The linear combination $\mathbf{x}^T\boldsymbol{\beta}$ is replaced with a nonparametric component, $\eta(\mathbf{x}^T\boldsymbol{\beta})$. By reducing the dimensionality from multivariate predictors to a univariate index $\mathbf{x}^T\boldsymbol{\beta}$, single-index models avoid the so-called "curse of dimensionality" (Bellman, 1961) while still capturing important features in high-dimensional data. The "curse of dimensionality" is a term applied to the problem caused by the rapid increase in volume associated with adding extra dimensions to a mathematical space. Single-index models avoid the "curse of dimensionality" because the index $\mathbf{x}^T\boldsymbol{\beta}$ aggregates the dimensions of $\mathbf{x}$.

Typical inferential procedures for single-index models follow two-steps, estimating $\boldsymbol{\beta}$ first to form $z = \mathbf{x}^T\boldsymbol{\beta}$ and then η. Hardle *et al.* (1993) considered a kernel estimator of η. Li and Duan (1989) proposed an alternative method that uses "sliced inverse regression" to estimate the index parameter. Extensions to the generalized partially linear single-index models using splines were put forward by Carroll *et al.* (1997), and by Yu and Ruppert (2002).

1.4 Dissertation Outline

In this dissertation, two semiparametric extensions of the GLM will be developed. In the first, the estimate of the distribution of the random component has a nonparametric form while the estimate of the systematic part has a parametric form. The corresponding model developed is called a semiparametric generalized linear model with a log-concave component (SGLM-L). In the second, the single-index model is incorporated into the GLM and further

extended to GLMM assuming that the random component follows a parametric distribution. Thus, we have a semiparametric estimate of the systematic part. The model of interest is called a generalized semiparametric single-index mixed model (GSSIMM).

The outline of the dissertation is as follows. In chapter 2, we propose a semiparametric generalized linear model with a log-concave random component. Chapter 3 covers the generalized semiparametric single-index mixed model. In both chapters, the literature review, model description, estimation method, inference, numerical results and a summary will be presented for each model respectively. Finally, in chapter 4, we present a summary of our conclusions and discuss open research questions.

Chapter 2

Semiparametric Generalized Linear Model with a Log-concave Random Component

2.1 Introduction

In this section, we will review the literature related to the semiparametric generalized linear model with a log-concave random component (SGLM-L). This model will be introduced in section 2.2.

A probability density f on the real line of y is called log-concave if it is written as $f(y) = \exp\{\varphi(y)\}$ for some concave function $\varphi : \mathbb{R} \to (-\infty, \infty)$. The log-concave density includes many well-known parametric classes. Families of distributions that always have log-concave density functions include the uniform distribution, the normal distribution, the logistic distribution, the chi-square distribution, the chi distribution, the exponential distribution, and the Laplace distribution. Some families of distributions have log-concave density functions for certain parameter values and not for others. Such families include the gamma distribution, the beta distribution, the Weibull distribution, and the power function distribution. Table 2.1 shows the density functions and parameter spaces that ensure log-concave property for these families of distributions. The Students t distribution, the Cauchy distribution, the F distribution, the Pareto distribution, and the log-normal distribution do not have log-concave density functions for any parameter values.

It is known that the nonparametric maximum likelihood estimator (NPMLE) of a log-concave density is always a piecewise linear function with at most as many knots as observations, but typically fewer (Rufibach and Dumbgen, 2004). An example of NPMLE of a log-concave density is shown in Figure 2.1 (see section 2.5.1). It is shown that this property

Distribution	Density Function	Log-concave
gamma	$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{x\beta}$	$\beta \geq 1$
beta	$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$\alpha \geq 1$ and $\beta \geq 1$
Weibull	$f(x) = cx^{c-1}e^{-x^c}$	$c \geq 1$
power function	$f(x) = \beta x^{\beta-1}$	$\beta \geq 1$

Table 2.1: Distribution families having log-concave density functions for certain parameter values and not for others.

can be exploited to design a linearly constrained optimization problem whose iteratively calculated solution yields the estimator (Rufibach, 2007).

Finding a statistical estimator typically requires maximization of an objective function, *e.g.* a (log-)likelihood. Often such a problem can be embedded in the framework of a linearly constrained optimization problem; see Groeneboom *et al.* (2003) for a short survey on recent developments in the use of optimization algorithms in statistics. It has been shown that the NPMLE of a log-concave density exists and is unique (Rufiback, 2007). Moreover, the NPMLE is a piecewise linear continuous function with changes of slope only at order statistics drawn from data. This will be explained in more detail in section 2.3.1. Computational tools, such as the iterative convex minorant algorithm, the active set algorithm (Groeneboom *et al.*, 2001) and the expectation maximization algorithm (EM) among many others, have been used to obtain the NPMLE of a log-concave density. An active set algorithm is a useful tool from optimization theory with many potential applications in statistical computing (Fletcher, 1987). Dumbgen (2007) proposed an active set algorithm to compute the maximum likelihood estimator of a log-concave density (see section 2.3.2). It is similar in spirit to the vertex direction and vertex reduction algorithms described by Groeneboom *et al.* (2001), who consider the special setting of mixture models.

The outline of this chapter is as follows. We introduce the semiparametric generalized linear model with a log-concave random component in section 2.2. The estimation method is proposed in section 2.3. Section 2.4 explains how to make an inference using the log-likelihood ratio test. A simulation study and a data application are presented in section 2.5. A brief conclusion is drawn in the final section.

2.2 Model

In the classical GLM modeling framework, the response Y is assumed to follow a distribution within the exponential family and can be written as the following expression

$$f_Y(y;\theta) = \exp\{y\theta - \psi(\theta) + C(y)\}$$

for specific functions $\psi(\cdot)$ and $C(\cdot)$. For known canonical parameter θ, the density for Y conditional on θ is $f_Y(y;\theta)$. The density f may be with respect to either Lebesgue measure or counting measure. Consider the probability density $f_Y(y;\theta)$ with moment generating function $M(t) = E\{\exp(Yt)\}$ and cumulative generating function $K(t) = \log M(t)$. For any value of t for which the moment generating function M is defined, it is immediately apparent that

$$f_Y(y;\theta+t) = \exp\{\varphi(y;\theta) + yt - K(t)\} = f(y;\theta)\exp\{yt - K(t)\}$$

is also a density, where $\varphi(y;\theta) = \log f_Y(y;\theta)$. This method is generally referred as "exponential tilting". To make the density of $(y;\theta+t)$ proper, $f(y;\theta)\exp\{yt - K(t)\}$ is divided by an integral which is a constant in terms of y. That is

$$f_Y(y;\theta+t) = \frac{f(y;\theta)\exp\{yt - K(t)\}}{\int f(y;\theta)\exp\{yt - K(t)\}dy} = \frac{f(y;\theta)\exp(yt)}{\int f(y;\theta)\exp(yt)dy}. \tag{2.1}$$

Taking a log on the both sides of equation (2.1), we have $\varphi(y;\theta+t) = \varphi(y;\theta) + yt - \text{constant}$. Thus, the log-density of $(y;\theta+t)$ is shifted from the log-density of $\varphi(y;\theta)$. If we have a baseline estimate of $\varphi(y;\theta)$, then it is easy to estimate $(y;\theta+t)$. Based on this concept, we develop the semiparametric generalized linear model with a log-concave random component.

Applying the concept of "exponential tilting" (2.1), our model has the form

$$f(y_i;\mathbf{x}_i,\boldsymbol{\beta}) = \frac{f(y_i)\exp(y_i\mathbf{x}_i^T\boldsymbol{\beta})}{\int f(y)\exp(y\mathbf{x}_i^T\boldsymbol{\beta})dy} \tag{2.2}$$

where $Y = \{y_1,\ldots,y_n\}^T$ is the vector of response variables, $X = \{\mathbf{x}_1,\ldots,\mathbf{x}_p\}$ is a $n \times p$ matrix of covariates, and $\boldsymbol{\beta}$ is the vector of unknown parameters of interest.

In model (2.2), the unknown distribution function $f(\cdot)$ is estimated via a nonparametric method while the linear predictor has a parametric form. Thus, model (2.2) is a semiparametric extension of classical GLM. In the context of GLM, most of the literature deals with a parametric random component, i.e. the random component follows a parametric distribution, such as binomial, Poisson, and gamma. There are corresponding link functions for these parametric distributions. We relax the parametric assumption for the distribution of the random component in the GLM and impose a log-concave constraint. We consider a GLM model without assuming the form of the link function but with the log-concave constraint for the distribution function of the random component. The major reasons to use this strategy are flexibility and robustness since the SGLM-L is not required to assume a specific distribution. The SGLM-L may be used for various data sets whose responses follow a distribution with the log-concave property. This concave shape constraint is a basic assumption in model (2.2). Note that if $f(y)$ is log-concave, so is $f(y;\mathbf{x})$ for any x.

We have noted that the nonparametric maximum likelihood estimator for the log-concave density is always a piecewise linear function (Rufibach and Dumbgen, 2004). This property establishes a nice connection between exponential tilting and the SGLM-L. When we estimate

the random part (for some fixed values of the covariates) by finding the maximum likelihood solution under the log-concave constraint, the solution is a density whose logarithm is piecewise linear. This means that when we apply exponential tilting to such a density, the tilted density still has a piecewise linear logarithm. Thus, we have found a natural exponential family in which to carry out the estimation of the effects of covariates. For instance, the NPMLE of the log-density of the random part given the fixed value of covariates, $\mathbf{x} = \mathbf{0}$, is a piecewise linear function, $\hat{\varphi}(y; \mathbf{0}, \boldsymbol{\beta})$. The NPMLE of a log-density given $\mathbf{x} \neq 0$ is still piecewise linear because $\hat{\varphi}(y; \mathbf{x}, \boldsymbol{\beta}) = \hat{\varphi}(y; \mathbf{0}, \boldsymbol{\beta}) + y\mathbf{x}^T\boldsymbol{\beta} + \text{constant}$.

2.3 Method

2.3.1 Likelihood Function

Let $\mathbf{Y}$ be the vector of independent response variables with a log-concave distribution, and $\mathbf{X}$ be the matrix of covariates. We observe $(y_1, \ldots, y_N)^T$, the vector of N observed outcomes for a response variable $\mathbf{Y}$. It happens sometimes that $y_i = y_j$ for $i \neq j$. The data are arranged as strictly increasing order statistics $y_{(1)} < y_{(2)} < \ldots < y_{(n)}$ drawn from the observed outcomes. It is clear that $n \leq N$. We define the probability weight $p_i = N^{-1} \times$ the number of$\{j : y_j = y_{(i)}\}$. For instance, if there are 3 observed y's having the same value of $y_{(i)}$, the weight is $p_i = 3/N$. In general, most of the weights have the value of $1/N$.

In the semiparametric generalized linear model with a log-concave component, our goal is to estimate the underlying density function $f(\cdot)$ nonparametrically, or equivalently estimate $\varphi(\cdot) = \log(f)$ nonparametrically, and estimate the parameter vector $\boldsymbol{\beta}$ using the order statistics $y_{(1)} < y_{(2)} < \ldots < y_{(n)}$ and probability weights $p_1, p_2, \ldots, p_n > 0$, i.e. $\sum_{i=1}^{n} p_i = 1$.

Based on model (2.2), we can construct a log-likelihood function for $(\boldsymbol{\varphi}, \boldsymbol{\beta})$ of the form

$$L(\boldsymbol{\varphi}, \boldsymbol{\beta}) = \sum_i p_i[\varphi\{y_{(i)}\} + y_{(i)}\mathbf{x}_i^T\boldsymbol{\beta}] - \sum_i p_i \log \int \exp\{\varphi(y) + y\mathbf{x}_i^T\boldsymbol{\beta}\}dy \qquad (2.3)$$

where, $\boldsymbol{\varphi} = (\varphi\{y_{(1)}\}, \varphi\{y_{(2)}\}, \ldots, \varphi\{y_{(n)}\})^T$. The log density $\varphi(\cdot) = \log(f)$ is estimated using a nonparametric method; thus the vector $\boldsymbol{\varphi}$ is sufficient to determine the function because the nonparametric maximum likelihood estimator for the log-concave density is always a piecewise linear function (Rufibach and Dumbgen, 2004).

2.3.2 Estimation Procedure

We perform the following procedure to obtain the maximum likelihood estimators of $(\boldsymbol{\varphi}, \boldsymbol{\beta})$ in the semiparametric generalized linear model with a log-concave random component,

Step 0. Initialize $\hat{\boldsymbol{\beta}}$. These initial values are obtained using the normal linear model in this dissertation.

Step 1. For given $\hat{\boldsymbol{\beta}}$, find $\hat{\boldsymbol{\varphi}}$ by maximizing

$$Ł(\boldsymbol{\varphi}) = \sum_i p_i[\varphi\{y_{(i)}\} + y_{(i)}\mathbf{x}_i^T\hat{\boldsymbol{\beta}}] - \sum_i p_i \log \int \exp\{\varphi(y) + y\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}dy. \tag{2.4}$$

Step 2. Update $\boldsymbol{\beta}$ by maximizing

$$L(\boldsymbol{\beta}) = \sum_i p_i[\hat{\varphi}\{y_{(i)}\} + y_{(i)}\mathbf{x}_i^T\boldsymbol{\beta}] - \sum_i p_i \log \int \exp\{\hat{\varphi}(y) + y\mathbf{x}_i^T\boldsymbol{\beta}\}dy. \tag{2.5}$$

Step 3. Continue Steps 1 and 2 until convergence.

Step 1 is an optimization problem with a nonlinear objective function and a concave constraint on the result. It can be solved by a variety of methods, including the active set and EM algorithms. The application of active set and EM algorithms for log-concave densities has been explicitly formulated (Dumbgen *et al.*, 2007). The related R code can be downloaded from Dumbgen's personal webpage. However, the algorithm by Dumbgen *et al.* is only applicable for estimating a log-concave density for a univariate variable without consideration of the covariates. We tried to extend their algorithm to work for our case in which the covariate is included. However, we found that the modified algorithm is not stable, and does not converge for our study. Hence, we exclude the active set algorithm in this dissertation and develop a steepest ascent method based on the calculation of the gradient. Step 2 is a nonlinear optimization problem and can be solved using the Newton-Raphson method.

2.3.2.1 Maximization Based on Gradient

In this section, the method of steepest ascent used for maximizing the likelihood (2.4) is presented.

First, we briefly introduce the active set algorithm for estimating a log-concave density introduced by Dumbgen (2007). Since the NPMLE $\hat{\varphi}(\cdot)$ of $\varphi(\cdot)$ is a piecewise linear function with at most as many knots as observations, the active set algorithm for maximizing likelihood (2.4) can be thought of as consisting of two basic procedures. The goal of the first procedure is to find the piecewise linear function maximizing the likelihood given the knots at which the slopes of $\varphi(\cdot)$ are changing. The second procedure is used to find an additional knot. The likelihood $L(\boldsymbol{\varphi})$ (2.4) will be increased by adding this knot. Iterating these two procedures will converge to the NPMLE $\hat{\varphi}(\cdot)$.

In this dissertation, we maximize the likelihood (2.4) using the steepest ascent method based on the calculation of the gradient. Let us consider order statistics $y_{(1)} < y_{(2)} < \ldots <$

$y_{(n)}$ and nonparametric log-density $\varphi\{y_{(1)}\}, \varphi\{y_{(2)}\}, \ldots, \varphi\{y_{(n)}\}$, where $\varphi(y)$ is a piecewise linear function. To ensure concavity, the inequality $\left[[\varphi\{y_{(i+1)}\} - \varphi\{y_{(i)}\}]/\{y_{(i+1)} - y_{(i)}\}\right] \leq \left[[\varphi\{y_{(i)}\} - \varphi\{y_{(i-1)}\}]/\{y_{(i)} - y_{(i-1)}\}\right]$ for $i = 2, \ldots, n-1$ has to be satisfied. That is, the slopes of the piecewise linear parts are non-increasing. It will be handy to rewrite this as:

$$\alpha_i = -\frac{1}{\delta_{i-1}}\varphi\{y_{(i-1)}\} + (\frac{1}{\delta_{i-1}} + \frac{1}{\delta_i})\varphi\{y_{(i)}\} - \frac{1}{\delta_i}\varphi\{y_{(i+1)}\} \geq 0 \; for \; i = 2, \ldots, n-1 \qquad (2.6)$$

where $\delta_i = y_{(i+1)} - y_{(i)}$. Expression (2.6) can be written in matrix form

$$\boldsymbol{\alpha} = B\boldsymbol{\varphi}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n)$, and

$$B = \begin{pmatrix} 1 & -\frac{1}{\delta_1} & 0 & & & & \\ -\frac{1}{\delta_1} & \ddots & \ddots & & & 0 & \\ & \ddots & \ddots & \ddots & & & \\ & & -\frac{1}{\delta_i} & \frac{1}{\delta_{i-1}} + \frac{1}{\delta_i} & -\frac{1}{\delta_i} & & \\ & & & \ddots & \ddots & \ddots & \\ & 0 & & & \ddots & \ddots & -\frac{1}{\delta_{n-1}} \\ & & & & & -\frac{1}{\delta_{n-1}} & 1 \end{pmatrix}.$$

Since B^{-1} exists in general, we can have $\boldsymbol{\varphi} = B^{-1}\boldsymbol{\alpha}$, and expression (2.4) can be represented as $L_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) = L(B^{-1}\boldsymbol{\alpha})$. Thus, maximizing $L(\boldsymbol{\varphi})$ (2.4) in terms of $\boldsymbol{\varphi}$ is equivalent to maximizing $L_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$ in terms $\boldsymbol{\alpha}$ with the constraint $\alpha_i \geq 0$ for $i = 2, 3, \ldots, n-1$.

Let $\bigtriangledown L_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) = \partial L_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$. By elementary calculus and the use of the chain rule, we have

$$\bigtriangledown L_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) = B^{-1}\frac{\partial L(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}}.$$

Consequently, $\bigtriangledown L_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$ can be obtained by knowing $\partial L(\boldsymbol{\varphi})/\partial\boldsymbol{\varphi}$. See Appendix A for calculating $\partial L(\boldsymbol{\varphi})/\partial\boldsymbol{\varphi}$.

To maximize $L_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$, we iteratively update $\boldsymbol{\alpha}$ based on $\bigtriangledown L_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$. The one step update of $\boldsymbol{\alpha}$ is found by

$$\boldsymbol{\alpha}_{\text{new}} = \boldsymbol{\alpha}_{\text{old}} + t \bigtriangledown L_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})|_{\boldsymbol{\alpha}_{\text{old}}}$$

where t is chosen so that constraint (2.6) is satisfied, and the updated $\boldsymbol{\alpha}_{new}$ ensures the increase in likelihood. In the beginning, $\boldsymbol{\alpha}_{\text{new}}$ is calculated based on $t = 1$. If the $\boldsymbol{\alpha}_{\text{new}}$ does not result in the increase of likelihood (2.4), we continuously update $\boldsymbol{\alpha}_{\text{new}}$ by using a smaller value of t until the likelihood (2.4) has a larger value. See Appendix B for details on maximizing (2.4).

2.3.2.2 Estimation

In this section, the Newton-Raphson method is explained for estimating the unknown vector of parameters, $\boldsymbol{\beta}$, given a NPMLE of the distribution of the random component in the SGLM-L. At the end of step 1 in the iterative algorithm, we have an estimate of $\varphi(\cdot)$, $\hat{\varphi}(\cdot)$. The objective function for estimating $\boldsymbol{\beta}$ is:

$$L(\boldsymbol{\beta}) = \sum_i p_i[\hat{\varphi}\{y_{(i)}\} + y_{(i)}\mathbf{x}_i^T\boldsymbol{\beta}] - \sum_i p_i \log \int \exp\{\hat{\varphi}(y) + y\mathbf{x}_i^T\boldsymbol{\beta}\}dy.$$

The $\boldsymbol{\beta}$ can be estimated using the Newton-Raphson method: We first initialize $\boldsymbol{\beta}$ as $\boldsymbol{\beta}^0$, then update using the following equation

$$\boldsymbol{\beta}_{\text{new}} = \boldsymbol{\beta}_{\text{old}} - \{\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}|_{\boldsymbol{\beta}_{\text{old}}}\}^{-1}\frac{\partial L(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}|_{\boldsymbol{\beta}_{\text{old}}}.$$

We then iterate until $\boldsymbol{\beta}$ converges. The initial $\boldsymbol{\beta}^0$ are chosen to be the estimate, $\hat{\boldsymbol{\beta}}$, in the previous iteration obtained using the iterative estimation procedure introduced in the beginning of section 2.3.2.

Here, the j^{th} element of the gradient $\partial L(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ is

$$\frac{\partial L(\boldsymbol{\beta})}{\partial\beta_j} = \sum_i y_{(i)}x_{ij} - \sum_i \frac{\int yx_{ij}\exp\{\varphi(y) + y\mathbf{x}_i^T\boldsymbol{\beta}\}dy}{\int \exp\{\varphi(y) + y\mathbf{x}_i^T\boldsymbol{\beta}\}dy},$$

and the jk^{th} element of the Hessian is

$$\frac{\partial^2 L(\beta)}{\partial\beta_j\partial\beta_k} = \sum_i \{\frac{\int yx_{ij}\exp\{\varphi(y) + y\mathbf{x}_i^T\boldsymbol{\beta}\}dy \int yx_{ik}\exp\{\varphi(y) + y\mathbf{x}_i^T\boldsymbol{\beta}\}dy}{[\int \exp\{\varphi(y) + y\mathbf{x}_i^T\boldsymbol{\beta}\}dy]^2}$$

$$-\frac{\int y^2x_{ij}x_{ik}\exp\{\varphi(y) + y\mathbf{x}_i^T\boldsymbol{\beta}\}dy}{\int \exp\{\varphi(y) + y\mathbf{x}_i^T\boldsymbol{\beta}\}dy}\}.$$

A more detailed explanation is given in Appendix C.

2.4 Log-likelihood Ratio Test

Since our model is based on maximum likelihood estimation, we perform hypothesis testing based on likelihood inference. Likelihood inference proceeds by fitting a full model and a series of reduced models which are nested. This means that each reduced model in the sequence is contained within the previous one. Let $\mathbf{Y}$ and $\mathbf{X}$ be the sets of response and

covariate respectively, and $\boldsymbol{\beta}$ be the set of parameters of interest in the model. An interesting hypothesis, or reduced model, is that $H_0 : \boldsymbol{\beta} = 0$. The difference between this reduced model and the full model with no restriction on $\boldsymbol{\beta}$ can be examined by calculating the likelihood ratio test statistic, which is defined as

$$G = 2\{\log L(\hat{\boldsymbol{\beta}}|\mathbf{Y}, \mathbf{X}) - \log L(\hat{\boldsymbol{\beta}}_0|\mathbf{Y}, \mathbf{X})\} \tag{2.7}$$

where $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}$ are the maximum likelihood estimates of $\boldsymbol{\beta}$ under the reduced model and the full model, respectively. Assuming that the reduced model is correct, the sampling distribution of G is approximately a chi-squared distribution with number of degrees of freedom equal to the difference between the number of parameters specified under the reduced model and the full model (Shao, 2003).

