2010

Spline-based sieve semiparametric generalized estimating equation for panel count data

Lei Hua
University of Iowa

Recommended Citation

Hua, Lei. "Spline-based sieve semiparametric generalized estimating equation for panel count data." PhD (Doctor of Philosophy) thesis, University of Iowa, 2010.
http://ir.uiowa.edu/etd/517.

SPLINE-BASED SIEVE SEMIPARAMETRIC GENERALIZED ESTIMATING EQUATION FOR PANEL COUNT DATA

by

Lei Hua

An Abstract

Of a thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Biostatistics
in the Graduate College of
The University of Iowa

May 2010

Thesis Supervisor: Associate Professor Ying Zhang

ABSTRACT

In this thesis, we propose to analyze panel count data using a spline-based sieve generalized estimating equation method with a semiparametric proportional mean model $E(\mathbb{N}(t)|Z) = \Lambda_0\,(t)\,e^{\beta_0^T Z}$. The natural log of the baseline mean function, $log\Lambda_0\,(t)$, is approximated by a monotone cubic B-spline function. The estimates of regression parameters and spline coefficients are the roots of the spline based sieve generalized estimating equations (sieve GEE). The proposed method avoids assuming any parametric structure of the baseline mean function and the underlying counting process. Selection of an appropriate covariance matrix that represents the true correlation between the cumulative counts improves estimating efficiency.

In addition to the parameters existing in the proportional mean function, the estimation that accounts for the over-dispersion and autocorrelation involves an extra nuisance parameter σ^2, which could be estimated using a method of moment proposed by Zeger (1988). The parameters in the mean function are then estimated by solving the pseudo generalized estimating equation with σ^2 replaced by its estimate, $\hat{\sigma}_n^2$. We show that the estimate of (β_0, Λ_0) based on this two-stage approach is still consistent and could converge at the optimal convergence rate in the nonparametric/semiparametric regression setting. The asymptotic normality of the estimate of β_0 is also established. We further propose a spline-based projection variance estimating method and show its consistency.

Simulation studies are conducted to investigate finite sample performance of the sieve semiparametric GEE estimates, as well as different variance estimating methods with different sample sizes. The covariance matrix that accounts for the overdispersion generally increases estimating efficiency when overdispersion is present in the data. Finally, the proposed method with different covariance matrices is applied to a real data from a bladder tumor clinical trial.

Abstract Approved: ______________________________
Thesis Supervisor

____
Title and Department

____
Date

SPLINE-BASED SIEVE SEMIPARAMETRIC GENERALIZED ESTIMATING EQUATION FOR PANEL COUNT DATA

by

Lei Hua

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Biostatistics
in the Graduate College of
The University of Iowa

May 2010

Thesis Supervisor: Associate Professor Ying Zhang

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Lei Hua

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Biostatistics at the May 2010 graduation.

Thesis Committee: ______________________________
Ying Zhang, Thesis Supervisor

____
Jiang Huang

____
Michael Jones

____
Kung-Sik Chan

____
Kathryn Chaloner

To My Husband and My Parents

ACKNOWLEDGEMENTS

I cannot believe I am at this stage of my dissertation. It is my professors with their guidance and my family with their support who make this bumpy journey rewarding, enjoyable, valuable and memorable.

First, I owe my deepest gratitude to my advisor, Dr. Ying Zhang for his guidance and encouragement during my research and study at the University of Iowa. His perpetual energy and enthusiasm in research had always motivated me. He was always accessible and willing to help me in every aspect of my professional development. Without him, this thesis wouldn't be possible and I wouldn't be the statistician that I am today.

It is also my pleasure to thank the members of my committee, Drs. Huang, Jones, Chaloner, and Chan for their valuable suggestions and comments. I appreciate their time and expertise in my research. Thanks also to the Department of Biostatistics for the financial support and all the faculty members in the department who have taught me and encouraged my in any aspect of my study here.

Finally my husband, Zhenzhou Lei stands behind me and accompanies me through my sorrow and happiness all along the way. I dedicate this dissertation to him and our daughter Rachel with all my love.