The likelihood ratio statistic (2.7) compares the unrestricted maximum $\log L(\hat{\boldsymbol{\beta}}|\mathbf{Y}, \mathbf{X})$ of the log-likelihood with the maximum $\log L(\hat{\boldsymbol{\beta}}_0|\mathbf{Y}, \mathbf{X})$ obtained for the restricted MLE $\hat{\boldsymbol{\beta}}_0$, computed under the restriction $H_0 : \boldsymbol{\beta} = 0$. If the unrestricted maximum $\log L(\hat{\boldsymbol{\beta}}|\mathbf{Y}, \mathbf{X})$ is significantly larger than $\log L(\hat{\boldsymbol{\beta}}_0|\mathbf{Y}, \mathbf{X})$, implying that G is large, H_0 will be rejected in favor of H_1.

For the semiparametric generalized linear model with a log-concave component, the null hypothesis is $H_0 : \boldsymbol{\beta} = 0$. The corresponding likelihood ratio test statistic is given by

$$G = 2[\sum_i p_i\{y_{(i)}\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\} - \sum_i p_i \log \int \exp\{\hat{\varphi}(y) + y\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}dy + \sum_i p_i \log \int \exp\{\hat{\varphi}(y)\}dy].$$

2.5 Numerical Results

2.5.1 Simulation

A simulation study is conducted to understand the performance of our SGLM-L approach. We generate data based on three cases:

Case 1: data sets are generated based on the simple linear regression model with normally distributed errors. The results will show us the performance of the two step estimation procedure introduced in section 2.3.2.

Case 2: data sets are generated based on the simple linear regression model with normally distributed errors to assess whether the log-likelihood ratio statistic (2.7) approximately follows a chi-square distribution and the power of the test.

Case 3: the random error is simulated no longer from a normal distribution, but instead a gamma distribution. The results let us know if the SGLM-L works for non-normal data.

For case 1, the simulation is based on the regression function $y_i = 2x_i + \varepsilon_i$, where, $i \in (1, 2, \ldots, 40)$, and the x_i's are random points in the interval $[-2, 2]$. 1000 data sets are

generated for each case of $\varepsilon_i \sim N(0,1)$, $\varepsilon_i \sim N(0,1.5)$, and $\varepsilon_i \sim N(0,2)$. The average estimated values of β's in SGLM-L (2.2) are 2.29, 1.03, and 0.59 with respect to $\varepsilon_i \sim N(0,1)$, $\varepsilon_i \sim N(0,1.5)$, and $\varepsilon_i \sim N(0,2)$. The estimated nonparametric log-density curves for the random components of three randomly selected data sets are shown in Figure 2.1. The likelihood ratio test statistic is shown in Table 2.2. In Figure 2.1, the dashed lines represent the normal distributions (in log) used to generate the data sets, and the solid lines represent the estimated nonparametric log-densities conditional on $x = 0$. In Figure 2.1, the NPMLEs of the distributions of the random component are fitted fairly well in the mean area, but are not in the tail areas. This is because the most data are located in mean area and only a few data are in tail area. The simulation results show that the estimation procedure introduced in section 2.3.2 is a straightforward method for estimating the coefficients in the SGLM-L.

Case 1	$\varepsilon \sim N(0,1)$	$\varepsilon \sim N(0,1.5)$	$\varepsilon \sim N(0,2)$
$L(\hat{\beta})$	-42.97	-65.94	-78.69
$L(\beta_0)$	-126.15	-120.32	-110.31
G	166.32	108.75	63.23

Table 2.2: Summary for average likelihood tatio test statistic for simulated data sets based on the regression function $y_i = 2x_i + \varepsilon_i,\ i = 1, 2, \cdots, 40$. G is the log-likelihood ratio test statistic defined by expression (2.7).

For case 2, we conducted a study for the distribution of log-likelihood ratio test statistic based on simulation from the simple linear regression $y_i = \alpha x_i + \varepsilon_i$, where, $i \in (1, 2, \ldots, 40)$, the x_i's are random points in the interval $[-2, 2]$, and $\varepsilon_i \sim N(0,1)$. Log-likelihood ratio inference is based on the technique introduced in section 2.4. 1000 data sets are generated using $\alpha = 0$. To estimate the Type I error rate, we assume that the unknown true parameter β in SGLM-L (2.2) is equal to α, and the log-likelihood ratio test statistic is assumed to have a χ^2_1 distribution. The null hypothesis of interest is $H_0 : \beta = 0$. The estimated type I error rate, Prob(Reject $H_0|\beta = 0$), is shown in Table 2.3. The estimated type I errors is 0.049 for the simple linear regression, which is close to nominal value 0.05. The estimated type I errors is 0.069 for the SGLM-L, which is not close to 0.05 compared to the estimate of simple linear regression. We conduct a Monte Carlo power analysis based on a common type I error rate 0.05. The details for the process of this power analysis is explained in Appendix D. 1000 data sets are generated for each of the five values of α varied from 0.1 to 0.5 as shown in Table 2.3. The estimated power of the test, Prob(Reject$H_0|\beta \neq 0$), is shown in Table 2.3. The estimated power using the SGLM-L increased from 0.135 to 0.964 with increasing of the α from 0.1 to the values above 0.5. The estimated power using the simple linear regression increased from 0.141 to 0.964 with increasing of the α from 0.1 to the values above 0.5. For each value of α, the estimated power using the SGLM-L is close to the estimated power using the simple linear regression. The results show us that the log-likelihood ratio test statistic may approximately follow a chi-square distribution. However, this distribution assumption is not proved theoretically in this dissertation.

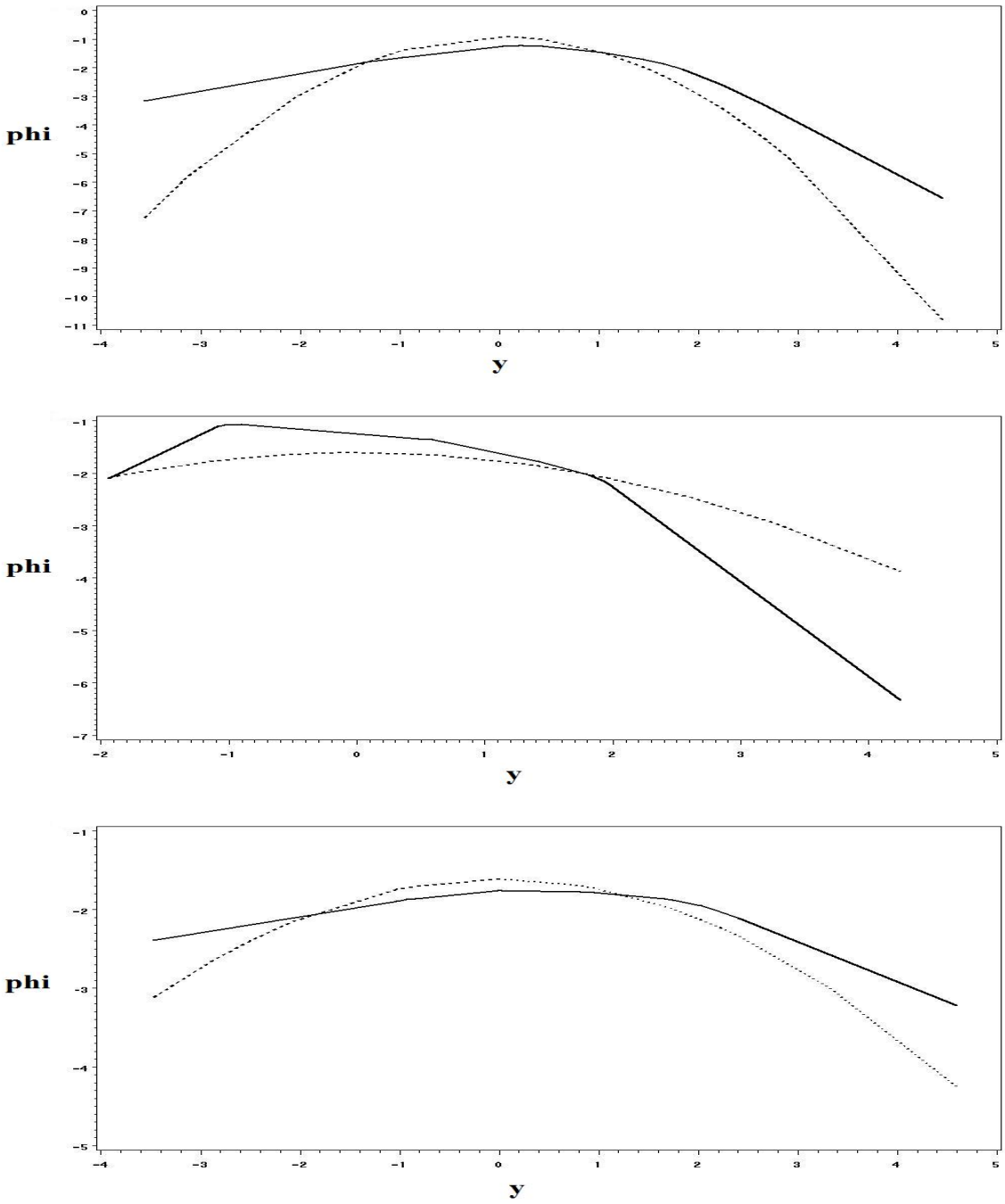

Figure 2.1: Plot of estimated nonparametric log-density of random component. The solid lines are the estimated curves. The dashed lines are the true densities used for generating the data sets. Data sets are generated using $y_i = 2x_i + \varepsilon_i, \varepsilon_i \sim N(0,1)$ top, $\varepsilon_i \sim N(0,1.5)$ middle, and $\varepsilon_i \sim N(0,2)$ bottom.

Case 2	Type I	Power*				
	Error Rate	$\alpha = 0.1$	0.2	0.3	0.4	0.5
SGLM-L	0.069	0.135	0.299	0.493	0.794	0.964
SLR	0.049	0.141	0.298	0.490	0.788	0.964

*The power values were computed using a simulated cutoff selected to achieve a 0.05 type I error rate.

Table 2.3: Estimated type I error rate and power of the log-likelihood ratio test in simulation. SGLM-L is the semiparametric generalized linear model with a log-concave component. SLR is the simple linear regression.

For case 3, 1000 data sets are generated based on the regression function $y_i = 2x_i + \varepsilon_i$, where, $i \in (1, 2, \ldots, 40)$, the x_i's are random points in the interval $[-2, 2]$, and $\varepsilon_i \sim$ Gamma$(2, 1)$. Table 2.4 shows the mean log-likelihoods are -30.86, -61.12, and -67.78 using the generalized linear model, the SGLM-L, and the simple linear regression recpectively. The mean log-likelihood of the SGLM-L is larger than the mean log-likelihood of the simple linear regression. It suggests that the SGLM-L works well for non-normal data.

Case 3	Log-linear Model	SGLM-L	Simple Linear Regression
Likelihood	-30.86	-61.12	-67.78

Table 2.4: Results of mean log-likelihood over 1000 datasets using three different models: the generalized linear model with gamma random component, the semiparametric generalized linear model with a log-concave component, and the simple linear regression.

2.5.2 Application

For an example of the application of the SGLM-L, we consider once again the North Carolina environmental data. We fit the SGLM-L to the data set including the daily concentration of $PM_{2.5}$ and ozone in New Hanover county of North Carolina in the years 2004 and 2005. In this study, $PM_{2.5}$ is the response variable, and ozone is the independent variable. From Figure 2.2, we note that a strong linear association exists between $PM_{2.5}$ and ozone. There are 98 daily observations available. The estimate of the shift of conditional maximum log-likelihood is $\hat{\beta} = 1.22$ for the SGLM-L. The log-likelihood ratio statistic is 40.44. This application gives us an example of a SGLM-L fit to real data. There may be better way to fit this data, such as a nonparametric regression.

The difference of estimates between our model and the simple linear regression exists. The estimated slope based on the simple linear regression, $PM_{2.5} = \alpha_0 + \alpha_1 \text{ozone}$, is 1.874. The estimated slope in the simple linear regression is larger than the estimate of our two

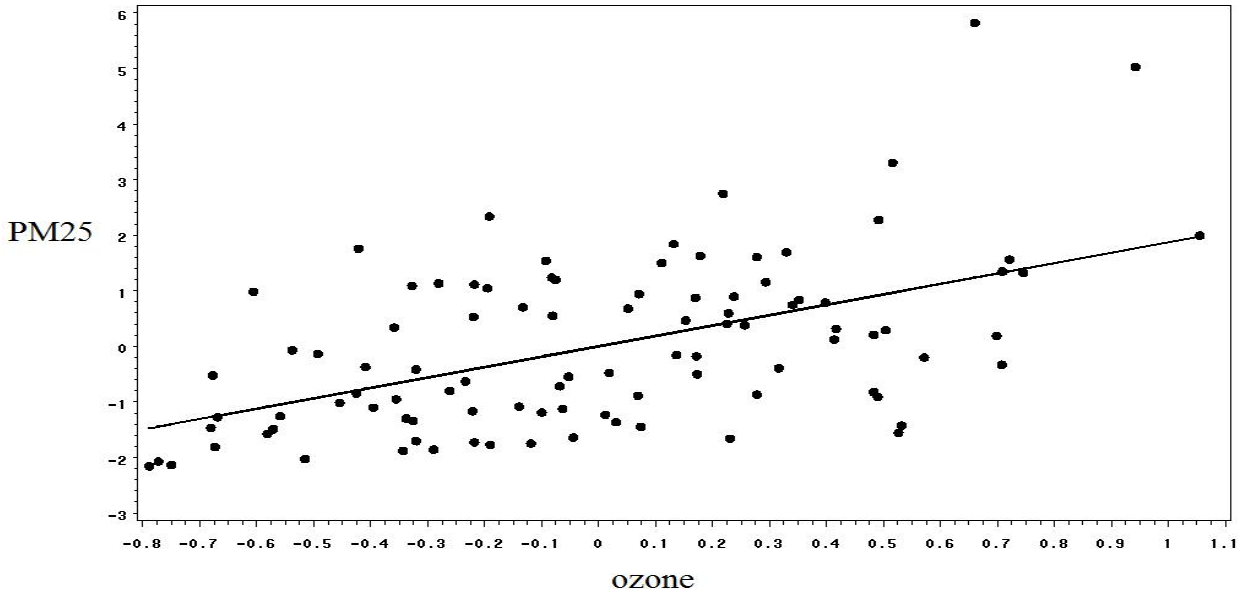

Figure 2.2: Plot of scaled New Hanover data: The solid line is fitted based on simple linear regression.

step estimation procedure in the SGLM-L by about 50%. This is due to the existence of different interpretations of the parameters between the models. Compared to the simple linear regression, the SGLM-L provides an alternative way for fitting the data. However, we do not compare the quality of fit between the SGLM-L and the simple linear regression for this data since the interpretation of the parameters is different between the models.

As we introduced in the application of NC data, the parameters in our SGLM-L have a different interpretation compared to the parameters in traditional methods, such as linear regression. Let us consider the univariate response Y with one independent variable X. Mathematically, a simple linear regression can be described by an equation of the form $y = \alpha_0 + \alpha_1 x$. For any fixed value of the variable X, Y is assumed to be a random variable with a normal distribution. And the mean value of Y, $\mu_{Y|X}$, is a straight-line function of X with slope α_1. The plot on the top of Figure 2.3 illustrates the normal distribution and straight-line assumptions in the simple linear regression. However, the parameter β in SGLM-L (2.2) is just a tilting factor of conditional likelihood of Y given X. The conditional maximum log-likelihood of Y, $\tilde{l}(Y|X)$, is a straight-line function of X with slope β. And the likelihood is estimated nonparametrically without assuming any specific distribution; rather, we just assume the log-concave property. This concept is illustrated in the plot at the bottom of Figure 2.3.

When faced with the task of building a univariate regression model, one of the first steps is the construction of a scatterplot of the data at hand. The scatterplot can be used to suggest

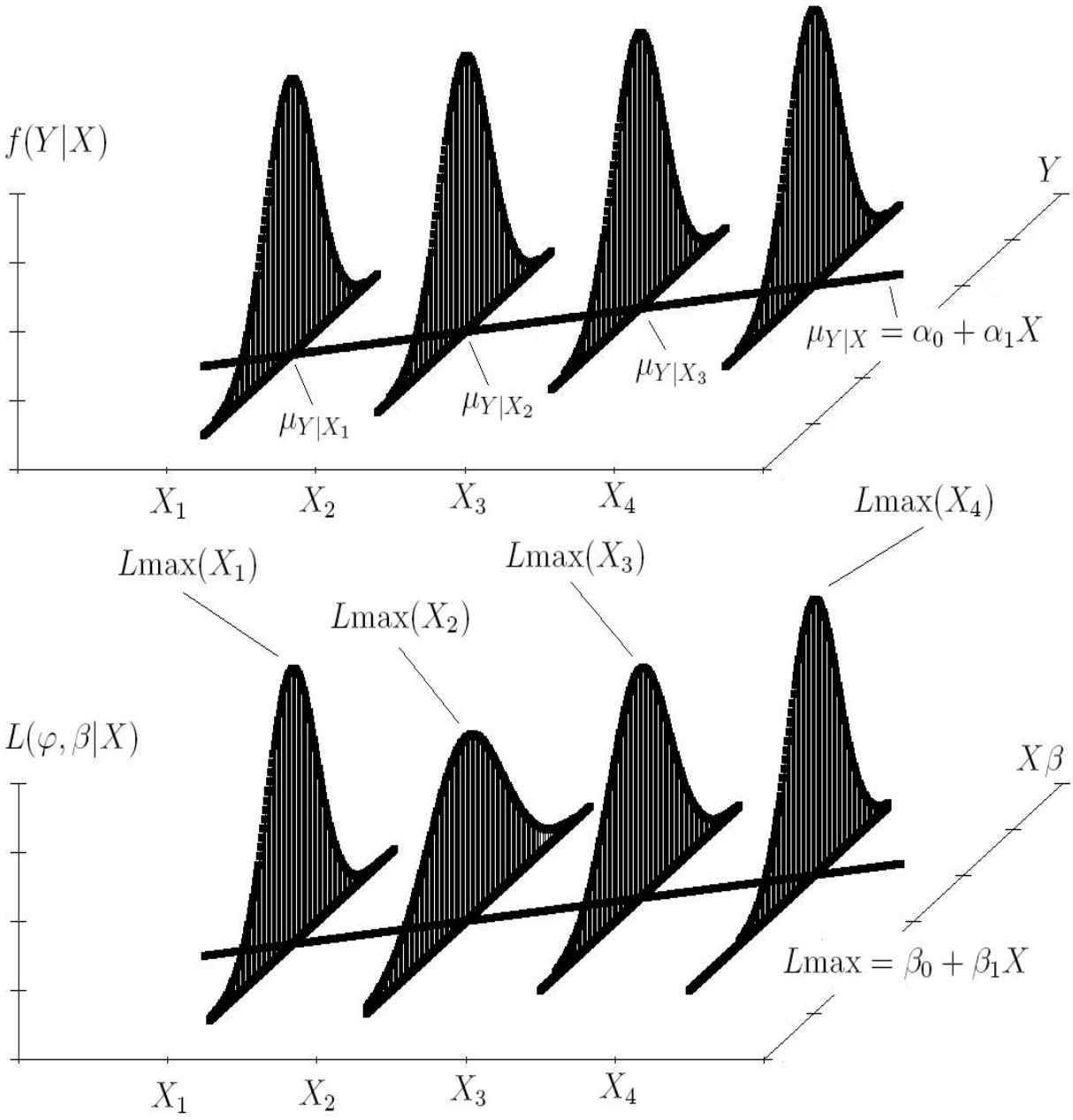

Figure 2.3: Comparison between simple linear regression and SGLM-L: the plot on the top is for the simple linear regression, the plot at the bottom is for SGLM-L.

an appropriate form for the model. However, a traditional scatterplot is not appropriate for visualizing contributions to a likelihood function for non-normal data in generalized linear models. Eno and Terrell (1999) presented a method of constructing a plot analogous to a scatterplot for logistic regression. They used grayscale graphics to visualize the contributions to a likelihood function. It would be helpful to extend a similar method for visualizing the contribution to a likelihood function in the semiparametric generalized linear model with a log-concave component in further research. This visualization method will also be useful in understanding the interpretation of the parameters in the SGLM-L.

2.6 Summary

The results based on the simulation and application of North Carolina air pollution data indicate that the semiparametric generalized linear model with a log-concave random component (SGLM-L) provides an alternative way for modeling linear associations. The iterative two step estimation procedure is a straightforward method for estimating the coefficients in the model. However, the estimates of our SGLM-L are different from the estimates of a traditional method, such as the simple linear regression. This leads to different interpretations of the estimates.

No assumption of the form of link function provides flexibility in modeling various kinds of data. The SGLM-L can be used in the case that the distribution of response belongs to log-concave families, including the normal, Laplace, and logistic distributions among many others (see section 2.1). For the data in which the response follows a gamma distribution, the simulation study shows that the SGLM-L had a higher likelihood relative to the one for a simple linear regression using normal distribution. This example suggests that the SGLM-L method may work well for other non-normal members of the log-concave family.

Chapter 3

Generalized Semiparametric Single-Index Mixed Model

3.1 Introduction

Generalized linear mixed models (GLMMs) are a natural extension of generalized linear models. The GLMMs enable the accommodation of nonnormally distributed responses, and they can model overdispersion and correlation by incorporating random effects. Unlike generalized linear models, the integration of the likelihood function in the GLMMs does not have a closed form and is numerically intractable for complicated problems involving irreducibly high-dimensional integrals. This problem leads to difficulties in making inference about the parameters. To overcome the computation limitation, various approximate methods for inference in the GLMMs have been proposed (Stirateli *et al.*, 1984; Breslow and Clayton, 1993). Bayesian methods also provide an attractive approach to inference in the GLMMs (McCulloch, 1997; Zeger and Karim, 1991; Natarajan and Kass, 2000).

An additional complicating factor is that the fixed effect component does not adequately describe the relationship between the response and the associated covariates. The assumption that the relationship is linear is sometimes violated. This has led to research on various generalized nonparametric mixed models. Lin and Zhang's paper (1999) proposed using generalized additive mixed models (GAMM), which are an additive extension of the GLMMs. They estimate the nonparametric functions using smoothing splines and jointly estimate the smoothing parameters and the variance components using marginal quasi-likelihood. A double penalized quasi-likelihood (DPQL) is used to make approximate inference. Zhang (2004) proposed using generalized linear mixed models with varying coefficients for longitudinal data. The parametric functional form in the linear predictor is routinely assumed in the GLMMs. Zhang relaxes this assumption by representing the covariate effects by a smooth but otherwise arbitrary function of time, with random effects used to model the correlation

induced by between-subject and within-subject variation in longitudinal data.

Assume we have an input vector $\mathbf{x}$ with p components, and a target $\mathbf{Y}$. Let $\boldsymbol{\beta}_m, m = 1, 2, \ldots, M$, be unite p-vectors of unknown parameters. The projection pursuit regression (PPR) model has the form $f(\mathbf{x}) = \sum_{m=1}^{M} f_m(\boldsymbol{\beta}_m^T \mathbf{x})$ (Friedman and Stuetzle, 1981). This is an additive model. The functions $f_m(\cdot)$ are unspecified and estimated along with the directions $\boldsymbol{\beta}_m$. The univariate variable $\mathbf{v}_m = \boldsymbol{\beta}_m^T \mathbf{x}$ is the projection of $\mathbf{x}$ onto the unit vector $\boldsymbol{\beta}_m$, and we seek $\boldsymbol{\beta}_m$ so that the model fits well, hence the name "projection pursuit". The $M = 1$ PPR is known as the single-index model.

The single-index model (Stoker, 1986) is a useful tool in multivariate nonparametric regression. When the number of covariates is relatively large, a problem arises is the well-known "curse of dimensionality" which implies the performance of nonparametric smoothing techniques deteriorates as the dimensionality increases (Bellman, 1961). By reducing the dimensionality from multivariate predictors to a univariate index $\mathbf{x}^T \boldsymbol{\beta}$, single-index models avoid the "curse of dimensionality".

The motivation, importance, and broad potential applications of single-index model are widely discussed in the literature. Hardle and Stoker (1989) and Ichimura (1993) have given examples of classical regression, discrete regression, and censored regression that can all be classified as single-index models. Various methods are available for fitting single-index models. Ichimura (1993) and Hardle *at al.* (1993) used kernel smoothing and discussed the empirical rule for bandwidth selection. Carroll *et al.* (1997) used local linear methods. Stoker (1986) and Hardle and Stoker (1989) used the average derivatives method.

Carroll *et al.* (1997) extended the generalized linear models to the generalized partially linear single-index models, where the mean function has the form $\eta(\mathbf{x}^T \boldsymbol{\beta}) + \boldsymbol{\alpha}^T \mathbf{w}$. This model allows for some of the predictors to be modeled linearly, with others being modeled nonlinearly. They proposed an iterative estimation procedure based on the quasi-likelihood approach, where the local linear methods were applied. Problems with the stability of the algorithm of Carroll *et al.* (1997) led Yu and Ruppert (2002) to propose penalized spline estimation for the partially linear single-index models. In an unpublished technical report by Yu (2008), the penalized spline estimation for the generalized partially linear single-index models is developed based on the quasi-likelihood method, and the asymptotic properties of the estimator are developed.

Combining the GLMMs and single-index models, we propose the generalized semiparametric single-index mixed model (GSSIMM) in this chapter. We use a single-index model to generalize the GLMM to have the linear combination of covariates enter the model via a nonparametric link function. This new class of model uses a nonparametric component, $\eta(\mathbf{x}^T \boldsymbol{\beta})$, to model covariate effects while accounting for overdispersion and correlation by adding random effects.

In this chapter, we address methodological issues related with the GSSIMM. We use penalized quasi-likelihood (PQL) to make approximate inference since the integration of full

likelihood is often intractable. An iterative two step estimation procedure is proposed to estimate the parameters in the model. We estimate the nonparametric function using P-splines and jointly estimate the smoothing parameters and the variance components using marginal quasi-likelihood. Asymptotic properties of the estimator are provided. We discuss inference based on the sandwich formula (Yu and Ruppert, 2002). The GSSIMM is applicable to longitudinal and spatial data.

This research is motivated by a study of association between daily mortality and air pollutants adjusted with climate variables. Outdoor and indoor air quality are important to human health. It is now well documented that high levels of many airborne pollutants can adversely affect many of the body's systems. Although much research has been conducted to study the association between daily mortality and daily concentration of air pollutants (Calder *et al.*, 2008; Curtis *et al.*, 2006; Kleem *et al.*, 2000), the true underlying relation is still unknown. Hence, in this chapter, we model it via GSSIMM. As we introduced in chapter one, we analyze the daily effect of $PM_{2.5}$ and ozone concentration on mortality using data from twelve counties in North Carolina in the years 2004 and 2005. For data analysis, we let the response observation y_{ij} be the mortality count due to cardiovascular disease in the ith county on the jth day. The vector of the jth daily observation for the independent variables in the ith county is $\mathbf{x}_{ij}$, including $PM_{2.5}$, ozone, average temperature, and average wind speed. The random effect for the ith county is denoted as $\mathbf{z}_i^T\mathbf{b}$, with random effect variable $\mathbf{b} \sim N(\mathbf{0}, \mathbf{\Sigma_b})$, and a vector, $\mathbf{z}_i$, associated with the random effect for the ith county. The reason to treat the county effect as random is that 12 counties are randomly selected from 100 counties in North Carolina.

For the data described above, it is straightforward to fit a loglinear mixed model as

$$\log\{\mu(\mathbf{x}, \mathbf{z})\} = \beta_0 + \mathbf{x}^T\boldsymbol{\beta} + \mathbf{z}^T\mathbf{b}, \tag{3.1}$$

where β_0 is the intercept, and $\boldsymbol{\beta}$ is the vector of unknown parameters. Model (3.1) has the advantage of both computational convenience and interpretation of the model parameters. However, the linear model (3.1) is not complex enough to capture the curvature of the relationship between daily mortality and the associated air pollutants.

In recent years, the generalized additive models (Hastie and Tibshirani, 1990) have been widely used to describe the relationship between air pollution and the health outcome of interest, adjusted for relevant seasonal and weather confounders. It is natural for us to use the following generalized additive mixed model

$$\log\{\mu(\mathbf{x}, \mathbf{z})\} = \beta_0 + f_1(\mathbf{x}_1) + \cdots + f_p(\mathbf{x}_p) + \mathbf{z}^T\mathbf{b}$$

to analyze the North Carolina air pollution data. Here, for $i \in (1, \cdots, p)$, $f_i(\cdot)$ is a unknown function, and β_0 is an intercept. However, there are two important drawbacks in using the GAMM to study the health effects of air pollution. First, the GAMM can not be used to model interaction. Since there is correlation among air pollutants and climate variables, it is important to consider the interactions among them. For instance, carbon monoxide is a

product of incomplete combustion, and is correlated with many other air pollutants. Second, the GAMM can not be used to model non-additive synergistic effects of air pollutants because of the additive assumption among predictors in the GAMM. Laboratory and epidemiologic studies have demonstrated effects of combined or sequential exposures that were greater than the effects of either exposure given singly (Mauderly and Samet, 2009).

Based on the above considerations, we suggest the GSSIMM for studying the health effects of air pollution. The GSSIMM allows curvature in the predictors by adding a nonparametric component, models correlation by incorporating random effects, and allows for non-additive synergistic effects of air pollutants.

The rest of this chapter is organized as follows. The GSSIMM is introduced in section 3.2. Section 3.3 presents the method of estimation. Variance components estimation and asymptotic properties of the estimates are explored in Section 3.4. A simulation and the study of health effects of air pollution at North Carolina are conducted to compare our approach with the GLMM and GAMM. The results are in Section 3.5 and 3.6 respectively. A summary of our conclusions and implications of this work are presented in the final section.

3.2 Model

We consider a semiparametric version of the generalized linear mixed model. This model predicts response Y using covariates $(\mathbf{X}, \mathbf{Z})$, where $\mathbf{X}$ represents the $n \times p$ vector-valued covariates associated with fixed effects, and $\mathbf{Z}$ represents the $n \times m$ vector-valued covariates associated with random effects. The conditional density of Y given $(\mathbf{X}, \mathbf{Z}) = (\mathbf{x}, \mathbf{z})$ has a distribution in the canonical exponential family

$$f_{Y|\mathbf{X},\mathbf{Z}}(y|\mathbf{x}, \mathbf{z}) = \exp[y\theta(\mathbf{x}, \mathbf{z}) - \psi\{\theta(\mathbf{x}, \mathbf{z})\} + C(y)]$$

for known functions $\psi(\cdot)$ and $C(\cdot)$. In the parametric generalized linear mixed models, the unknown regression function $\mu(\mathbf{x}, \mathbf{z}) = E(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = \psi'\{\theta(\mathbf{x}, \mathbf{z})\}$ is modeled linearly via a link function g by

$$g\{\mu(\mathbf{x}, \mathbf{z})\} = \mathbf{x}^T\boldsymbol{\beta} + \mathbf{z}^T\boldsymbol{b} \tag{3.2}$$

where $\boldsymbol{b}$ is the random effect.

A natural extension of (3.2) is to allow the fixed part of the mean function to be modeled nonlinearly. This leads us to consider the generalized semiparametric single-index mixed model

$$g\{\mu(\mathbf{x}, \mathbf{z})\} = \eta(\mathbf{x}^T\boldsymbol{\beta}) + \mathbf{z}^T\mathbf{b}, \quad \text{with} \quad \|\boldsymbol{\beta}\| = 1 \tag{3.3}$$

where the restriction $\|\boldsymbol{\beta}\| = 1$ is required for identifiability (see section 1.3), $\boldsymbol{b}$ is assumed to be distributed as $N(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{b})$, $\boldsymbol{\Sigma}_\mathbf{b} = \boldsymbol{\Sigma}_\mathbf{b}(\boldsymbol{\theta})$ depends on an unknown vector $\boldsymbol{\theta}$ of variance components, and $\eta(\cdot)$ is an unspecified function.

3.3 Method

3.3.1 Penalized Quasi-Likelihood

The integrated log-likelihood of $\{\eta(\cdot), \boldsymbol{\beta}, \boldsymbol{\theta}\}$ for the GSSIMM is (compare to Breslow and Clayton (1993), equation (2))

$$\exp[l_M\{\mathbf{y}; \eta(\cdot), \boldsymbol{\beta}, \boldsymbol{\theta}\}] \propto |\boldsymbol{\Sigma}_\mathbf{b}|^{-1/2} \int \exp\left[\sum_{i=1}^{n} l_i\{y; \eta(\cdot), \boldsymbol{\beta}, \boldsymbol{\theta}\} - \frac{1}{2}\mathbf{b}^T\boldsymbol{\Sigma}_\mathbf{b}^{-1}\mathbf{b}\right] d\mathbf{b}, \tag{3.4}$$

where $\mathbf{y} = (y_1, \cdots, y_n)^T$ and $l_i\{y; \eta(\cdot), \boldsymbol{\beta}, \boldsymbol{\theta}\} = y_i\{\eta(\mathbf{x}_i^T\boldsymbol{\beta}) + \mathbf{z}_i^T\boldsymbol{b}\} - \psi\{\eta(\mathbf{x}_i^T\boldsymbol{\beta}) + \mathbf{z}_i^T\boldsymbol{b}\}$. Here, we denote design matrices with rows $\mathbf{x}_i^T$ and $\mathbf{z}_i^T$ by $\mathbf{X}$ and $\mathbf{Z}$. The unknown univariate function $\eta(\cdot)$ can be estimated by a P-spline (Ruppert and Carroll, 1997; Ruppert, 2002). Assume that

$$\eta(t) = \tau_0 + \tau_1 t + \cdot + \tau_p t^p + \sum_{k=1}^{K} \tau_{p+k}(t - h_k)_+^p, \tag{3.5}$$

where $\{h_k\}_{k=1}^K$ are spline knots. In the next section we discuss the choice of the number of knots K and the knot locations. The basis $\mathbf{B}(t) = (1, t, \cdots, t^p, (t-h_1)_+^p, \cdots, (t-h_K)_+^p)^T$ is known as the truncated power basis of degree p. Define the spline coefficient vector $\boldsymbol{\tau} = (\tau_0, \tau_1, \cdots, \tau_{p+K})^T$. Then our spline model is $\eta(t) = \boldsymbol{\tau}^T\mathbf{B}(t)$. If the degree $p = 1$, $\eta(t)$ are linear splines and a linear combination of piecewise linear functions. In this dissertation, we choose quadratic spline bases for $p = 2$. The quadratic spline function has a continuous first derivative, and does not have a sharp corner like the linear spline function does. Higher value of p lead to smoother spline functions with $p - 1$ continuous derivatives.

The P-spline estimators of the $\eta(\cdot)$ maximizes the penalized log-likelihood

$$l_M\{\mathbf{y}; \eta(\cdot), \boldsymbol{\beta}, \boldsymbol{\theta}\} - \frac{n}{2}\lambda\boldsymbol{\tau}^T\mathbf{D}_\tau\boldsymbol{\tau}, \tag{3.6}$$

where λ is the smoothing parameter and $\mathbf{D}_{\boldsymbol{\tau}}$ is the penalty matrix, which is a semi-positive definite matrix. Since the numerical integration required for maximizing expression (3.6) is often intractable, we approximate it by applying the Laplace method to $l_M\{\mathbf{y}; \eta(\cdot), \boldsymbol{\beta}, \boldsymbol{\theta}\}$ in expression (3.4) (Tierney and Kadane, 1986; Lin, 1999). Some calculations show that the approximate P-spline estimators $\hat{\eta}(\cdot)$, and $\hat{\boldsymbol{\beta}}$ can be obtained by maximizing the following penalized quasi-likelihood (PQL) with respect to $\eta(\cdot), \boldsymbol{\beta}$ and $\mathbf{b}$:

$$\sum_{i=1}^{n}[y_i\{\eta(\mathbf{x}^T\boldsymbol{\beta}_i) + \mathbf{z}_i^T\boldsymbol{b}\} - \psi\{\eta(\mathbf{x}_i^T\boldsymbol{\beta}) + \mathbf{z}_i^T\boldsymbol{b}\}] - \frac{1}{2}\mathbf{b}^T\boldsymbol{\Sigma}_\mathbf{b}^{-1}\mathbf{b} - \frac{n}{2}\lambda\boldsymbol{\tau}^T\mathbf{D}_{\boldsymbol{\tau}}\boldsymbol{\tau}. \tag{3.7}$$

See Appendix E for a detailed derivation of expression (3.7).

3.3.2 Estimation Procedure

The procedure for estimating $\eta(\cdot)$, $\boldsymbol{\beta}$ and $\boldsymbol{b}$ in model (3.3) is as follows:

Step 0 (Initialization step). Fit a generalized linear mixed-effect model to obtain initial values $\hat{\boldsymbol{\beta}}_1$ and set $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_1/\|\hat{\boldsymbol{\beta}}_1\|$.

Step 1. Find $\hat{\eta}$ and $\hat{\boldsymbol{b}}$ by maximizing the PQL

$$\sum_{i=1}^{n}[y_i\{\eta(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}) + \mathbf{z}_i^T\boldsymbol{b}\} - \psi\{\eta(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}) + \mathbf{z}_i^T\boldsymbol{b}\}] - \frac{1}{2}\boldsymbol{b}^T\boldsymbol{\Sigma}_{\boldsymbol{b}}^{-1}\boldsymbol{b} - \frac{n}{2}\lambda\boldsymbol{\tau}^T\mathbf{D}_{\boldsymbol{\tau}}\boldsymbol{\tau}. \tag{3.8}$$

In this dissertation, we choose $\mathbf{D}_{\boldsymbol{\tau}} = \text{diag}(\mathbf{0}_{p+1}, \mathbf{1}_K)$ (Ruppert *et al.*, 2003).

Step 2. Update $\boldsymbol{\beta}$ by maximizing

$$\sum_{i=1}^{n}[y_i\{\hat{\eta}(\mathbf{x}_i^T\boldsymbol{\beta}) + \mathbf{z}_i^T\hat{\boldsymbol{b}}\} - \psi\{\hat{\eta}(\mathbf{x}_i^T\boldsymbol{\beta}) + \mathbf{z}_i^T\hat{\boldsymbol{b}}\}] + \gamma(\|\boldsymbol{\beta}\|^2 - 1), \tag{3.9}$$

where γ is the Lagrange multiplier.

Step 3. Continue Steps 1 and 2 until convergence.

Calculations in step 1 can be easily implemented by fitting the generalized additive mixed model. The SAS procedure GLIMMIX can be used to estimate the parameters in the GAMM. The objective function (3.9) is the log-likelihood plus a constraint. See Appendix F for details on maximizing (3.9). For more details concerning constrained maximum likelihood estimation, see Aitchison and Silvey (1960).

3.3.3 Choosing the Knots

It is recommended (Yu and Ruppert, 2002) that the knots be placed at equally-spaced sample quantiles of the predictor variable, which in this context is the index $\mathbf{x}^T\boldsymbol{\beta}$, when $\eta(\cdot)$ is modeled by a spline. For example, if there are 9 knots, then they would be at the 10th percentile, 20th percentile, etc. of the index values. Ruppert (2002) has a detailed study on the choice of total number of knots K. For smooth and either monotonic or unimodal regression functions, 10 to 20 knots seems to be quite adequate. If the regression function has a discontinuity, then it is important to have a knot near it. More than 20 knots would be needed if $\eta(\cdot)$ has many local minima and maxima, but that is unlikely in applications of the single-index model.

The idea is to choose enough knots to resolve the essential structure in the underlying regression function. But for some complicated penalized spline models there are computational advantages to keeping the number of knots relatively low. For small sample data, a

reasonable default is to choose the knots to ensure that there are a fixed number of unique observations, say 4-5, between each knot (Ruppert *et al.*, 2003). For large data sets this can lead to an excessive number of knots, so a maximum number of allowable knots (say, 20-40 toal) is recommended. In this dissertation, we choose 10 knots.

3.3.4 Selection of λ

Selecting a suitable value of smoothing parameter λ is crucial for curve fitting. There are a number of ways to choose λ, including minimizing the cross-validation (CV) score, the generalized cross validation (GCV) score, Mallow's C_p, and Akaike's information criterion (AIC). The GCV that approximates the CV is a computationally expedient criterion, and is popular in spline literature (Ruppert *et al.*, 2002).

The generalized nonparametric regression using splines can be represented as the GLMM (Ruppert *et al.*, 2002). In the GLMM representation, the smoothing parameter λ in the generalized nonparametric regression is the inverse of the variance component. Thus, the maximizer of (3.8) can be easily obtained by fitting the working GLMM. The SAS procedure GLIMMX can be used to estimate $(\eta(\cdot), \mathbf{b})$. The SAS procedure GLIMMIX already incorporates the automatic selection of λ.

3.4 Inference

3.4.1 Variance Component Estimation

By letting $\mathbf{t} = \mathbf{X}\boldsymbol{\beta}$, model (3.3) has the generalized linear mixed model representation as follows:

$$g(\boldsymbol{\mu}) = \mathbf{T}_0\boldsymbol{\alpha} + \mathbf{B}\mathbf{a} + \mathbf{Z}\mathbf{b} \tag{3.10}$$

where $\mathbf{T}_0$ and $\mathbf{B}$ were defined in expression (3.5), $\mathbf{T}_0 = (\mathbf{1}, \mathbf{t}, \cdots, \mathbf{t}^P)$, $\mathbf{B} = ((\mathbf{t}-h_1)_+^P, \cdots, (\mathbf{t}-h_K)_+^P)$, and $\mathbf{Z}$ represents the covariates associated with random effects, $\mathbf{b}$. The vector $\boldsymbol{\alpha} = (\tau_0, \cdots, \tau_p)^T$ is a $(p+1) \times 1$ vector of regression coefficients, and $\mathbf{a}$ and $\mathbf{b}$ are independent random effects with distributions $\mathbf{a} \sim N(\mathbf{0}, \boldsymbol{\Lambda})$ and $\mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{b})$. $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}(\boldsymbol{\delta})$ depends on unknown parameter $\boldsymbol{\delta}$, and $\boldsymbol{\Sigma}_\mathbf{b} = \boldsymbol{\Sigma}_\mathbf{b}(\boldsymbol{\theta})$ depends on an unknown vector $\boldsymbol{\theta}$ of variance components. In other words, $\eta(\mathbf{x}^T\boldsymbol{\beta})$ in model (3.3) is represented as $\mathbf{T}_0\boldsymbol{\alpha} + \mathbf{B}\mathbf{a}$ in model (3.10). Specifically, $\mathbf{a}$ are the penalized coefficients on the truncated power basis in a penalized spline model. However, they are treated as random effects within the mixed model representation (Ruppert *et al.*, 2003). In this dissertation, the random effect depends on a single component of dispersion. That is, $\boldsymbol{\Lambda} = \delta\mathbf{I}$ and $\boldsymbol{\Sigma}_\mathbf{b} = \theta\mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

Given $\boldsymbol{\beta}$, PQL (3.7) becomes (compare to Lin and Zhang, 1999, equation (9))

$$\sum_{i=1}^{n}\{y_i(\mathbf{T}_{0_i}^T\boldsymbol{\alpha}+\mathbf{B}_i^T\mathbf{a}+\mathbf{z}_i^T\boldsymbol{b})-\psi(\mathbf{T}_{0_i}^T\boldsymbol{\alpha}+\mathbf{B}_i^T\mathbf{a}+\mathbf{z}_i^T\boldsymbol{b})\}-\frac{1}{2}\mathbf{b}^T\boldsymbol{\Sigma}_{\mathbf{b}}^{-1}\mathbf{b}-\frac{1}{2}\mathbf{a}^T\boldsymbol{\Lambda}^{-1}\mathbf{a}, \quad (3.11)$$

where $\mathbf{T}_{0_i}^T$ and $\mathbf{B}_i^T$ are the ith rows of $\mathbf{T}_0$ and $\mathbf{B}$. We define the working vector $\mathbf{Y}$ to have components $Y_i=\zeta_i+(y_i-\mu_i)g'(\mu_i)$, where $\zeta_i=\mathbf{T}_{0_i}^T\boldsymbol{\alpha}+\mathbf{B}_i^T\mathbf{a}+\mathbf{z}_i^T\boldsymbol{b}$, and $g(\mu_i)=\zeta_i$. The solution to (3.11) via Fisher scoring may be expressed as the iterative solution to the system

$$\begin{pmatrix}\mathbf{T}_0^T\mathbf{W}\mathbf{T}_0 & \mathbf{T}_0^T\mathbf{W}\mathbf{B} & \mathbf{T}_0^T\mathbf{W}\mathbf{Z}\\ \mathbf{B}^T\mathbf{W}\mathbf{T}_0 & \mathbf{B}^T\mathbf{W}\mathbf{B}+\boldsymbol{\Lambda}^{-1} & \mathbf{B}^T\mathbf{W}\mathbf{Z}\\ \mathbf{Z}^T\mathbf{W}\mathbf{T}_0 & \mathbf{Z}^T\mathbf{W}\mathbf{B} & \mathbf{Z}^T\mathbf{W}\mathbf{Z}+\boldsymbol{\Sigma}_{\mathbf{b}^{-1}}\end{pmatrix}\begin{pmatrix}\boldsymbol{\alpha}\\ \mathbf{a}\\ \mathbf{b}\end{pmatrix}=\begin{pmatrix}\mathbf{T}_0^T\mathbf{W}\mathbf{Y}\\ \mathbf{B}^T\mathbf{W}\mathbf{Y}\\ \mathbf{Z}^T\mathbf{W}\mathbf{Y}\end{pmatrix}, \quad (3.12)$$

where $\mathbf{W}=\mathrm{diag}\{\psi''(\zeta_i)\}$. An examination of equation (3.12) shows that it corresponds to the best linear unbiased estimation (BLUE) of $\mathbf{a}$ and $\mathbf{b}$ in the associated normal theory model $\mathbf{Y}=\mathbf{T}_0\boldsymbol{\alpha}+\mathbf{B}\mathbf{a}+\mathbf{Z}\mathbf{b}+\boldsymbol{\varepsilon}$, where $\mathbf{a}$ and $\mathbf{b}$ are independent random effects with $\mathbf{a}\sim N(\mathbf{0},\boldsymbol{\Lambda})$ and $\mathbf{b}\sim N(\mathbf{0},\boldsymbol{\Sigma}_{\mathbf{b}})$ and $\boldsymbol{\varepsilon}\sim \mathrm{N}(\mathbf{0},\mathbf{W}^{-1})$.