ABSTRACT

In this thesis, we propose to analyze panel count data using a spline-based sieve generalized estimating equation method with a semiparametric proportional mean model $E(\mathbb{N}(t)|Z) = \Lambda_0\,(t)\, e^{\beta_0^T Z}$. The natural log of the baseline mean function, $log\Lambda_0\,(t)$, is approximated by a monotone cubic B-spline function. The estimates of regression parameters and spline coefficients are the roots of the spline based sieve generalized estimating equations (sieve GEE). The proposed method avoids assuming any parametric structure of the baseline mean function and the underlying counting process. Selection of an appropriate covariance matrix that represents the true correlation between the cumulative counts improves estimating efficiency.

In addition to the parameters existing in the proportional mean function, the estimation that accounts for the over-dispersion and autocorrelation involves an extra nuisance parameter σ^2, which could be estimated using a method of moment proposed by Zeger (1988). The parameters in the mean function are then estimated by solving the pseudo generalized estimating equation with σ^2 replaced by its estimate, $\hat{\sigma}_n^2$. We show that the estimate of (β_0, Λ_0) based on this two-stage approach is still consistent and could converge at the optimal convergence rate in the nonparametric/semiparametric regression setting. The asymptotic normality of the estimate of β_0 is also established. We further propose a spline-based projection variance estimating method and show its consistency.

Simulation studies are conducted to investigate finite sample performance of the sieve semiparametric GEE estimates, as well as different variance estimating methods with different sample sizes. The covariance matrix that accounts for the overdispersion generally increases estimating efficiency when overdispersion is present in the data. Finally, the proposed method with different covariance matrices is applied to a real data from a bladder tumor clinical trial.

TABLE OF CONTENTS

CHAPTER

APPENDIX

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1
INTRODUCTION

1.1 Motivating Examples

Panel count data are often seen in clinical trials, reliability studies, and epidemiological studies. A well known example is the superficial bladder tumor clinical trail studied by Byar et al. (1980), Wei et al. (1989), Wellner & Zhang (2000), Sun & Wei (2000), Zhang (2002), Wellner & Zhang (2007) and Lu et al. (2009) among others. Patients with superficial bladder tumors were enrolled and randomized into one of three treatment groups: pyridocine pills, thiotepa instillation or placebo group. The number and size of the bladder tumors were measured for each subject at their enrollment. Superficial bladder tumor has a high recurrent rate. During the follow-up visits, the newly formed bladder tumors for each subject were counted and removed and the assigned treatment was continued. The primary research interest was to evaluate and compare the effectiveness of the three different treatments and their abilities on suppressing the recurrence of the bladder tumor while controlling for other covariates. The number of visits and the time between visits varied from subject to subject.

Another interesting example is the data coming from the National Cooperative Gallstone Study (NCGS), which is a 10-year, multicenter, double-blinded, placebo-controlled clinical trial on the use of natural bile acid chenodeoxycholic acid (chenodiol) for the dissolution of cholesterol gallstones (Thall & Lachin (1988)). Patients were randomly assigned to one of three treatment groups: high dose, low dose or placebo. Although they were scheduled to follow-up at 1, 2, 3, 6, 9, and 12 months, the number of observations and each observation time differed from patient to patient. The actual successive visit times and the associated counts of nausea were recorded. The objective of this study was to estimate chenodiol's effect on the incidence of nausea.

Other interesting examples of panel count data include the number of seizures in epileptics, number of damaged joints in patients with psoriatic arthritis (Gladman et al. 1995), etc.

1.2 Literature Review

Panel count data share some special features with longitudinal data, survival data and categorical data; as a mixture of these three, they also impose more challenges for the analysis. First of all panel count data are a special case of longitudinal data, where subjects experience some events of interest multiple times. The resulting data are usually referred to as event history data. The event history data can be further classified into two types. One monitors the process continuously, records the exact event time and thus produces the recurrent event data (Byar et al. (1980); Prentice et al. (1981); Pepe & Cai (1993)). Panel count data is the other type, in which subjects are only observed at discrete observation time points. Instead of the exact event time, only the numbers of events between observation times are known. For the analysis of recurrent event data, a number of methods have been proposed. Prentice et al. (1981) generalized the Cox proportional hazard function (Cox 1972) and developed a conditional likelihood model for subjects with multiple events. Andersen & Gill (1982) used an intensity-based counting process modeling techniques and derived the asymptotic properties of the estimators based on the Martingale theory. Lawless (1987) analyzed the data based on a nonhomogeneous Poisson process plus some random effects. In the framework of marginal model, Wei et al. (1989) analyzed the multivariate failure time model. Lawless & Nadeau (1995) discussed the special case when the failure times were discrete. Other examples include Pepe & Cai (1993), Lin et al. (2000) and Lin & Ying (2001).