In the rest of this section, we explain the variance estimation in terms of the reparameterized version of the general model in equation (3.10). Under the GLMM representation (3.10), we may treat $\boldsymbol{\delta}$ as extra variance components in addition to $\boldsymbol{\theta}$. Then, we construct the marginal quasi-likelihood of $(\boldsymbol{\delta},\boldsymbol{\theta})$ by assuming a flat prior for $\boldsymbol{\alpha}$ (Harville, 1974) and integrating $\mathbf{a}$, $\mathbf{b}$, and $\boldsymbol{\alpha}$ out as follows:

$$\exp\{l_M(\mathbf{y};\boldsymbol{\delta},\boldsymbol{\theta})\}\propto|\boldsymbol{\Sigma}_{\mathbf{b}}|^{-\frac{1}{2}}|\boldsymbol{\Lambda}|^{-\frac{1}{2}}\int\exp\left\{\sum_{i=1}^{n}l_i(y;\boldsymbol{\alpha},\mathbf{a},\mathbf{b})-\frac{1}{2}\mathbf{b}^T\boldsymbol{\Sigma}_{\mathbf{b}}^{-1}\mathbf{b}-\frac{1}{2}\mathbf{a}^T\boldsymbol{\Lambda}^{-1}\mathbf{a}\right\}d\mathbf{b}d\mathbf{a}d\boldsymbol{\alpha}, \quad (3.13)$$

where $l_i(y;\boldsymbol{\alpha},\mathbf{a},\mathbf{b})=y_i(\mathbf{T}_{0_i}^T\boldsymbol{\alpha}+\mathbf{B}_i^T\mathbf{a}+\mathbf{z}_i^T\boldsymbol{b})-\psi(\mathbf{T}_{0_i}^T\boldsymbol{\alpha}+\mathbf{B}_i^T\mathbf{a}+\mathbf{z}_i^T\boldsymbol{b})$.

Since the evaluation of the marginal likelihood $l_M(\mathbf{y};\boldsymbol{\delta},\boldsymbol{\theta})$ in expression (3.13) often involves high dimensional integration, we approximate it by using the Laplace method. The statistic $\sum_{i=1}^{n}l_i\{y;\boldsymbol{\alpha},\mathbf{a},\mathbf{b}\}$ is proportional to the scaled deviance (McCullagh and Nelder, 1989, section 2.1.2). Thus, taking a quadratic expansion of the exponent of the integrand of expression (3.13) about its mode before integration and approximating the statistic $l_i(y;\mu_i)$ by the Pearson χ^2- statistic (Breslow and Clayton, 1993), derivations similar to those in Appendix E give the approximate marginal log-quasi-likelihood as

$$l_M(\mathbf{y};\boldsymbol{\delta},\boldsymbol{\theta})\approx-\frac{1}{2}\log|\mathbf{V}|-\frac{1}{2}\log|\mathbf{T}_0^T\mathbf{V}^{-1}\mathbf{T}_0|-\frac{1}{2}(\mathbf{Y}-\mathbf{T}_0\hat{\boldsymbol{\alpha}})^T\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{T}_0\hat{\boldsymbol{\alpha}}), \quad (3.14)$$

where $\mathbf{V}=\mathbf{B}\boldsymbol{\Lambda}\mathbf{B}^T+\mathbf{Z}\boldsymbol{\Sigma}_{\mathbf{b}}\mathbf{Z}^T+\mathbf{W}^{-1}$. We define $\mathbf{P}=\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{T}_0(\mathbf{T}_0^T\mathbf{V}^{-1}\mathbf{T}_0)^{-1}\mathbf{T}_0^T\mathbf{V}^{-1}$ and differentiate expression (3.14) with respect to the component of $\boldsymbol{\vartheta}=(\boldsymbol{\delta},\boldsymbol{\theta})$ to obtain estimating equations for the variance parameters (Harville, 1977):

$$\frac{1}{2}\mathrm{tr}(\mathbf{P}\frac{\partial\mathbf{V}}{\partial\boldsymbol{\vartheta}_j})-\frac{1}{2}(\mathbf{Y}-\mathbf{T}_0\hat{\boldsymbol{\alpha}})^T\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\boldsymbol{\vartheta}_j}\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{T}_0\hat{\boldsymbol{\alpha}})=0.$$

The corresponding Fisher information matrix of the approximate marginal quasi-likelihood estimators $\hat{\boldsymbol{\vartheta}} = (\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\theta}})$ can be approximated by

$$\mathcal{I}(\boldsymbol{\vartheta}) = \begin{pmatrix} \mathcal{I}_{\boldsymbol{\delta}\boldsymbol{\delta}} & \mathcal{I}_{\boldsymbol{\delta}\boldsymbol{\theta}} \\ \mathcal{I}_{\boldsymbol{\delta}\boldsymbol{\theta}}^T & \mathcal{I}_{\boldsymbol{\theta}\boldsymbol{\theta}} \end{pmatrix}, \tag{3.15}$$

where the (j,k)th element of $\mathcal{I}(\boldsymbol{\vartheta})$ is $\mathcal{I}_{\boldsymbol{\vartheta}_j\boldsymbol{\vartheta}_k} = -\frac{1}{2}\text{tr}(\mathbf{P}\frac{\partial \mathbf{V}}{\partial \boldsymbol{\vartheta}_j}\mathbf{P}\frac{\partial \mathbf{V}}{\partial \boldsymbol{\vartheta}_k})$. Note that we mainly use equation (3.15) to construct an approximate covariance matrix of $\hat{\boldsymbol{\theta}}$ and are not interested in using it to make inference on $\hat{\boldsymbol{\delta}}$. As we introduced in the beginning of this section, $\mathbf{a}$ in model (3.10) are the penalized coefficients on the truncated power basis in a penalized spline model, and are not of interest for interpretation. Consequently, we are not interested in the estimation of the corresponding variance component $\boldsymbol{\delta}$. Although the estimate of $\boldsymbol{\theta}$ has been demonstrated to be reasonably good for discrete data problems with moderate to large cell frequencies, it is less satisfactory for sparse data. Breslow and Lin (1995) and Lin and Breslow (1996) have proposed a good bias correction method to improve the estimate of $\boldsymbol{\theta}$. Although the bias correction method was not employed in this research it will be included in future research.

3.4.2 Asymptotic Properties

In this section, we present results on the consistency and asymptotic normality for PQL estimators of the generalized nonparametric single-index mixed models. We consider asymptotics when the smoothing parameter $\lambda_n \to 0$ as sample size grows as well as when λ is fixed. When addressing the asymptotic properties of PQL estimators, we condition on the number of knots being fixed. Define $\boldsymbol{\omega} = (\boldsymbol{\beta}^T, \boldsymbol{b}^T, \boldsymbol{\tau}^T)^T$, and $\mathbf{v}_i = (\mathbf{x}_i, \mathbf{z}_i, y_i)$. We call $\boldsymbol{\omega}_0 = (\boldsymbol{\beta}_0^T, \boldsymbol{b}_0^T, \boldsymbol{\tau}_0^T)^T$ the "true" parameter for the remainder of the section.

3.4.2.1 Asymptotics with $\lambda_n \to 0$

We denote λ by λ_n to indicate dependence on the sample size.

Theorem 1 Consider the maximizer of the quasi-likelihood (3.7). Under regularity conditions 1-3 in Appendix G, if the smoothing parameter $\lambda_n = o(1)$, then a sequence of penalized likelihood estimators $\hat{\boldsymbol{\omega}} = (\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{b}}^T, \hat{\boldsymbol{\tau}}^T)$ maximizing expression (3.7) exists and is a consistent estimator of $\boldsymbol{\omega}_0$ such that $\|\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}_0\| = O_p(n^{-1/2} + \lambda_n)$.

Proof See Appendix G.

Theorem 2: Consider the maximizer of the quasi-likelihood (3.7). Under regularity conditions 1-3 in Appendix G, if the smoothing parameter $\lambda_n = o(n^{-1/2})$, then a sequence of penalized likelihood estimators $\hat{\boldsymbol{\omega}}$ exists, is consistent, and is asymptotically normally

distributed. That is,

$$\sqrt{n}(\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}_0) \xrightarrow{D} N[0, \{I(\boldsymbol{\omega}_0)\}^{-1}]$$

where $I(\boldsymbol{\omega}_0)$ is the Fisher information matrix defined in the Appendix G.

Proof See Appendix H.

3.4.2.2 Asymptotics With λ Fixed and Inference Using the Sandwich Formula

The asymptotic variance in Theorem 2 does not involve λ since λ goes to 0 sufficiently fast as n tends to infinity. For finite sample inference, one would expect this asymptotic variance to over-estimate the variance of $\hat{\boldsymbol{\omega}}$ because some terms are assumed to vanish with infinite samples when approximating the asymptotic variance in Theorem 2. Therefore, for purpose of inference, we give the asymptotic distribution of $\hat{\boldsymbol{\omega}}$ when λ is fixed.

The PQL maximizer of (3.7) is a solution to a set of estimating equations

$$\sum_{i=1}^{n}\{y_i - \mu(\mathbf{v}_i;\boldsymbol{\omega})\}\frac{\partial}{\partial\boldsymbol{\omega}}m(\mathbf{v}_i;\boldsymbol{\omega}) - \boldsymbol{\Sigma}\boldsymbol{\omega} - n\lambda\mathbf{D}\boldsymbol{\omega} = 0 \tag{3.16}$$

where $g\{\mu(\mathbf{v}_i;\boldsymbol{\omega})\} = m(\mathbf{v}_i;\boldsymbol{\omega}) = \boldsymbol{\tau}^T\mathbf{B}(\mathbf{x}^T\boldsymbol{\beta}_i) + \mathbf{z}_i^T\mathbf{b}$, $\boldsymbol{\Sigma}$ and $\mathbf{D}$ are defined in Appendix G. The gradient matrix of the mean function is

$$\frac{\partial}{\partial\boldsymbol{\omega}}m(\mathbf{v}_i;\boldsymbol{\omega}) = \begin{pmatrix} \boldsymbol{\tau}^T\mathbf{B}'(\mathbf{x}^T\boldsymbol{\beta}_i)\mathbf{x}_i \\ \mathbf{z} \\ \mathbf{B}(\mathbf{x}^T\boldsymbol{\beta}_i) \end{pmatrix}.$$

Equation (3.16) is an estimating equation and can be rewritten using the notation

$$\varphi_i(\boldsymbol{\omega};\lambda) = \{y_i - \mu(\mathbf{v}_i;\boldsymbol{\omega})\}\frac{\partial}{\partial\boldsymbol{\omega}}m(\mathbf{v}_i;\boldsymbol{\omega}) - \frac{1}{n}\boldsymbol{\Sigma}\boldsymbol{\omega} - \lambda\mathbf{D}\boldsymbol{\omega}$$

as $\sum_{i=1}^{n}\varphi_i(\boldsymbol{\omega};\lambda) = 0$. Let $\hat{\boldsymbol{\omega}}(\lambda)$ solve $\sum_{i=1}^{n}\mathsf{E}\varphi_i\{\boldsymbol{\omega}(\lambda);\lambda\} = 0$. If we think of $\hat{\boldsymbol{\omega}}(\lambda)$ as an estimator of $\boldsymbol{\omega}(\lambda)$, the covariance matrix of $\hat{\boldsymbol{\omega}}(\lambda)$ can be estimated by the sandwich formula (Huber, 1967; Carroll *et al.*, 1995). Define

$$H(\boldsymbol{\omega}) = \sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\omega}^T}\varphi_i(\boldsymbol{\omega}) \text{ and } G(\boldsymbol{\omega}) = \sum_{i=1}^{n}\varphi_i(\boldsymbol{\omega})\varphi_i(\boldsymbol{\omega})^T$$

Assume that $n^{-1}H(\boldsymbol{\omega})$ and $n^{-1}G(\boldsymbol{\omega})$ have limits as $n \to \infty$. One can show that $\hat{\boldsymbol{\omega}}(\lambda)$ is consistent for $\boldsymbol{\omega}(\lambda)$ and that

$$n^{1/2}\{\hat{\boldsymbol{\omega}}(\lambda) - \boldsymbol{\omega}(\lambda)\} \xrightarrow{D} \text{normal}[0, H\{\boldsymbol{\omega}(\lambda)\}^{-1}G\{\boldsymbol{\omega}(\lambda)\}[H\{\boldsymbol{\omega}(\lambda)\}^{-1}]^T].$$

These asymptotics justify the sandwich estimator V_{sw} of the covariance matrix of $\hat{\boldsymbol{\omega}}(\lambda)$, which is

$$V_{sw} = H\{\hat{\boldsymbol{\omega}}(\lambda)\}^{-1} G\{\hat{\boldsymbol{\omega}}(\lambda)\}[H\{\boldsymbol{\omega}(\lambda)\}^{-1}]^T. \tag{3.17}$$

The sandwich estimate can be used for joint confidence regions and for hypothesis testing using a Wald test. For example, if a test of the null hypothesis of $H_0 : \mathbf{R}\boldsymbol{\omega}_0 - \mathbf{q}_0 = 0$ is desired, where $\mathbf{R}$ is a $d_1 \times \dim(\boldsymbol{\omega})$ matrix of full rank $d_1 \leq \dim(\boldsymbol{\omega})$, then the test can be based on the Wald statistic,

$$W = (\mathbf{R}\hat{\boldsymbol{\omega}} - \mathbf{q}_0)^T (\mathbf{R}V_{sw}\mathbf{R}^T)^{-1} (\mathbf{R}\hat{\boldsymbol{\omega}} - \mathbf{q}_0)$$

which has a chi-square limiting distribution with d_1 degrees of freedom.

In the GSSIMM, our focus is on the d-dimensional single-index parameter $\boldsymbol{\beta}$. Then the upper-left $d \times d$ block matrix of V_{sw} can be used in calculating the Wald statistic.

3.5 Simulation

We conduct simulation study to understand the performance of our approach by considering the following three cases.

3.5.1 Case 1: Count Data Generated from a Nonlinear Mixed Model

A simulation study was conducted similar to the one in Carroll *et al.* (1997), but for a count response instead of a binary response. In this simulation example, the systematic components are generated from a "sine-bump" model

$$\log(\mu_{ij}) = 1 + \sin \frac{\pi(\mathbf{x}_{ij}^T \boldsymbol{\beta} - A)}{B - A} + \mathbf{z}_i^T \boldsymbol{b} \tag{3.18}$$

where the $\mathbf{x}_{ij}$ are trivariate with independent uniform [0,1] components, $\boldsymbol{b} \sim N(\mathbf{0}, 0.5\mathbf{I})$, $i \in (1, 2, \cdots, 12)$, $j \in (1, 2, \cdots, n_i)$ for each i, and $n_i = 200$ for all i's. Define $\mathbf{z}_i$ is a vector, where the ith element is 1, otherwise zero. The vector of parameters is $\boldsymbol{\beta} = (1, 1, 1)/\sqrt{3} = (0.577, 0.577, 0.577)$. We set $A = \sqrt{3}/2 - 1.645/\sqrt{12}$ and $B = \sqrt{3}/2 + 1.645/\sqrt{12}$ to ensure that the function (3.18) was relatively thick in the tails. 1000 data sets were generated based on model (3.18). In each data set, we generate the response using $y_{ij} \sim \text{Poisson}(\mu_{ij})$.

The simulated data sets were fit using both GSSIMM and the parametric log-linear mixed model (3.1). In this simulation, the GSSIMM estimates are much more accurate than the log-linear mixed model, which are badly biased. The simulation results of case 1 are summarized in Table 3.1. Since the GSSIMM estimates are better than the log-linear mixed model

estimates, they also do a reasonably effective job of fitting the data as we see Figure 3.1. Let $\eta(\mathbf{x}_{ij}^T\boldsymbol{\beta})$ be $1+\sin\frac{\pi(\mathbf{x}_{ij}^T\boldsymbol{\beta}-A)}{B-A}$. Then model (3.18) can be represented in the form of a GSSIMM using the following expression $\log(\mu_{ij}) = \eta(\mathbf{x}_{ij}^T\boldsymbol{\beta}) + \mathbf{z}_i^T\boldsymbol{b}$, where $\eta(\cdot)$ is a nonlinear function. For the data sets generated based on the "sine-bump" model (3.18) we would expect that the fit of the GSSIMM is to be better than the fit of GLMM since the GLMM can not be used to capture the nonlinear curvature in the data. The simulation results shown in Table 3.1 and Figure 3.1 agree well with our expectation.

	Log-linear Mixed model				GSSIMM			
Case 1	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\theta}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\theta}$
True Value	0.577	0.577	0.577	0.5	0.577	0.577	0.577	0.5
Estimate	-0.0564	-0.0161	0.0111	0.27	0.5813	0.5754	0.5742	0.32
Standard error	0.0369	0.0369	0.0368	0.54	0.0159	0.0151	0.0166	0.42
MSE	0.3438	0.3394	0.3272	0.18	0.0007	0.0004	0.0004	0.15

Table 3.1: Parameter estimation and average mean square errors obtained from log-linear mixed model and GSSIMM when the true model is $\log(\mu_{ij}) = 1 + \sin\frac{\pi(\mathbf{x}_{ij}^T\boldsymbol{\beta}-A)}{B-A} + \mathbf{z}_i^T\boldsymbol{b}$.

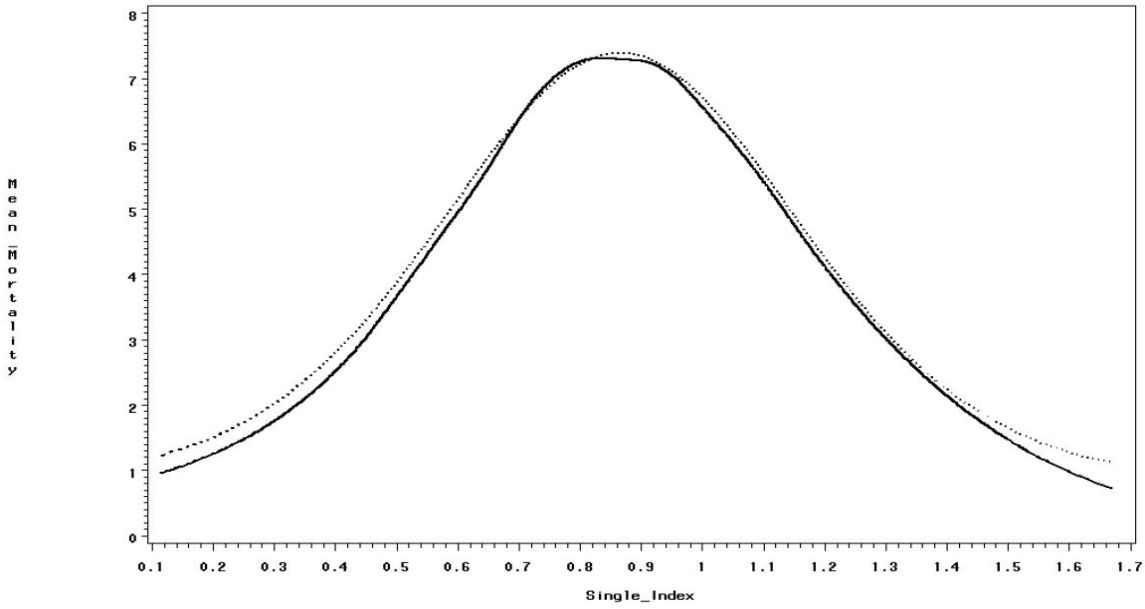

Figure 3.1: The estimated mean function when the true model is $\log(\mu_{ij}) = 1+\sin\frac{\pi(\mathbf{x}_{ij}^T\boldsymbol{\beta}-A)}{B-A} + \mathbf{z}_i^T\boldsymbol{b}$: Sold line represents true mean function and dotted line stands for the estimated mean function using GSSIMM.

The GSSIMM fit is also compared to the generalized additive mixed model (GAMM) fit.

The form of the GAMM is as the following expression

$$\log\{\mu(\mathbf{x},\mathbf{z})\} = \alpha_0 + f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) + f_3(\mathbf{x}_3) + \mathbf{z}^T\mathbf{b}. \tag{3.19}$$

The P-spline method is used to estimated the unknown functions for each of $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ in the GAMM where each P-spline is of the form (3.5) with $p = 2$. In this case, the GSSIMM has a smaller mean estimated deviance than the GAMM. Table 3.2 shows the mean estimated deviances of the simulations. The deviance is defined as $D(\mathbf{Y};\hat{\boldsymbol{\mu}}) = 2l(\mathbf{Y};\mathbf{Y}) - 2l(\hat{\boldsymbol{\mu}},\mathbf{Y})$, where $l(\mathbf{Y};\mathbf{Y})$ is the maximum likelihood achievable in a full model with n parameters, and $l(\hat{\boldsymbol{\mu}},\mathbf{Y})$ is the likelihood that achieved by the model under investigation (McCullagh and Nelder, 1989). The true deviance in Table 3.2 is computed based on model (3.18) used for generating the data, it can be treated as a reference deviance to assess the quality of fit for the other models. The fit of our GSSIMM is better than the fit of the GAMM for the data sets generated from the "sine-bump" model (3.18). Although the GAMM can be used to capture the nonlinear curvature in a nonlinear model, it can not be used to capture the single-index feature in the "sine-bump" model (3.18). The GSSIMM works well in capturing both nonlinear and single-index features in the data sets generated from the "sine-bump" model (3.18).