In spite of the abundant discussions of the recurrent event history data, the analysis of panel count data has started to attract attention in the past two decades. Several parametric approaches have been discussed. Kalbfleisch & Lawless (1985) and Gentleman et al. (1994) discussed the analysis of panel count data based on a finite continuous Markov model. Breslow (1984) discussed the parametric analysis using Poison regression. However, in medical settings, the disease progression is often unknown. Parametric assumptions relating the outcomes and observation times are susceptible to serious violations. In addition, the observation times vary from patient to patient in panel count data. Even in clinical trial with scheduled follow-up times like in NCGS, patients can still be early, late or absent. Neglecting the actual different visit times and using the scheduled time may introduce bias and such analysis is questionable.

Nonparametric and semiparametric analysis can relax parametric assumptions and is applied to longitudinal data. Zhang et al. (1998), Lin & Zhang (1999) and

Rice & Wu (2001) among others adopted the linear mixed effect to analyze ordinary longitudinal data nonparametrically. Random effects are included in the model to account for part of the correlation between repeated measurement. A stochastic process, such as nonhomogeneous Ornstein-Uhlenbeck (NOU) process, Weiner process, integrated Weiner process, an integrated OU (IOU) or an ante-dependence process for equally spaced time points, is specified in the regression model to account for the autocorrelation. However, the choice of the stochastic process is arbitrary, and it is unknown how these assumptions will influence the inference of the mean function. In addition, the linear mixed effect model cannot deal with random observation times and it dose not address the monotone constraint imposed by the counting process in panel count data.

To deal with the special features of panel count data, Thall & Lachin (1988) studied the data from NCGS using a marginal model. They proposed a nonparametric estimation of the rate of the counting process. Sun & Kalbfleisch (1995) first discussed the estimation of mean function of a specific counting process directly. They applied the isotonic regression method and estimated the nonparametric mean function of the counting process at specific time points. Wellner & Zhang (2000) proposed a nonparametric maximum pseudolikelihood estimator(NPMPLE) and a nonparametric maximum likelihood estimator(NPMLE) assuming the underlying counting process as a nonhomogeneous Poisson process. They found out that NPMPLE is exactly the nonparametric estimator proposed by Sun & Kalbfleisch (1995). They proved the consistency of NPMPLE and NPMLE and derived their convergence rate. They showed the robustness of both estimators against the Poisson assumption. The NPMPLE is based on the pseudolikelihood, which neglects the correlation between consecutive counts and treats them as if they were independent. The NPMLE is based on the likelihood, which incorporates the correlation between consecutive counts based on Poisson assumption. In general, the NPMLE is more efficient than the NPMPLE at the cost of more computing times.

In many clinical trials, comparison between treatment groups is of primary interest. Based on their nonparametric estimation, Thall & Lachin (1988) proposed a K-variate statistic to compare the intensities of two treatment groups. Sun & Fang (2003) proposed a nonparametric test to compare the estimates from different counting processes and proved the asymptotic normality of the test statistics. Later, Zhang (2006) discussed a similar test comparing the mean functions of K populations based on the asymptotic normality of a smoothing functional of the NPMPLE studied in Wellner & Zhang (2000).

Recently, there are more interests in analyzing panel count data using a semiparametric model. In the literature of repeated measurement and longitudinal data, a semiparametric model is often assumed,

$$E\left(Y|Z,T\right)=\mu\left(Z^{T}\beta+\gamma\left(T\right)\right) \tag{1.1}$$

where μ is a known link function, β is the regression parameter and γ is the unknown function. In the panel count setting, when μ is chosen as the exponential function, and γ is chosen as the logarithm of the baseline mean function, e.g., $\gamma = log\Lambda_0\left(t\right)$, this model reduces to a proportional mean model

$$E(\mathbb{N}(t)|Z) = \Lambda_0\,(t)\,e^{\beta_0^T Z}. \tag{1.2}$$

This model is widely studied in the literature, for example, in Lawless & Nadeau (1995), Sun & Wei (2000), and Lin et al. (2000). The baseline mean function Λ_0 is monotone nondecreasing due to the nature of the counting process. The estimation of β_0 and Λ_0 often involve complicated algorithms with heavy computing effort.