Case 1	GAMM	GSSIMM	True Model
Deviance	3980.9	2492.9	2530.9

Table 3.2: Mean deviation values between GAMM and GSSIMM when the true model is $\log(\mu_{ij}) = 1 + \sin\frac{\pi(\mathbf{x}_{ij}^T\boldsymbol{\beta}-A)}{B-A} + \mathbf{z}_i^T\boldsymbol{b}$.

A Monte Carlo study of asymptotic inference was conducted based on the sandwich formula (3.17). This study is designed in the similar way to the one in Yu and Ruppert's paper (2002). The simulation is conducted using the "sine-bump" model (3.18) we introduced previously. For each simulated covariate matrix $\mathbf{X}$, we simulated 200 response vectors $\mathbf{y}$, and then computed estimates of the parameter vector $\boldsymbol{\omega}$ using the method described in section 3.3. Let $\{\boldsymbol{\omega}\}_1^{200}$ denote the Monte Carlo sample of parameter estimates and $\mathbf{V}_{sample}$ the sample covariance matrix. Then $\mathbf{V}_{sample}$ is compared to the 200 sandwich covariance estimators $\mathbf{V}_{sw}$ computed using formula (3.17). $\mathbf{V}_{sample}$ is a Monte Carlo estimate of the true covariance matrix and is used in place of the true covariance matrix in assessing the accuracy of the sandwich formula. Table 3.3 shows the average estimated covariance matrix over 200 simulations and the Monte Carlo sample covariance matrix using the "sine-bump" model (3.18). As shown in Table 3.3, the estimated covariance matrix using the sandwich formula captures the independence among single-index coefficients since the off-diagonal elements are close to zero. The relative difference between the estimated covariance and the sample covariance is measured by

$$D = \frac{\frac{1}{200}\sum_{i=1}^{200}\text{norm}(\mathbf{V}_{sw,i} - \mathbf{V}_{sample})}{\text{norm}(\mathbf{V}_{sample})}.$$

The relative difference for this Monte Carlo study is 1.03. We use Frobenius norm for calculating the relative difference in this dissertation.

	Sample Covariance Matrix			Estimated Covariance Matrix		
	β_1	β_2	β_3	β_1	β_2	β_3
β_1	0.0046			0.0002		
β_2	-1.37e-05	0.0132		1e-04	0.0002	
β_3	-7.8e-06	1.59e-04	0.0129	9.03e-05	9.04e-05	0.0002

Table 3.3: Comparison between sample covariance matrix and estimated covariance matrix using sandwich method when the true model is $\log(\mu_{ij}) = 1 + \sin\frac{\pi(\mathbf{x}_{ij}^T\boldsymbol{\beta}-A)}{B-A} + \mathbf{z}_i^T\boldsymbol{b}$.

3.5.2 Case 2: Count Data Generated from Generalized Linear Mixed Model

A simulation was conducted based on the GLMM using the following expression

$$\log(\mu_{ij}) = 1 + \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_i^T\boldsymbol{b} \tag{3.20}$$

where the $\mathbf{x}_{ij}$ are trivariate with independent uniform [0,1] components, $\boldsymbol{b} \sim N(\mathbf{0}, 0.5\mathbf{I})$, $i \in (1, 2, \cdots, 12)$, $j \in (1, 2, \cdots, n_i)$ for each i, and $n_i = 200$ for all i's. Let $\mathbf{z}_i$ is a vector, where the ith element is 1, otherwise zero. The vector of parameters is $\boldsymbol{\beta} = (0.8, -0.5, 0.4)$. 1000 data sets were generated based on model (3.20).

We fit the data sets using both GSSIMM and GLMM. The results are shown in Table 3.4. The GLMM does a better job in estimating the regression coefficients with smaller MSEs and smaller bias. The estimate of the variance components using the GSSIMM is close to the true value compared to the estimate using the GLMM. Since the MSE of GSSIMM is quite small, the GSSIMM works well for estimating the parameters since model (3.20) is a special case of the GSSIMM. However, the MSE obtained from GLMM is 5 times smaller than that of GSSIMM, suggesting that GLMM is more effieicnet than GSSIM in this case 2.

We compare the GSSIMM fit to the generalized additive mixed model (3.19) fit. As in the previous case, a P-spline method is used to estimate the unknown functions for each of $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ in the GAMM (3.19) where each P-spline is of the form (3.5) with $p = 2$. The GAMM has a smaller mean estimated deviance than the GSSIMM as shown in Table 3.5 where the deviance is computed as described for case 1.

	Log-linear Mixed model				GSSIMM			
Case 2	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\theta}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\theta}$
True Value	0.8	-0.5	0.4	0.5	0.8	-0.5	0.4	0.5
Estimate	0.8012	-0.5004	0.4010	0.27	0.8122	-0.5563	0.4465	0.36
s.e.	0.0208	0.0207	0.0198	0.15	0.0506	0.0448	0.0497	0.14
MSE	0.0004	0.0004	0.0004	0.13	0.0022	0.0062	0.0051	0.06

Table 3.4: Parameter estimation and average mean square errors obtained from log-linear mixed model and GSSIMM when the true model is $\log(\mu_{ij}) = 1 + \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_i^T\boldsymbol{b}$.

Case 2	GAMM	GSSIMM	True Model
Deviance	2433.2	2487.5	2463.3

Table 3.5: Mean deviation values between GAMM and GSSIMM when the true model is $\log(\mu_{ij}) = 1 + \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_i^T\boldsymbol{b}$.

3.5.3 Case 3: Count Data Generated from a Generalized Additive Mixed Model

A third simulation was conducted based on the GAMM similar to Lin and Zhang (1999). 1000 data sets are generated from the following model

$$\log(\mu_{ij}) = 1 + x_{1ij} + f_1(x_{2ij}) + f_2(x_{3ij}) + \mathbf{z}_i^T\boldsymbol{b} \tag{3.21}$$

where the $\mathbf{x}_{ij}$ are trivariate with independent uniform [0,1] components, $\boldsymbol{b} \sim N(\mathbf{0}, 0.5\mathbf{I})$, $i \in (1, 2, \cdots, 12)$, $j \in (1, 2, \cdots, n_i)$ for each i, and $n_i = 200$ for all i's. Define $\mathbf{z}_i$ is a vector, where the ith element is 1, otherwise zero. The true functions are defined as mixtures of Beta functions $f_1(x) = 1/3\{2F_{8,8}(x) + F_{5,5}(x)\}$, and $f_2(x) = 1/10\{6F_{30,17}(x) + 4F_{3,11}(x)\}$, where $F_{p,q}(x) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}x^{p-1}(1-x)^{q-1}$.

The data sets were fit using the GAMM, GSSIMM, and GLMM. The mean deviances are shown in Table 3.6, where the deviance is computed as described for case 1. The mean deviance of GAMM has the smallest value. The mean deviance of GLMM has the largest value. Figure 3.2 shows the estimated nonparametric curves using GAMM. The GAMM provides the best fit for the data generated from model (3.21). Our GSSIMM fits the data better than the GLMM. The results suggest that the GSSIMM does not fit well to the data generated from a GAMM compared to the fit of the GAMM. However, the GSSIMM fit is better than the parametric GLMM fit.

Case 3	GAMM	GSSIMM	GLMM	True Model
Deviance	2431.2	3745.6	6213.9	2452.4

Table 3.6: Comparison of deviance values of GAMM, GSSIMM, GLMM when the true model is $\log(\mu_{ij}) = 1 + x_{1ij} + f_1(x_{2ij}) + f_2(x_{3ij}) + \mathbf{z}_i^T \boldsymbol{b}$.

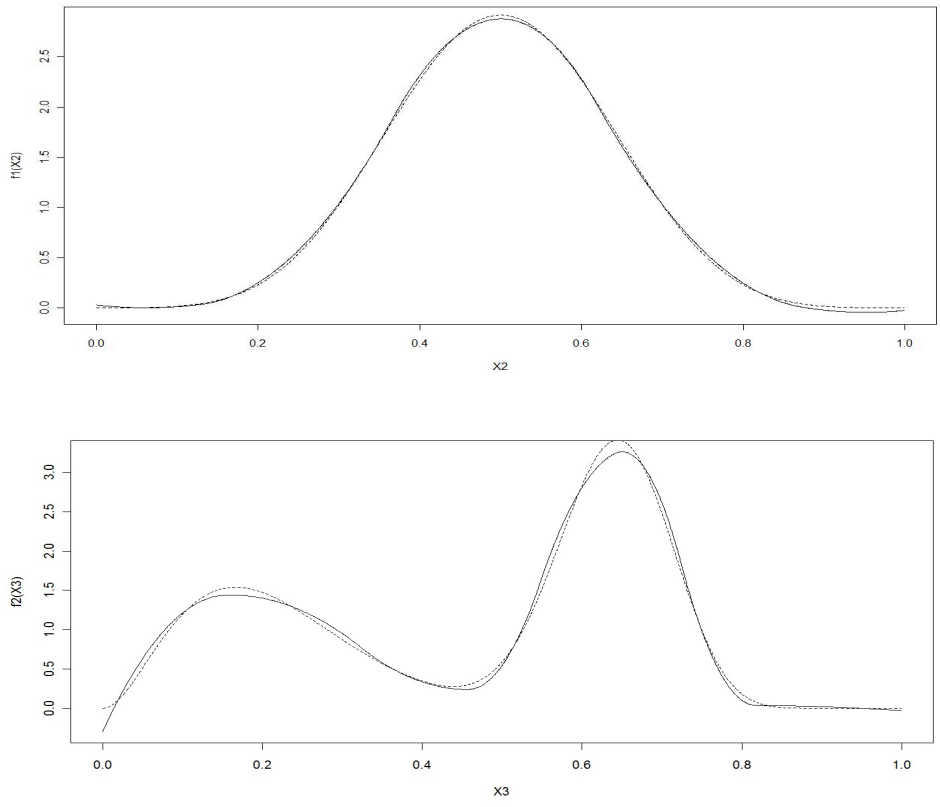

Figure 3.2: Estimated nonparametric curves for a single replication of the datasets generated from the GAMM. The solid curve correspond to the estimates of the underlying mean function. The dashed curve is the true mean function.

3.6 Application

3.6.1 Health Effect of Air Pollution

In the past 30 years, outdoor levels of some pollutants such as particulates, sulfur oxides, and carbon monoxide, have been declining in many US and western European cities, thanks to emission controls on vehicles, heating, power generation and industry (US EPA, 1999). However, many outdoor air quality problems still exist in the developed world and may be worsened by increased use of motor vehicles and industrial chemicals. Severe outdoor air pollution problems exist in the developing world, especially in large cities such as Beijing, Chongqing, Bombay, Karachi, Cairo, Sao Paulo and Mexico City.

Levels of priority air pollutants often exceed safety limits in many parts of the world, especially in large cities in developing countries. For example, it was estimated in 2004 that 18% of the world's urban areas (cities with a population of over 100,000) have ambient air containing an annual mean of over 100 or more than twice the US EPA limit (Cohen *et al.*, 2005). It should be noted that adverse health effects have been documented at levels well below these official US EPA standards.

Outdoor air quality plays an important role in human health. Air pollution causes large increases in medical expenses, morbidity, and is estimated to cause about 800,000 annual premature deaths worldwide (Cohen *et al.*, 2005). The outdoor air often contains biologically significant levels of many pollutants including particulates, ozone, carbon monoxide, oxides of nitrogen and sulfur, bioaerosols, metals, volatile organics, and pesticides. A large percentage of these pollutants are produced by anthropogenic activities. While most people spend the majority of their time indoors, outdoor air quality can affect indoor air quality to a large degree. In addition, many patients such as asthmatics, patients with allergies and chemical sensitivities, pregnant women, the elderly and children are especially susceptible to poor outdoor and indoor air quality. It is now well documented that higher levels of many airborne pollutants can affect many of the body's systems adversely, including respiratory, cardiovascular (including heart and brain), resproductive/developmental, and neurological/neuropsychiatric systems. Air pollution has also been shown in some studies to increase rates of infection, cancer, and mortality. Especially well documented are the respiratory and cardiovascular effects of common air pollutants ($PM_{10}/PM_{2.5}$, O_3, NO_2, SO_2, and CO), which are well below standards set by US EPA, WHO, and other agencies. In addition, the outdoor air often contains significant levels of many other pollutants such as metals (lead, mercury, cadmium, manganese, and nickel), isocyanates, ethylene oxide, aldehydes (acrolein, formaldehyde), and other volatile organic chemicals. The health effects (such as cancer or asthma) of occupational exposure to these chemicals are well known, however few studies have looked at the health effects of ambient air exposures from metals or volatile organics. Combinations of many of these air pollutants can have additive or synergistic adverse health effects.

As an example of our methodology, we analyze the daily effect of $PM_{2.5}$ and ozone concentration on mortality using data from twelve counties in North Carolina in the years of 2004 and 2005 (Fig. 3.3).

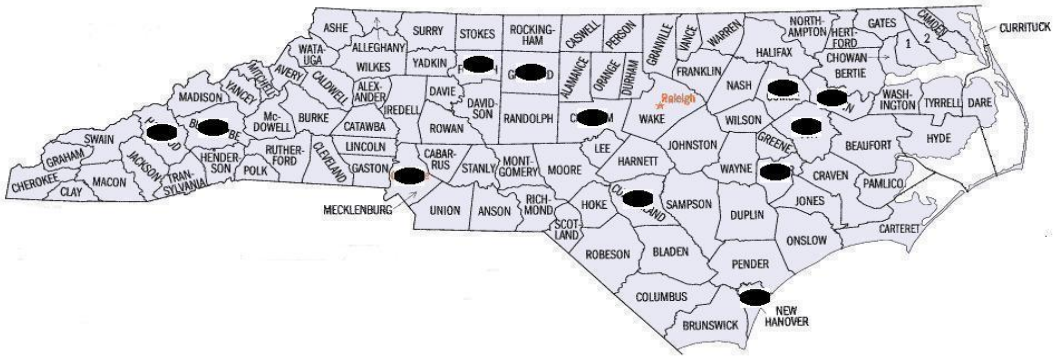

Figure 3.3: The Counties in North Carolina included in the study. Include (solid dots) Buncombe, Chatham, Cumberland, Edgecombe, Forsyth, Guilford, Haywood, Lenoir, Martin, Mecklenburg, New Hanover, and Pitt.

3.6.2 Application for Generalized Semiparametric Single-Index Mixed Model

The response variable of interest is the daily mortality with the major cause of death being cardiovascular diseases. Since the data on daily mortality is count data, a generalized linear model is used to model the count data.

Although much research has been conducted to study the association between daily mortality and daily concentration of air pollutants, the true underlying relation is still unknown. Nonparametric regression can be of substantial value in the solution of complex problems such as one described in this example. We chose the technique of nonparametric regression to describe the association between daily mortality and daily concentration of air pollutants.

Ozone can be produced by a number of processes such as lightening and by atmospheric reactions involving volatile organic chemicals, nitrogen, oxides, and sunlight. This indirect ozone production is most efficient during warm weather. Carbon monoxide is a product of incomplete combustion. It is correlated with many other air pollutants including sulfur oxides. Vehicle combustion is known to produce carbon monoxide, sulfur oxides, and nitrogen oxides. Since correlations exist among air pollutants and climate variables, the single-index model is considered. A single-index, composed of the linear combination of environmental factors (air pollutants and climate factors), will be used in the GSSIMM. This single-index provides

information on how these environmental factors are related with mortality synergistically.

Based on the above concerns for modeling North Carolina mortality and air pollution data, GSSIMM (3.3) is applied to study the association between daily mortality and daily air pollutants adjusted for climate variables in North Carolina. As we explained in the introduction, the response observation y_{ij} is the jth daily mortality with major cause of death as cardiovascular disease in the ith county. The $\mathbf{x}_{ij}$ is the vector of the jth daily observations for the independent variables in the ith county. There are four independent variables to be used in the single-index: $PM_{2.5}$, Ozone, average temperature, and wind speed.

If the response variable Y follows a Poisson distribution, model (3.3) has the form

$$\log\{\mu_{ij}\} = \eta(\mathbf{x}_{ij}^T\boldsymbol{\beta}) + \mathbf{z}_i^T\boldsymbol{b}, \quad \text{with} \quad \|\boldsymbol{\beta}\| = 1,$$

where $y_{ij} \sim \text{Poisson}(\mu_{ij})$. The unknown function is $\eta(\cdot)$, and the unknown vector of single-index coefficients is $\boldsymbol{\beta}$. The single random effect for the ith county is denoted as $\mathbf{z}_i^T\boldsymbol{b}$ with latent variable $\boldsymbol{b} \sim N(\mathbf{0}, \boldsymbol{\Sigma_b})$, and $\boldsymbol{\Sigma_b}$ is the covariance matrix, $\mathbf{z}_i$ is the vector associate with the random effect for the ith county. To obtain the maximum likelihood estimate, we apply the iterative estimation procedure describes in section 3.3.2. We present the results of fitting a quadratic P-spline with 10 knots to the North Carolina data.

In the first example, mortality, lagged by one day, is modeled as a function of $PM_{2.5}$, ozone, average temperature, and average speed, and their effects are tested . We denote by $x_{PM_{ij}}$, $x_{O_{ij}}$, $x_{T_{ij}}$, and $x_{W_{ij}}$ as the jth daily observation in the ith county for $PM_{2.5}$, ozone, average temperature, and average wind speed respectively. The model has the following form

$$\log\{\mu_{ij}\} = \eta(x_{PM_{ij}}\beta_{PM} + x_{O_{ij}}\beta_O + x_{T_{ij}}\beta_T + x_{W_{ij}}\beta_W) + \mathbf{z}_i^T\boldsymbol{b}, \tag{3.22}$$

where β_{PM}, β_O, β_T, and β_W are the single-index coefficients for $PM_{2.5}$, ozone, average temperature, and average wind speed respectively.

The plot of mean mortality vs. the index for all 12 selected counties is shown in Figure 3.4, 3.5, and 3.6. We notice that there is a non-monotone increasing trend in the plots. The mean daily mortality increases with increasing index for all 12 counties of interest. However, the magnitudes of the mean daily mortality and the rates of increasing mean daily mortality are varied for different counties. These results indicate the existence of a random effect in the GSSIMM for North Carolina data. From top curve on Figure 3.6, it is noted that Mecklenburg county has the highest daily mortality level and the largest increasing rate of mortality corresponding to the increasing index. Martin county, the bottom curve on Figure 3.6, has the lowest daily mortality level and the smallest increasing rate of mortality as a function of the index. The difference between Mecklenburg and Martin county may be due to the fact that Mecklenburg county has the largest city in North Carolina, Charlotte, and that Martin county is rural. Mecklenburg county has a higher population density and more industry which could produce more air pollutants compared to Martin county.

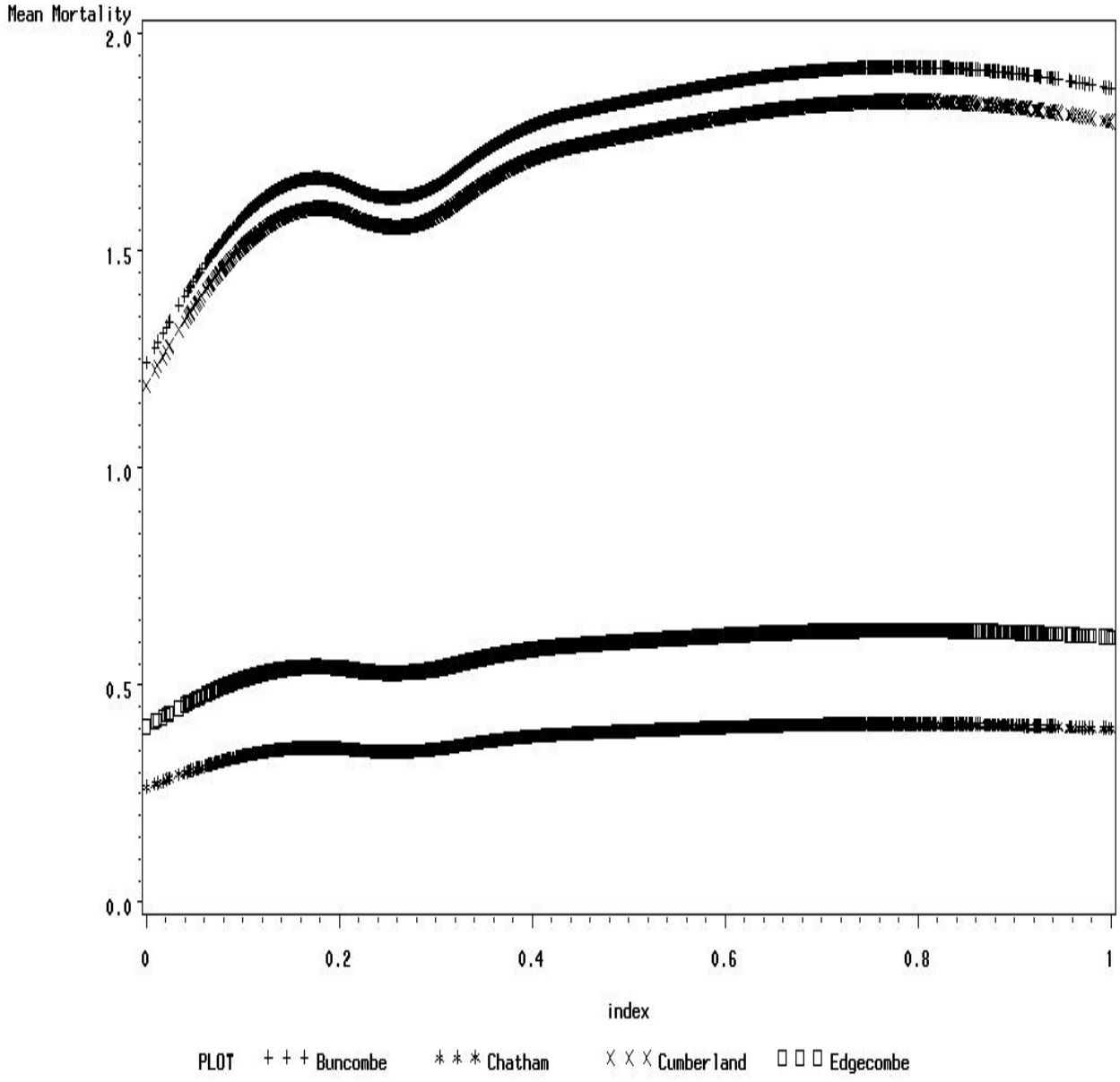

Figure 3.4: Plot of mean mortality vs. index for Buncombe, Chatham, Cumberland, and Edgecombe Counties in model (3.22).

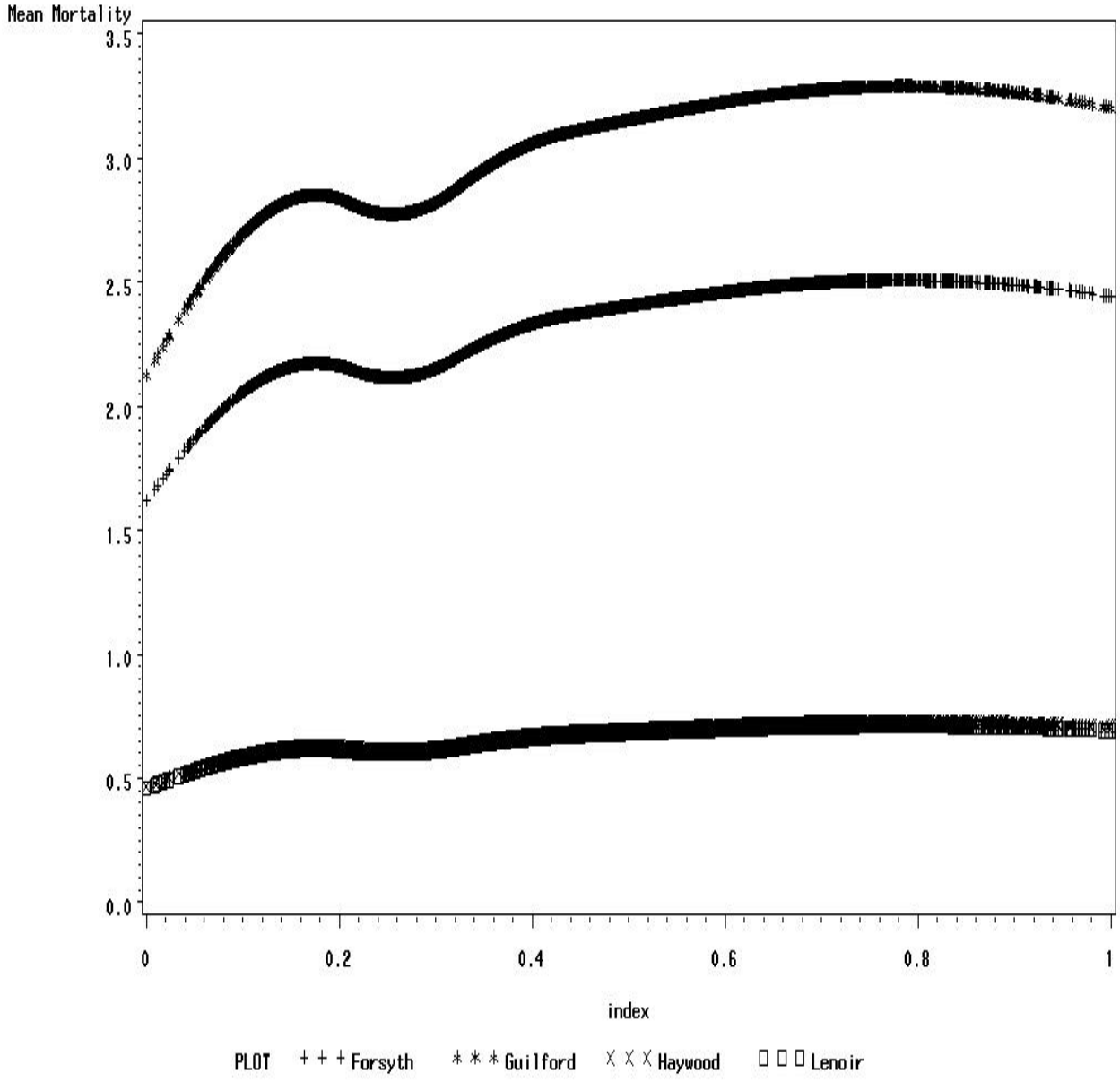

Figure 3.5: Plot of mean mortality vs. index for Forsyth, Guilford, Haywood, and Lenoir Counties in model (3.22).

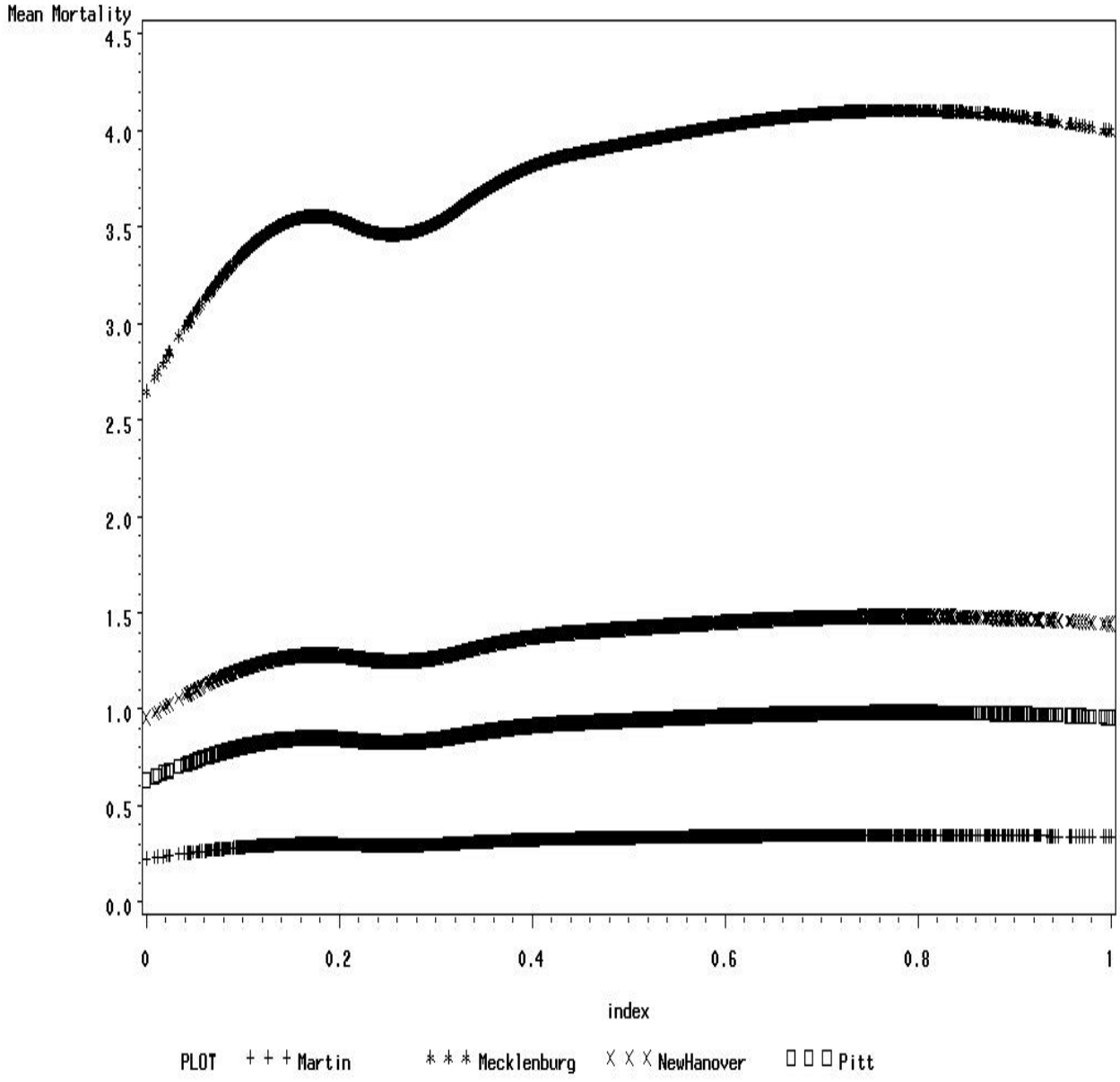

Figure 3.6: Plot of mean mortality vs. index for Martin, Mecklenburg, NewHanover, and Pitt Counties in model (3.22).

The estimated curve, $\hat{\eta}(\cdot)$, for Mecklenburg county is presented in plot on the top of Figure 3.7. There is a non-monotone increasing trend in mean daily mortality with increasing index. The 95% bootstrap confidence limit for the estimated mean function is shown in the plot on the bottom of Figure 3.7. The confidence band is wide in the tail areas of single-index values since relative small number of observations in these areas. It is useful to develop a method for calculating the confidence band for the estimated mean function in future research.

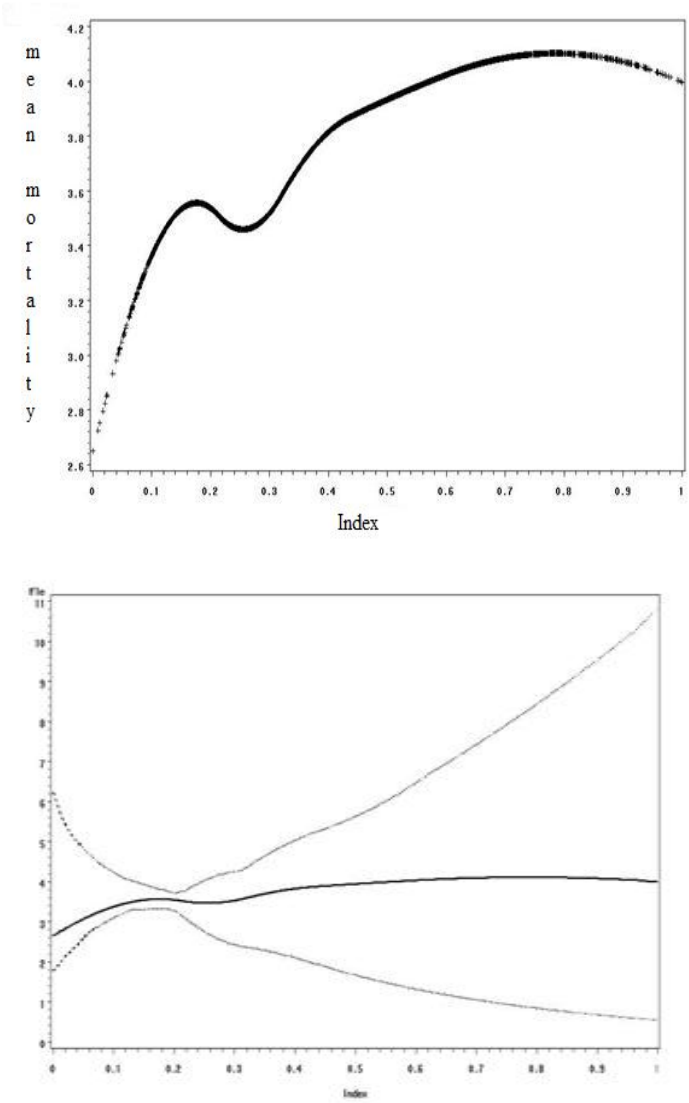

Figure 3.7: Estimate for function $\eta(\cdot)$ for Mecklenburg county in model (3.22).

The estimated single-index coefficients for $PM_{2.5}$, ozone, average temperature, and average wind speed, and their 95% bootstrap confidence intervals are shown in Table 3.7. The significant coefficient for $PM_{2.5}$ suggest that a higher level of $PM_{2.5}$ contributes to the in-

creasing of the index, consequently resulting in increased mean mortality. The estimated single-index coefficient for ozone is positive, but is non-significant. The negative single-index coefficient for average temperature is significant, which indicates larger mortality in cold weather. In practice, lower wind speed results in higher levels of the air pollution. Thus, the coefficient for average wind speed is expected to have a different sign than the coefficient of $PM_{2.5}$ and ozone. However, our estimate is positive. This is an issue deserving further study. One possible reason may be that higher wind speed may also increase the mortality since North Carolina is a coastal state. For instance, hurricanes making landfall may result in an increase in mortality.

	β_{PM}	β_O	β_T	β_W
Estimation	0.1943	0.8432	-0.3669	0.3416
Bootstrap CI	[0.1366,0.2521]	[-1.0422,2.7286]	[-0.3818,-0.3519]	[0.2310,0.4522]

Table 3.7: Parameter estimation and bootstrap confidence interval of single-index coefficients for four environmental factors in model (3.22).

The Wald test was introduced in section 3.4.2 to test the significance of the single-index coefficients using sandwich formula (3.17). The results of the tests are shown in Table 3.8. The P-value in the test is computed assuming a χ^2_1 distribution for the test statistic under H_0. The single-index coefficient for ozone is statistically not significant while the single-index coefficients for three other environmental factors are statistically significant. These results agree well with results using bootstrap confidence intervals shown in Table 3.7.

H_0	$\beta_{PM} = 0$	$\beta_O = 0$	$\beta_T = 0$	$\beta_W = 0$
P-value	< 0.0001	0.84096	< 0.0001	< 0.0001

Table 3.8: Results of the Wald test for the significance of single-index coefficients in model (3.22) using sandwich formula.

Since the single-index coefficient for ozone in model (3.22) is not statistically significant, we remove ozone in model (3.22) and fit the following model

$$\log\{\mu_{ij}\} = \eta(x_{PM_{ij}}\beta_{PM} + x_{T_{ij}}\beta_T + x_{W_{ij}}\beta_W) + \mathbf{z}_i^T\boldsymbol{b}. \tag{3.23}$$

The estimated single-index coefficient and results of test of significance are shown in Table 3.9. In model (3.23), the single-index coefficients for $PM_{2.5}$ and average wind speed have significant positive estimates while the single-index coefficient for average temperature has a significant negative estimate. The P-values are computed based on the Wald test using sandwich formula (3.17). The estimated curve, $\hat{\eta}(\cdot)$, for Mecklenburg county in model (3.23) is presented in Figure 3.8. There is a non-monotone increasing trend in mean daily mortality with increasing index.

	β_{PM}	β_T	β_W
Estimation	0.5009	-0.4325	0.7497
P-value	< 0.0001	< 0.0001	< 0.0001

Table 3.9: Parameter estimation and P-value for testing the significance of single-index coefficients for three environmental factors in model (3.23).

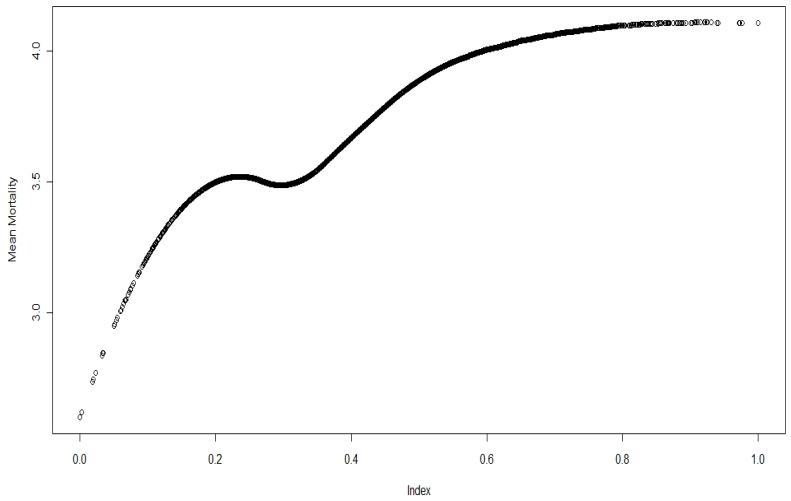

Figure 3.8: Plot of mean mortality vs. index for Mecklenburg county in model (3.23).

We use the deviance to assess the quality of the fit to the NC data using model (3.23). The deviance is defined as $D(\mathbf{Y};\hat{\boldsymbol{\mu}}) = 2l(\mathbf{Y};\mathbf{Y}) - 2l(\hat{\boldsymbol{\mu}},\mathbf{Y})$, where $l(\hat{\boldsymbol{\mu}},\mathbf{Y}) = \sum_i \log f_i\{y_i|\theta_i = g^{-1}(\mu_i)\}$ (McCullagh and Nelder, 1989). Here, $f(\cdot)$ is the distribution function in the exponential family for the response variable. The deviance has an exact χ^2 distribution for the normal linear model (under the assumption that the model is correct). In the GLM framework, the deviance is approximately distributed as χ^2_{n-p}, where n is the number of observations in the data, and p is the number of fitted parameters under H_0. This distribution assumption for the deviance is used to apply the deviance as a diagnostic for assessing the adequacy of the fitted model. To assess the adequacy of model (3.23), the H_0 is that model (3.23) has no lack-of-fit for fitting the NC air pollution data. The P-value is 0.4785. Thus, we conclude that model (3.23) provides an appropriate fit for NC data. The fitted curves overlayed on observed data are shown in Figure 3.9. The counties, which have large number of observations, tend to have better fits. For the counties with too many zeros, like Martin, a zero-inflated Poisson model may be needed in future research. In the GLM, the deviance residuals can be used to explore the adequacy of fit of a model, with respect to choice of variance function, link function and terms in the linear predictor. The deviance residual is defined as $r_D = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}$, where $d_i = 2l(y_i;y_i) - 2l(\hat{\mu}_i, y_i)$. Figure 3.10 are the

deviance residual vs. $\hat{\mu}$ plots for each county. Figure 3.10 shows that the deviance residuals are approximately equally dispersed around the horizontal axis.

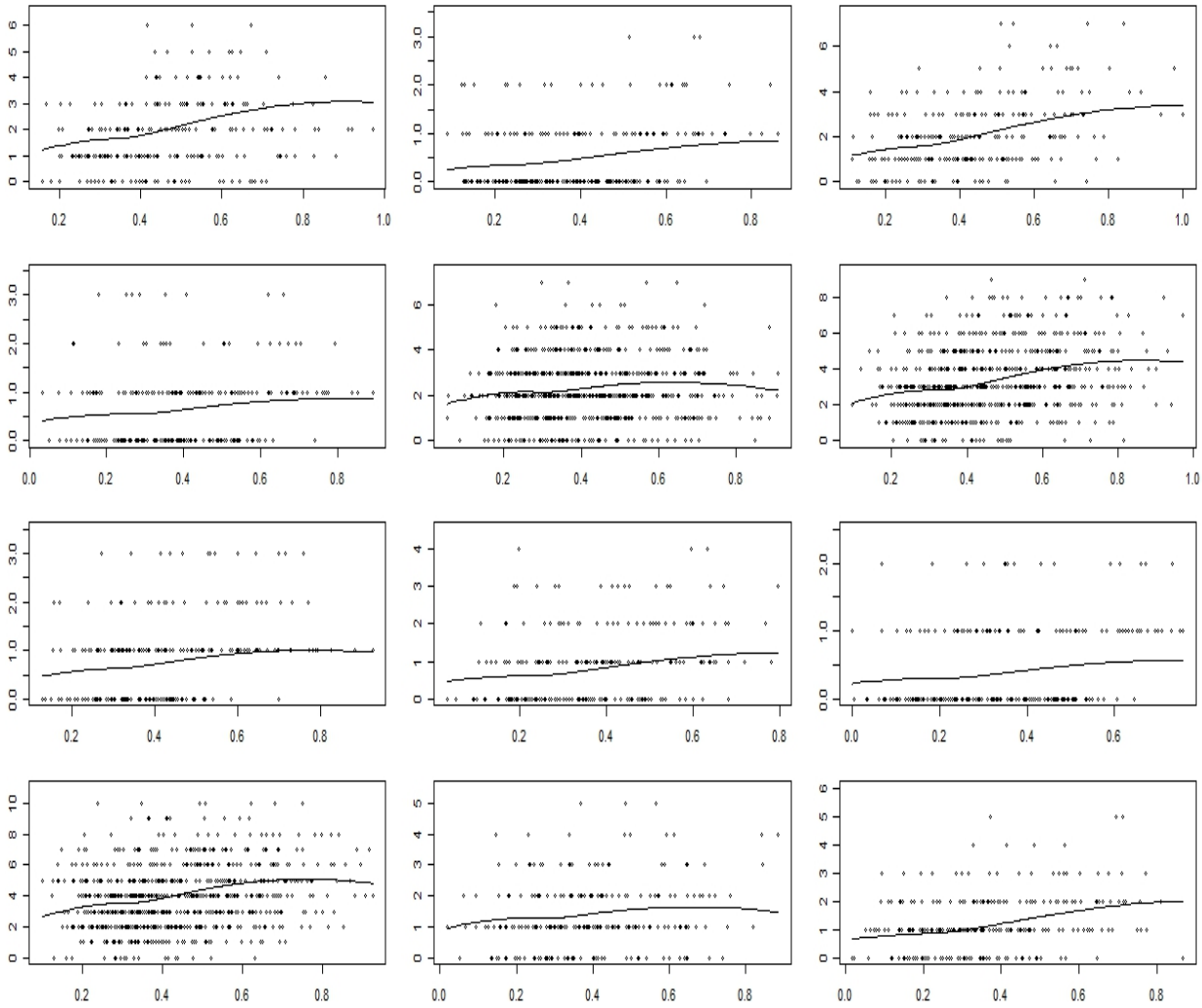

Figure 3.9: The fitted curves overlayed on observed data for each county for model (3.23). The counties on top panel: Buncombe, Chatham, Cumberland; the counties in the first middle panel: Edgecombe, Forsyth, Guilford; the counties in the second middle panel: Haywood, Lenoir, Martin; the counties in the bottom panel: Mecklenburg, New Hanover, Pitt.

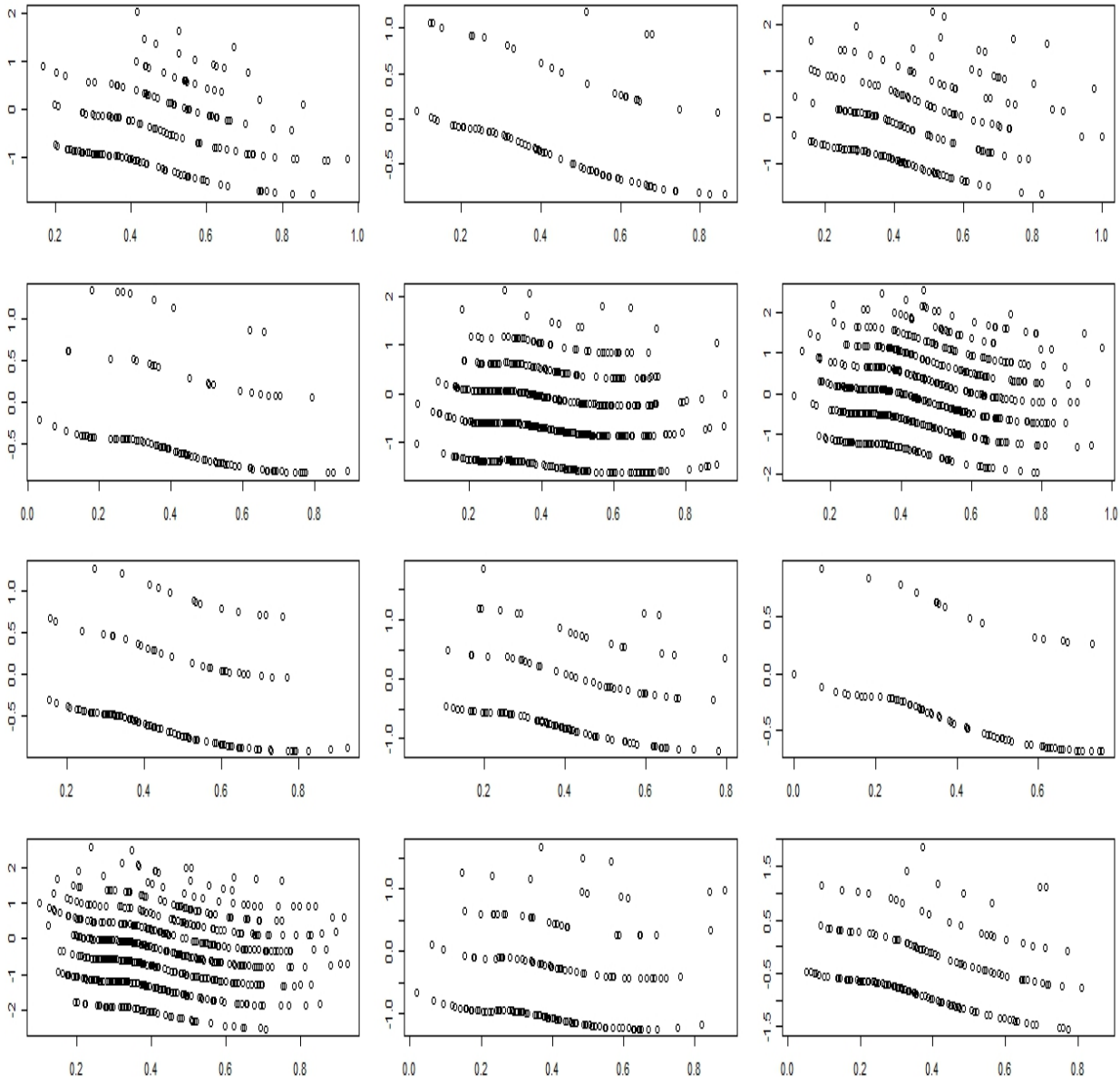

Figure 3.10: The deviance residual vs. $\hat{\mu}$ plots for each county for model (3.23). The counties on top panel: Buncombe, Chatham, Cumberland; the counties in the first middle panel: Edgecombe, Forsyth, Guilford; the counties in the second middle panel: Haywood, Lenoir, Martin; the counties in the bottom panel: Mecklenburg, New Hanover, Pitt.

We also use the GAMM to fit the NC data. The model has the following form

$$\log\{\mu_{ij}\} = \alpha_0 + f_1(x_{PM_{ij}}) + f_2(x_{T_{ij}}) + f_3(x_{W_{ij}}) + \mathbf{z}_i^T \boldsymbol{b}.$$

A P-spline, which has the form (3.5) with $p = 2$, is used to estimated the unknown functions for each of $f_1(\cdot)$, $f_2(\cdot)$, and $f_3(\cdot)$ in the GAMM. The estimated nonparametric curves are shown in Figure 3.11. There is no trend shown in the nonparametric curve for the covariate, average temperature. Although, there are increasing and decreasing trends in the nonparametric curves for the covariates $PM_{2.5}$ and average wind speed respectively, the curves are relatively flat in the mean area of the normalized covarates values. The deviances are 5320.87 and 4431.55 for the GAMM and GSSIMM respectively. The results suggest that the GSSIMM may provides a better fit to the NC data compared to the GAMM.

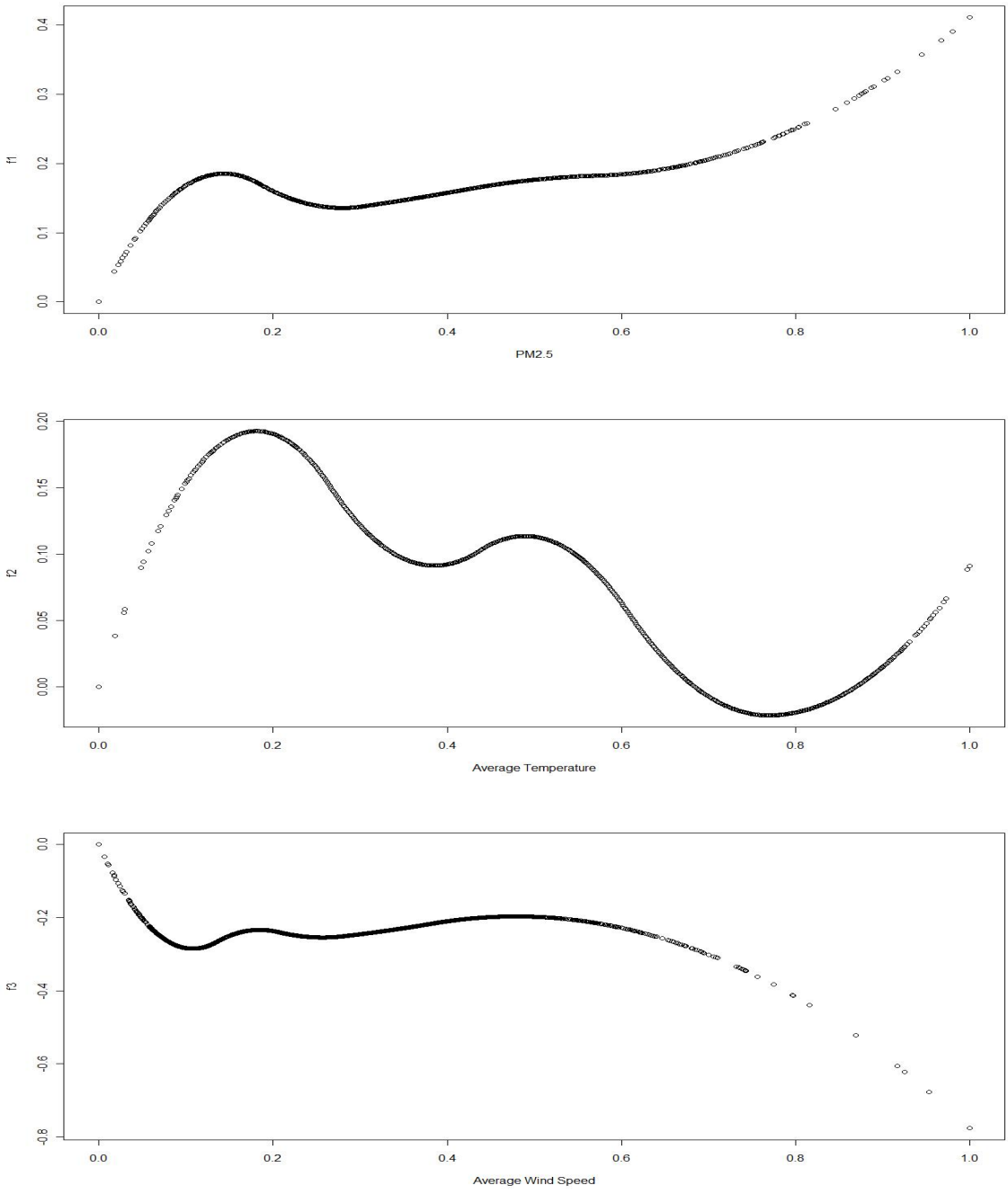

Figure 3.11: Estimated nonparametric curves in GAMM for NC data. The plot on the top is the estimated curve for the covariate, $PM_{2.5}$. The plot in the middle is the estimated curve for the covariate, average temperature. The plot at the bottom is the estimated curve for the covariate, average wind speed.

3.7 Summary

We have proposed the generalized semiparametric single-index mixed model. To avoid high dimensional integration, DPQL was proposed to make approximate inference. P-splines were used to estimate the nonparametric function and marginal quasi-likelihood was used to make inference on the variance components. The iterative two step estimation approach is computationally efficient and stable in practice. The implementation can be achieved by existing routines of the statistical software SAS and standard nonlinear optimization with constraints. Assuming a fixed number of knots, we show $\sqrt{n}$ consistency and asymptotic normality of the estimators. The sandwich estimate of the covariance matrix enables joint inference of all parameters. The simulation examples presented and the study of health effect of air pollution in North Carolina illustrate the effectiveness of our approach.

Chapter 4

Conclusions and Discussion

4.1 Conclusions

In this dissertation, we proposed two models which are nonparametric extensions of the GLM. In the semiparametric generalized linear model with log-concave random component (SGLM-L), the estimate of the distribution of the random component has a nonparametric form while the estimate of the systematic part has a parametric form. In the generalized semiparametric single-index mixed model (GSSIMM), the single-index model is incorporated into the GLM and further extended to GLMM assuming that the random component follows a parametric distribution. The SGLM-L and GSSIMM complement each other in the application of semiparametric methods for the GLM.

In both models, the estimators are obtained based on iterative maximum likelihood estimation procedure. We focus on maximum likelihood estimation with log-concave constraint on the estimator for SGLM-L, and penalized quasi-likelihood estimation for GSSIMM. They are computationally efficient and stable, and implemented by existing routines of the statistical software SAS and R.

The semiparametric generalized linear model with log-concave random component is proposed as an alternative method for modeling linear associations such as linear regression. There are different interpretation of parameters between the SGLM-L and the linear regression. The simulation study shows us that the SGLM-L approach fits well for non-normal data because we avoid any specific parametric distributional form for the random component. The SGLM-L may be used for various data sets whose response follow a distribution with the log-concave property.

In the generalized semiparametric single-index mixed model, the marginal quasi-likelihood was used to make inference on the variance components based on the connection between the nonparametric models and mixed models. Assuming a fixed number of knots, the $\sqrt{n}$

consistency and asymptotic normality of the estimators are provided. The sandwich estimator of the covariance matrix enables joint inference of all parameters. The simulation results show the performances of the GSSIMM, GAMM and GLMM in fitting different kind of simulated data sets. The study of health effect of air pollution in North Carolina show that GSSIMM is useful in multivariate nonparametric analysis.

4.2 Discussion

In the simulation study for SGLM-L, there is only one independent variable considered. However, it is feasible to extend the simulation to the cases including multiple independent variables since the related formulas are written in the vector or matrix forms. The log-likelihood ratio test is used in SGLM-L. The power study based on simulation presents evidence that the log-likelihood ratio test statistic (2.7) reasonably follows a chi-square distribution. However, the theoretical proof of this distribution assumption is not established. It will be helpful to further develop the theoretical justification for this distribution assumption. It would also be useful to extend a method for visualizing the contribution to a likelihood function in the SGLM-L in future research.

Our methodology for the GSSIMM can be further developed for other research such as Bayesian and spatial studys. The penalized splines can be viewed as hierarchical Bayesian models (Gu, 2001). By incorporating appropriate prior knowledge, the estimates can be obtained by using the Bayesian method. In this dissertation, the vector of random effects $\mathbf{b}$ in GSSIMM is assumed to follow a multi-normal distribution as $\mathrm{N}(\mathbf{0}, \mathbf{\Sigma})$. However, it will be useful to extend it to other distributions. An example is the spatial model with spatially correlated random effects.

Bibliography

Aitchison, J. and Silvey, S. D. (1960), "Maximum-likelihood Estimation Procedures and Associated Test of Significance", *Journal of the Royal Statistical Society*, Series B, **22**, 154-171.

Bellman, R. E. (1961), *Adaptive Control Processes, Princeton*, NJ: Princeton University Press.

Breslow, N. E., and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models", *Journal of the American Statistical Association*, **88**, 9-25.

Breslow, N. E., and Lin, X. (1995), "Bias Correction in Generalised Linear Mixed Models with a Single Component of Dispersion", *Biometrika*, **82**, 81-91.

Calder, C. A., Holloman, C. H., Bortnick, S. M., Strauss, W., and Morara, M. (2008), "Relating Ambient Particulate Matter Concentration Levels to Mortality Using Exposure Simulator", *Journal of American Statistical Association*, **103**, 137-148.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997), "Generalized Partially Linear Single-Index Models", *Journal of the American Statistical Association*, **92**, 477-488.

Cohen, A. J., Ross, A. H., Ostro, B, Pandey, K. D., Kryzanowski, M, Kunzail, N, *et al.* (2005), "The Global Burden of Disease Due to Outdoor Air Pollution", *Journal of Toxicol Environment Health*, **68**, 1-7.

Crouch, E. A. C., and Spiegelman, D. (1990), "The Evaluation of Integrals of the Form $\int f(t)\exp(-t^2)dt$: Application to Logistic Normal Models", *Journal of the American Statistical Association*, **85**, 464-469.

Curtis, L., Rea, W., Smith-Wills, P., Fenyves, E., and Pan, Y., (2006), "Adverse Health Effects of Outdoor Air Pollutants", *Environment International*, **32**, 815-830.

Diggle, P. J., Heagerty, P., Liang, K., and Zeger, S. L. (2002), *Analysis of Longitudinal Data*, Oxford: Oxford University Press.

Dumbgen, L., Husler, A., Rufibach, K. (2007), "Active Set and EM Algorithms for Log-Concave Densities Based on Complete and Censored Data", Preprint, University of Bern.

Eno, D. R., Terrell, G. R. (1999), "Scatterplots For Logistic Regression", *Journal of Computational and Graphical Statistics*, **8**, 413-425.

Fletcher, R. (1987), *Practical Methods of Optimization (2nd edition)*. New York: Wiley.

Friedman, J. H. and Stuetzle, W. (1981), "Projection Pursuit Regression", *Journal of the American Statistical Association*, **76**, 817-823.

Green, P. J. and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models*, New York: Chapman Hall.

Groeneboom, P., Jongbloed, G., and Wellner, J. A. (2001), "Estimation of Convex Function: Characterization and Asymptotic Theory", *Annals of Statistics*, **29**, 1653-1698.

Groeneboom, P., Jongbloed, G., and Wellner, J. A. (2003), "The Support Reduction Algorithm for Computing nonparametric function estimates in mixture models", Technical Report 2002-13, Department of Mathematics, Vrijie Universiteit Amsterdam.

Gu, C. (2001), *Smoothing Spline ANOVA Models*, New York: Springer.

Hardle, W., Hall, P., and Ichimura, H. (1993), "Optimal Smoothing in Single-Index Models", *The Annals of Statistics*, 21, 157-178.

Hardle, W., and Stoker, T. M. (1989), "Investing Smooth Multiple Regression by the Method of Average Derivatives", *Journal of the American Statistical Association*, **84**, 986-995.

Harville, D. A. (1977), "Maximum Likelihood Approaches to Variance Component Estimation and to Ralated Problems", *Journal of the American Statistical Association*, **72**, 320-340.

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, London: Chapman and Hall.

Hinde, J. (1982), "Compound Poisson Regression Models", in GLIM 82: *Proceedings of the International Conference on Generalized Linear Models*, ed. R. Gilchrist, Berline: Springer, pp. 109-121.

Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models", *Journal of Econometrics*, **58**, 71120.

Jiang, J. (2007), *Linear and Generalized Linear Mixed Models and Their Applicaitons*, New York: Springer.

Kleem, R. J., Mason, R. J., Heilig, C. M., Neas, L. M., and Dockery, D. W. (2000), "Is Daily Mortality Associated Specifically With Fine Particles? Data Reconstruction and Replication of Analysis", *Journal of Air and Waste Management Association*, **50**, 1215-1222.

Li, K. C., and Duan, N. (1989), "Regression Analysis Under Link Violation", *The Annals of Statistics*, **17**, 327-336.

Lin, X., and Breslow, N. E. (1996), "Bias Correction in Generalized Linear Mixed Models with Multiple Components of Dispersion", *Journal of the American Statistical Association*, **91**, 1007-1016

Lin, X. and Zhang, D. (1999),"Inference in Generalized Additive Mixed Models By Using Smoothing Splines", *Journal of Royal Statistical Society*, B, **92**, 162-190.

Mauderly, J. L., and Samet, J. M. (2009), "Is There Evidence for Synergy Among Air Pollutants in Causing Health Effects?", *Environmental Health Perspectives*, **117**, 1-6.

McCullagh, P. and Nelder, J. A. (1989), *Generalized linear models*, New York: Chapman Hall.

McCulloch, C. (1997), "Maximum Likelihood Algorithm for Generalized Linear Mixed Models", *Journal of the American Statistical Association*, **95**, 227-237.

Natarajan, R., and Kass, R. E. (2000), "Reference Bayesian Methods for Generalized Linear Mixed Models", *Journal of the American Statistical Association*, **90**, 141-150.

Rufibach, K. (2007), "Computing Maximum Likelihood Estimates of a Log-concave Density Function", *Journal of Statistical Computation and Simulation* **77**, 561-574.

Rufibach, K. and Dumbgen, L. (2004), "Maximum Likelihood Estimation of a Log-concave Density: Basic Properties and Uniform Consistency", Preprint, Department of Mathematical Statistics and Actuarial Science, University of Bern.

Ruppert, D. (2002), "Selecting the Number of Knots for Penalized Splines", *Journal of Computational and Graphical Statistics*, **11**, 735-757.

Ruppert, D., and Carroll, R. (1997), "Penalized Regression Splines", Preprint, Cornell University, School of Operations Research and Industrial Engineering.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge: Cambridge University Press.

Shao, J. (2003), *Mathematical Statistics*, New York: Springer.

Stoker, T. M. (1986), "Consistent Estimation of Scaled Coefficients", *Econometrica*, **54**, 1461-1481

Stiratelli, R., Laird, N., and Ware, J. (1984), "Random Effects Models for Serial Observations with Binary Response", *Journal of Statistical Computation and Simulation*, **42**, 1-20.

Tierney, L., and Kadane, J. B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities", *Journal of the American Statistical Association*, **81**, 8286.

Yu, Y., and Ruppert, D. (2002), "Penalized Spline Estimation for Partial Linear Single-Index Models", *Journal of the American Statistical Association*, **97**, 1042-1054.

Yu, Y. (2008),"Penalized Spline Estimation for Generalized Partielly Linear Single-Index Models", Preprint.

Zeger, S. L., and Karim, M. R. (1991), "Generalized Linear Models with Random Effects: a Gibbs Sampling Approach", *Journal of the American Statistical Association*, **86**, 79-86.

Zhang, D. (2004), "Generalized Linear Mixed Models with Varying Coefficients for Longitudinal Data", *Biometrics*, **60**, 8-15.

Appendix A

Calculation of the Gradient in the Semiparametric Generalized Linear Model with Log-concave Random Component

Any vector $\boldsymbol{\varphi} \in \mathbb{R}^m$ defines a function via

$$\varphi(y) = (1 - \frac{y - y_k}{\delta_k})\varphi_k + \frac{y - y_k}{\delta_k}\varphi_{k+1} \quad for\ y \in [y_k, y_{k+1}], 1 \leq k \leq n$$

where $\delta_k = y_{k+1} - y_k$, and $\varphi_k = \varphi(y_k)$. Here, the index set of k's consists of the points in which the slopes of $\varphi(\cdot)$ are changing. Then we may write (2.4) as $\mathrm{L}(\boldsymbol{\varphi}) = \sum_k p_k(\varphi_k + y_k\mathbf{x}_k^T\hat{\boldsymbol{\beta}}) - \sum_i p_i \log\{\sum_k \delta_k J(\varphi_k, \varphi_{k+1})\}, withto$
$\mathrm{J}(\varphi_k, \varphi_{k+1}) = \int_0^1 \exp\{(1-t)\varphi_k + t\varphi_{k+1} + (t\delta_k + y_k)\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}dt, J(r,s) = \exp(r)J(0, s-r)$.Let

$$\begin{aligned} J_{ab}(r,s) &= \frac{\partial^{a+b}}{\partial r^a \partial s^b} J(r,s), \\ J_{ab}(r,s) &= \int_0^1 (1-t)^a t^b \exp\{(1-t)r + ts + (t\delta_k + y_k)\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}dt, \\ J_{ab}(s,r) &= \exp(r)J_{ab}(0, s-r). \end{aligned}$$

One may write the auxiliary function $J(\varphi_k, \varphi_{k+1})$ explicitly

$$J(\varphi_k, \varphi_{k+1}) = \frac{\exp(y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})\{\exp(\varphi_{k+1} + \delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}) - \exp(\varphi_k)\}}{\varphi_{k+1} + \delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}} - \varphi_k}$$

or utilize the fact $J(r,s) = \exp(r)J(0, s-r)$.

Let $J_{ab}(r,s) = \partial^{a+b}/\partial r^a \partial s^b J(r,s)$, we have

$$J_{ab}(r,s) = \int_0^1 (1-t)^a t^b \exp\{(1-t)r + ts + (t\delta_k + y_k)\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}dt$$

and $J_{ab}(s,r) = \exp(r)J_{ab}(0,s-r)$. In the following, each J function is listed with its explicit form and Taylor expansion. Taylor expansion is used to make the computation stable.

$$\begin{aligned}
J(0,y) &= \frac{\exp(y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})\{\exp(y+\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})-1\}}{y+\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}} \\
&= \frac{\exp(y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}+\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})-\exp(y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}))}{y+\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}} \\
&\quad + \frac{(\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}-1)\exp(y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}+\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})+\exp(y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})}{(\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2}y \\
&\quad + \frac{\{(\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2-2\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}+2\}\exp(y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}+\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})-2\exp(y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})}{2(\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})^3}y^2+O(y^3).
\end{aligned}$$

Moreover, elementary calculations reveal that:

$$\begin{aligned}
J_{10}(0,y) &= \frac{\exp(y+\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}+y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})-(y+\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}+1)\exp(y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})}{(y+\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2} \\
&= \frac{\exp(y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}+\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})-(\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}+1)\exp(y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})}{(\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2} \\
&\quad + \frac{(\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}-2)\exp(y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}+\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})+(\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}+2)\exp(y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})}{(\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})^3}y \\
&\quad + \frac{\{(\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2-4\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}+6\}\exp(y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}+\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})-(2\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}+6)\exp(y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})}{2(\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})^4}y^2 \\
&\quad + O(y^3),
\end{aligned}$$

$$\begin{aligned}
J_{01}(0,y) &= \frac{(y+\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}-1)\exp(y+\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}+y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})+\exp(y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})}{(y+\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2} \\
&= \frac{(\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}+1)\exp(y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}+\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})+\exp(y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})}{(\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2} \\
&\quad + \frac{\{(\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2-2\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}+2\}\exp(y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}+\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})-2\exp(y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})}{(\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})^3}y \\
&\quad + \frac{\{(\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})^3-3(\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2+6\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}-6\}\exp(y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}+\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})+6\exp(y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})}{2(\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}})^4}y^2 \\
&\quad + O(y^3).
\end{aligned}$$

We have $\bigtriangledown L(\boldsymbol{\varphi}) = (\frac{\partial L(\boldsymbol{\varphi})}{\partial \varphi_1}, \cdots, \frac{\partial L(\boldsymbol{\varphi})}{\partial \varphi_n})$, and

$$\frac{\partial L(\boldsymbol{\varphi})}{\partial \varphi_j} = p_j - \sum_i p_i \{ \frac{\delta_j J_{10}(\varphi_j, \varphi_{j+1}) + \delta_{j-1} J_{01}(\varphi_{j-1}, \varphi_j)}{\sum_k \delta_k J_{00}(\varphi_k, \varphi_{k+1})} \}.$$

Appendix B

Quasi-code of Algorithm for Maximizing Likelihood (2.4)

The following quasi-code is used for maximizing likelihood 2.4 in SGLM-L based on the method of steepest ascent described in section 2.3.2.

iter1$\leftarrow 1$
while [max { abs ($\boldsymbol{\varphi}_0$-$\boldsymbol{\varphi}_{\text{new}}$)} > 0.001 and iter1<100]{
$\boldsymbol{\varphi}_0 \leftarrow \boldsymbol{\varphi}_{\text{new}}$, $\boldsymbol{\alpha}_0 \leftarrow \boldsymbol{\alpha}_{\text{new}}$,$L \leftarrow L_{\text{new}}$
$t \leftarrow$ choose $t(\boldsymbol{\alpha}, \bigtriangledown L_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0))$
$\boldsymbol{\alpha}_{\text{new}} \leftarrow \boldsymbol{\alpha}_0 + t \bigtriangledown (\boldsymbol{\alpha}_0)$
$\boldsymbol{\varphi}_{\text{new}} \leftarrow$ search$\varphi(B^{-1}\boldsymbol{\alpha}_{\text{new}})$
$\text{L}_{\text{new}} \leftarrow L_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_{\text{new}})$
if ($\text{L}_{new} > L$){reset the knots}, iter2 $\leftarrow 1$
while ($\text{L}_{\text{new}} < L$){
t=$(0.5)^{iter3}t$
$\boldsymbol{\alpha}_{\text{new}} \leftarrow \boldsymbol{\alpha}_0 + t \bigtriangledown L_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0)$
$\boldsymbol{\varphi}_{\text{new}} \leftarrow$ search $\varphi(B^{-1}\boldsymbol{\alpha}_{\text{new}})$
$\text{L}_{\text{new}} \leftarrow L_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_{\text{new}})$
iter2$\leftarrow$ iter2 $+ 1$
}
iter1$\leftarrow$ iter1 $+ 1$
}

Appendix C

Computation for Newton-Raphson Algorithm in the Semiparametric Generalized Linear Model With Log-concave Random Component

For computing $\partial L(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ and $\partial^2 L(\beta)/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T$, we need to find $\int yx_{ij}\exp\{\varphi(y)+y\mathbf{x}_i^T\boldsymbol{\beta}\}dy$ and $\int y^2x_{ij}x_{ik}\exp\{\varphi(y)+y\mathbf{x}_i^T\boldsymbol{\beta}\}dy$. It is sufficient to know these two integrations if we know $\int_0^1 t\exp\{(1-t)\varphi_k+t\varphi_{k+1}+(t\delta_k+y_k)\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}dt$ and $\int_0^1 t^2\exp\{(1-t)\varphi_k+t\varphi_{k+1}+(t\delta_k+y_k)\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}dt$, because $y=\delta_k t+y_k,$ for $y\in(y_k,y_{k+1})$.

$\int_0^1 t\exp\{(1-t)\varphi_k+t\varphi_{k+1}+(t\delta_k+y_k)\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}dt$ can be compute from $J_{01}(r,s)$. We also obtain the following equation

$$\begin{aligned}&\int_0^1 t^2\exp\{(1-t)\varphi_k+t\varphi_{k+1}+(t\delta_k+y_k)\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}dt\\&=\frac{\{(\varphi_{k+1}-\varphi_k)^2+2(\varphi_{k+1}-\varphi_k)(b+1)-b^2-2b+2\}e^{\varphi_{k+1}+a+b}-2e^{\varphi_k+a}}{\{(\varphi_{k+1}-\varphi_k)^2+2b(\varphi_{k+1}-\varphi_k)+b^2\}(\varphi_{k+1}-\varphi_k+b)}\end{aligned}$$

where $a=y_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}$, and $b=\delta_k\mathbf{x}_i^T\hat{\boldsymbol{\beta}}$.

Appendix D

Power Analysis for the Semiparametric Generalized Linear Model With Log-concave Random Component

For case 2 in the simulation study for the SGLM-L, the data sets are generated from the simple linear regression (SLR) $y_i = \alpha x_i + \varepsilon_i$, where, $i \in (1, 2, \ldots, 40)$, the x_i's are random points in the interval $[-2, 2]$, and $\varepsilon_i \sim N(0, 1)$.

The power study include 1000 Monte Carlo repetitions. In the first process, we generated two vectors, $\mathbf{x}$ and $\boldsymbol{\varepsilon}$. Using these two vectors, we construct 6 data sets based on SLR for six values of the slope, $\alpha \in (0, 0.1, 0.2, 0.3, 0.4, 0.5)$. Then we can compute the test statistic for these 6 data sets for both the SGLM-L and SLR. Thus, we have 12 test statistics.

The above process is repeated for 1000 times. The test statistics for this 1000 Monte Carlo repetitions could be summarized as the following Table D.1.

Iteration	SGLM-L						SLR					
	$\alpha = 0$	0.1	0.2	0.3	0.4	0.5	0	0.1	0.2	0.3	0.4	0.5
1	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1000	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12

Table D.1: Data structure for the test statistics in the power study for the SGLM-L.

Then the test statistics in Table D.1 are sorted for each column. For each of the techniques (SGLM-L and SLR), the 50 largest value ($\text{test}_{(50)}$) when $\alpha = 0$ is chosen to be the critical value. The empirical power will be the proportion of the test statistics which are less than or equal to the $\text{test}_{(50)}$ for different values of the slope, α.

Appendix E

Derivation of Expression (3.7)

Laplace's method for integrals as describes, for example, in De Bruijin (1961) provides an approximation for integrals of the form $\int e^{n\mathcal{L}(\boldsymbol{\nu})}d\boldsymbol{\nu}$ when n is large. The idea is that if $\mathcal{L}$ has a unique maximum at $\hat{\boldsymbol{\nu}}$, then for large n the value of this integral depends only on the behavior of the function $\mathcal{L}$ near its maximum. Thus if we set $\sigma^2 = -1/\mathcal{L}''(\hat{\boldsymbol{\nu}})$, then we can replace $\mathcal{L}(\boldsymbol{\nu})$ by $\mathcal{L}''(\hat{\boldsymbol{\nu}}) - (\boldsymbol{\nu} - \hat{\boldsymbol{\nu}})^2/(2\sigma^2)$. This produces the approximation

$$\int e^{n\mathcal{L}(\boldsymbol{\nu})}d\boldsymbol{\nu} = n\mathcal{L}(\hat{\boldsymbol{\nu}}) \int \exp[-\{\frac{n(\boldsymbol{\nu} - \hat{\boldsymbol{\nu}})^2}{2\sigma^2}\}]d\boldsymbol{\nu} = \sqrt{2\pi}\sigma n^{-1/2}\exp\{n\mathcal{L}(\hat{\boldsymbol{\nu}})\}$$

Writing Equation (3.4) in the form $c|\boldsymbol{\Sigma}|^{-1/2}\int e^{-\mathcal{L}(\mathbf{b})}d\mathbf{b}$, we apply the Laplace method for integral approximation (Tierney and Kadane 1986). Ignoring the multiplication constant c, the approximation yields

$$l_M\{y; \eta(\cdot), \boldsymbol{\beta}, \boldsymbol{\theta}\} \approx -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\mathcal{L}''(\tilde{\mathbf{b}}) - \mathcal{L}(\tilde{\mathbf{b}}), \tag{E.1}$$

where $\tilde{\mathbf{b}} = \tilde{\mathbf{b}}\{\eta(\cdot), \boldsymbol{\beta}, \boldsymbol{\theta}\}$ satisfies

$$-\sum_{i=1}^{n}[y_i\mathbf{z}_i - \psi'\{\eta(\boldsymbol{\beta}^T\mathbf{x}_i) + \mathbf{z}_i^T\boldsymbol{b}\}\mathbf{z}_i] + \boldsymbol{\Sigma}^{-1}\boldsymbol{b} = 0.$$

Adding approximation (E.1) into expression (3.6), penalized likelihood (3.6) can be approximated by

$$l_M\{y; \eta(\cdot), \boldsymbol{\beta}, \boldsymbol{\theta}\} \approx -\frac{1}{2}|I + \mathbf{Z}^T\mathbf{W}\mathbf{Z}\boldsymbol{\Sigma}| + \sum_{i=1}^{n} l_i\{y; \eta(\cdot), \boldsymbol{\beta}, \boldsymbol{\theta}, \tilde{\mathbf{b}}\} - \tilde{\mathbf{b}}^T\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{b}} \tag{E.2}$$

where $\mathbf{W} = \text{diag}[\psi'\{\eta(\boldsymbol{\beta}^T\mathbf{x}_i) + \mathbf{z}_i^T\tilde{\boldsymbol{b}}\}]$ that are recognizable as the GLM Fisher information (McCullagh and Nelder 1989).

Assuming that $\mathbf{W}$ varies slowly with $(\eta(\cdot), \boldsymbol{\beta})$ for fixed θ (Breslow and Clayton, 1993), we ignore the first term in expression (E.2) when maximizing it with respect to $\eta(\cdot)$. Thus the estimators $(\hat{\eta(\cdot)}, \hat{\boldsymbol{\beta}}, \hat{\mathbf{b}})$ can be equivalently obtained by maximizing the PQL in expression (3.7) jointly with respect to $(\eta(\cdot), \boldsymbol{\beta}, \mathbf{b})$.

Appendix F

One Step Updating of β for Maximizing Expression (3.9)

For maximizing expression (3.9), the one-step update of $\boldsymbol{\beta}$ is calculated using the following equation. For more details concerning constrained maximum likelihood estimation, see Aitchison and Silvey (1960).

$$\begin{pmatrix} \boldsymbol{\beta}_{new} \\ \gamma_{new} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_{old} \\ \gamma_{old} \end{pmatrix} + \begin{pmatrix} \mathbf{B} & -\mathbf{h} \\ -\mathbf{h}^T & 0 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{D}\text{log(likelihood)} + \mathbf{h}\gamma_{old} \\ h \end{pmatrix}$$

In this equation, $\mathbf{B} = E[-\partial^2 l/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T]$. As well, $h = (\| \boldsymbol{\beta} \|^2 - 1)$, $\mathbf{h} = \partial h/\partial\boldsymbol{\beta} = 2\boldsymbol{\beta}$, and $\mathbf{D}\text{log(likelihood)} = \partial l/\partial\boldsymbol{\beta}$.

Appendix G

Proof of Consistency in Theorem 1

Regularity Conditions:

Define $\boldsymbol{\omega} = (\boldsymbol{b}^T, \boldsymbol{\tau}^T)^T$.

1. The observations $\mathbf{v}_i = (\mathbf{x}_i, \mathbf{z}_i, y_i)$ are independent and identically distributed with probability density $f(\mathbf{v}, \tau|\boldsymbol{b})$ with respect to some measure μ. $f(\mathbf{v}, \tau|\boldsymbol{b})$ has a common support and the model is identifiable. Furthermore, the $L(\mathbf{v}, \boldsymbol{\tau}|\boldsymbol{b}) = \log f(\mathbf{v}, \boldsymbol{\tau}|\boldsymbol{b})$ (for convenience, we brief as $L(\mathbf{v}, \boldsymbol{\tau}) = L(\mathbf{v}, \boldsymbol{\tau}|\boldsymbol{b})$ satisfying the equations

$$E_{\boldsymbol{\tau}}\{\frac{\partial L(\mathbf{v}, \boldsymbol{\tau})}{\partial \boldsymbol{\tau}_j}\} = 0 \quad \text{for} \quad \text{all } j$$

and

$$I_{jk}(\boldsymbol{\tau}) = E_{\boldsymbol{\tau}}\{\frac{\partial L(\mathbf{v}, \boldsymbol{\tau})}{\partial \boldsymbol{\tau}_j}\frac{\partial L(\mathbf{v}, \boldsymbol{\tau})}{\partial \boldsymbol{\tau}_k}\} = E_{\boldsymbol{\tau}}\{-\frac{\partial^2 L(\mathbf{v}, \boldsymbol{\tau})}{\partial \boldsymbol{\tau}_j \partial \boldsymbol{\tau}_k}\}.$$

2. The Fisher information matrix

$$I(\boldsymbol{\tau}) = E[\{\frac{\partial}{\partial \boldsymbol{\tau}} L(\boldsymbol{\tau})\}\{\frac{\partial}{\partial \boldsymbol{\tau}} L(\boldsymbol{\tau})\}^T]$$

is finite and positive definite at $\boldsymbol{\omega} = \boldsymbol{\omega}_0$.

3. There exist an open subset ω of Θ containing the true parameter vector $\boldsymbol{\omega}_0$ such that for almost all $\mathbf{v}$ the density $f(\mathbf{v}, \boldsymbol{\tau})$ admits all third derivatives $\frac{\partial^3 L(\mathbf{v}, \boldsymbol{\tau})}{\partial \boldsymbol{\tau}_j \partial \boldsymbol{\tau}_k \partial \boldsymbol{\tau}_l}$ for all $\boldsymbol{\omega} \in \omega$. Further there exist function M_{jkl} such that

$$|\frac{\partial^3 L(\mathbf{v}, \boldsymbol{\tau})}{\partial \boldsymbol{\tau}_j \partial \boldsymbol{\tau}_k \partial \boldsymbol{\tau}_l}| \leq M_{jkl}(\mathbf{v}) \quad \text{for} \quad \text{all } \boldsymbol{\omega} \in \omega$$

where $m_{jkl} = E_{\boldsymbol{\omega}_0}\{M_{jkl}(\mathbf{v})\} < \infty$.

These regularity conditions guarantee asymptotic normality of the ordinary maximum likelihood estimates (Lehmann and Casella, 1998).

We can rewrite the DPQL (3.8) of model (3.4) as

$$Q_{n,\lambda}(\boldsymbol{\omega}) = L_{n,\lambda}(\boldsymbol{\tau}|\boldsymbol{b}) - \boldsymbol{\omega}^T\boldsymbol{\Sigma}\boldsymbol{\omega} - \frac{n}{2}\lambda\boldsymbol{\omega}^T\mathbf{D}\boldsymbol{\omega}$$

where $\boldsymbol{\Sigma} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\boldsymbol{b}}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}$ and $\mathbf{D} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{D}_{\boldsymbol{\tau}} \end{pmatrix}$.

Let $a_n = n^{-1/2} + \lambda_n$. We need to show that for any given $\varepsilon > 0$ that there exists a constant C such that $\mathrm{P}\{\sup_{\|\mathbf{u}\|=C} Q_{n,\lambda_n}(\boldsymbol{\omega}_0 + a_n\mathbf{u}) < Q_{n,\lambda_n}(\boldsymbol{\omega}_0)\} \geq 1 - \varepsilon$

That is, with probability at least $1-\varepsilon$, there exists a local maximum in the ball $\{\boldsymbol{\omega}_0 + a_n\mathbf{u} : \|\mathbf{u}\| \leq C\}$. Thus there exists local maximizer $\hat{\boldsymbol{\omega}}$ such that $\|\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}_0\| = O_p(a_n)$.

We have

$$\begin{aligned} Q_{n,\lambda_n}(\boldsymbol{\omega}_0 + a_n\mathbf{u}) - Q_{n,\lambda_n}(\boldsymbol{\omega}_0) &= L_{n,\lambda_n}(\boldsymbol{\omega}_0 + a_n\mathbf{u}) - L_{n,\lambda_n}(\boldsymbol{\omega}_0) - \frac{1}{2}(\boldsymbol{\omega}_0 + a_n\mathbf{u})^T\boldsymbol{\Sigma}(\boldsymbol{\omega}_0 + a_n\mathbf{u}) \\ &\quad -\frac{n}{2}\lambda_n(\boldsymbol{\omega}_0 + a_n\mathbf{u})^T\mathbf{D}(\boldsymbol{\omega}_0 + a_n\mathbf{u}) + \frac{1}{2}\boldsymbol{\omega}_0^T\boldsymbol{\Sigma}\boldsymbol{\omega}_0 + \frac{n}{2}\lambda_n\boldsymbol{\omega}_0^T\mathbf{D}\boldsymbol{\omega}_0 \\ &= L_{n,\lambda_n}(\boldsymbol{\omega}_0 + a_n\mathbf{u}) - L_{n,\lambda_n}(\boldsymbol{\omega}_0) - a_n\boldsymbol{\omega}_0^T\boldsymbol{\Sigma}\mathbf{u} \\ &\quad -n\lambda_n a_n\boldsymbol{\omega}_0^T\mathbf{D}\mathbf{u} - \frac{1}{2}a_n^2\mathbf{u}^T\boldsymbol{\Sigma}\mathbf{u} - \frac{n}{2}\lambda_n a_n^2\mathbf{u}^T\mathbf{D}\mathbf{u} \end{aligned}$$

Using the standard argument on the Taylor expansion of the (unpenalized) likelihood function, we have

$$\begin{aligned} Q_{n,\lambda_n}(\boldsymbol{\omega}_0 + a_n\mathbf{u}) - Q_{n,\lambda_n}(\boldsymbol{\omega}_0) &\leq a_n L'(\boldsymbol{\omega}_0)^T\mathbf{u} - \frac{1}{2}\mathbf{u}^T I(\boldsymbol{\omega}_0)\mathbf{u}\{a_n^2(1 + o_p(1))\} - a_n\boldsymbol{\omega}_0^T\boldsymbol{\Sigma}\mathbf{u} \\ &\quad -n\lambda_n a_n\boldsymbol{\omega}_0^T\mathbf{D}\mathbf{u} - \frac{1}{2}a_n^2\mathbf{u}^T\boldsymbol{\Sigma}\mathbf{u} - \frac{n}{2}\lambda_n a_n^2\mathbf{u}^T\mathbf{D}\mathbf{u} \end{aligned}$$

Thus

$$\begin{aligned} Q_{n,\lambda_n}(\boldsymbol{\omega}_0 + a_n\mathbf{u}) - Q_{n,\lambda_n}(\boldsymbol{\omega}_0) &\leq \{a_n L'(\boldsymbol{\omega}_0)^T - a_n\boldsymbol{\omega}_0^T\boldsymbol{\Sigma} - n\lambda_n a_n\boldsymbol{\omega}_0^T\mathbf{D}\}\mathbf{u} \\ &\quad -\frac{1}{2}\mathbf{u}^T I(\boldsymbol{\omega}_0)\mathbf{u}\{a_n^2(1 + o_p(1))\} \\ &\quad -\frac{1}{2}a_n^2\mathbf{u}\boldsymbol{\Sigma}\mathbf{u} - \frac{n}{2}\lambda_n a_n^2\mathbf{u}D\mathbf{u} \end{aligned} \tag{G.1}$$

where $L'(\boldsymbol{\omega}_0)$ is the likelihood gradient vector. Note $L'(\boldsymbol{\omega}_0) = O_p(1)$. The second term will dominate the first term of (G.2) uniformly in $\|\mathbf{u}\| = C$, by choosing sufficiently large C. The third and fourth terms of (G.2) are also dominated by the second term. Hence, by choosing sufficiently large C, (G.1) holds.

Appendix H

Proof of Theorem 2

Asymptotic normality is established under assumption 1-3 and the consistency proved previously. For consistent estimator $\hat{\boldsymbol{\omega}} = \hat{\boldsymbol{\omega}}_{n,\lambda_n}$, using a first order Taylor expansion of Q_{n,λ_n} near $\boldsymbol{\omega}_0$ yields,

$$0 = \frac{\partial Q_{n,\lambda_n}}{\partial \boldsymbol{\omega}}|_{\hat{\boldsymbol{\omega}}} = \frac{\partial Q_{n,\lambda_n}}{\partial \boldsymbol{\omega}}|_{\boldsymbol{\omega}_0} + \frac{\partial^2 Q_{n,\lambda_n}}{\partial \boldsymbol{\omega}\partial \boldsymbol{\omega}^T}|_{\bar{\boldsymbol{\omega}}}(\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}_0)$$

where $\bar{\boldsymbol{\omega}}$ is a vector between $\hat{\boldsymbol{\omega}}$ and $\boldsymbol{\omega}_0$. Consequently, we have

$$\sqrt{n}(\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}_0) = \{-\frac{\partial^2 Q_{n,\lambda_n}}{n\partial \boldsymbol{\omega}\partial \boldsymbol{\omega}^T}|_{\bar{\boldsymbol{\omega}}}\}^{-1}\frac{\partial Q_{n,\lambda_n}}{\sqrt{n}\partial \boldsymbol{\omega}}|_{\boldsymbol{\omega}_0}. \tag{H.1}$$

Next we show that, (i) the limit distribution of

$$\frac{\partial Q_{n,\lambda_n}}{\sqrt{n}\partial \boldsymbol{\omega}}|_{\boldsymbol{\omega}_0} \xrightarrow{D} N(0, I(\boldsymbol{\tau}_0)). \tag{H.2}$$

Note that

$$\frac{\partial Q_{n,\lambda_n}}{\sqrt{n}\partial \boldsymbol{\omega}}|_{\boldsymbol{\omega}_0} = \frac{\partial L_n}{\sqrt{n}\partial \boldsymbol{\tau}}|_{\boldsymbol{\tau}_0} - \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}\boldsymbol{b}_0 - \lambda_n\sqrt{n}\mathbf{D}\boldsymbol{\tau}_0.$$

By the central limit theorem, the first term in distribution to $N(0, I(\boldsymbol{\tau}_0))$; the second term of the right hand side goes to zero as $n \to \infty$; the third term goes to zero since $\lambda_n = o(n^{-1/2})$. Thus, (H.2) is established.

Then, we need show (ii)

$$-\frac{\partial^2 Q_{n,\lambda_n}}{n\partial \boldsymbol{\omega}\partial \boldsymbol{\omega}^T}|_{\bar{\boldsymbol{\omega}}} \xrightarrow{P} I(\boldsymbol{\tau}_0).$$

We have

$$-\frac{\partial^2 Q_{n,\lambda_n}}{n\partial \boldsymbol{\omega}\partial \boldsymbol{\omega}^T}|_{\bar{\boldsymbol{\omega}}} = -\frac{\partial^2 L_n}{n\partial \boldsymbol{\omega}\partial \boldsymbol{\omega}^T}|_{\bar{\boldsymbol{\tau}}} + \frac{1}{n}\boldsymbol{\Sigma} + \lambda_n\mathbf{D},$$

the first term converges in probability to $I(\boldsymbol{\tau}_0)$, the second term goes to zero as $n \to \infty$; the third term goes zero since $\lambda_n = o(n^{-1/2})$.

Thus, applying Slutsky's lemma to expression (H.1) and converting back to the original parameter space via the delta method, Theorem 2 follows when $\lambda_n = o(n^{-1/2})$.