Other work has been focused on the intensity, namely,

$$\lambda\,(t|Z) = \lambda_0\,(t)\,e^{\beta_0^T Z} \tag{1.3}$$

where $\lambda_0\,(t) = \frac{d}{dt}\Lambda_0\,(t)$. For example, Kalbfleisch & Lawless (1985) generalized their Markov model to handle the covariance analysis and used the proportional structure to model the transition intensities. Lee & Kim (1998) used the same Markov model for two or more correlated multi-state processes and modeled the correlation between these processes based on marginal models using the proportional model. When the outcome is a zero-one binary variable and λ is the intensity of the counting process, this model simplifies to the proportional hazard model of Cox (1972).

Although using the model in Equation (1.3) does not require nonnegativity and monotone nondecreasing constraint, in many cases, the mean function is of more interest and modeling it directly is desirable. Wellner & Zhang (2007) modeled the baseline mean function in Equation (1.2) directly and estimated the parameters using isotonic regression. Wellner & Zhang (2007) discussed the semiparametric maximum pseudolikelihood estimator and semiparametric maximum likelihood estimator based on a Poisson assumption in parallel to their two nonparametric estimators (Wellner & Zhang 2000). Both the maximum pseudolikelihood estimator and the maximum likelihood estimator are shown to be consistent regardless of the true underlying counting process. They studied the convergence rate of both estimators and showed that in spite of the fact that the nonparametric estimator of the baseline mean function converges at a slower rate, $n^{1/3}$, the regression parameter for the parametric part still converges at the standard rate, $n^{1/2}$ and is asymptotically normally distributed. The maximum likelihood estimator based on nonhomogeneous Poisson process assumption accommodates some correlation between consecutive cumulative counts, and in general is more efficient than the maximum pseudolikelihood estimator. However, neither method considers the overdispersion problem commonly seen in count data, and thus will not be very efficient when overdispersion is present. In this dissertation, we avoid assuming any underlying counting process and use a generalized estimating equation to estimate the parameters specified in the proportional mean function in Equation (1.2). Selection of an appropriate covariance matrix that accounts for the overdispersion will produce more efficient estimates when overdispersion is present in the data.

1.3 Outline of the Dissertation

The rest of the dissertation is organized as follows. Chapter 2 introduces the spline-based sieve semiparametric generalized estimating equation. Section 2.1 presents two commonly used smoothing techniques: kernel machine and splines in the analysis of longitudinal data. We use the regression splines to estimate the baseline mean function. Section 2.2 reviews the spline-based sieve M-estimators. They are sieve counterparts of the maximum pseudolikelihood estimator and the maximum likelihood estimator based on nonhomogeneous Poisson process studied by Wellner & Zhang (2007). Instead of maximizing some 'likelihood' function based on the assumption of the entire process, we propose to estimate the unknown parameters by only assuming the mean function of the counting process as shown in Equation (1.2), and assuming a working correlation matrix between the consecutive cumulative counts. The parameters are estimated by solving spline-based sieve semiparametric generalized estimating equations (sieve GEE). Section 2.3 presents the model in detail and discusses different choices of the covariance matrices that can be used in the estimating equation to accommodate different data structure. Chapter 3 discusses the asymptotic properties of the spline-based sieve GEE estimator proposed in Chapter 2 using modern empirical process theory. Some basic terms and theorems in empirical process theory are summarized in Section 3.1. General theorems of the consistency and convergence rate of the estimates of both the baseline mean function and the regression parameters as well as the asymptotic normality of the estimated regression parameters in the presence of a nuisance parameter are then developed in Section 3.2. In Section 3.3, these general theorems are further applied to the special structure of the Gamma-Frailty Poisson model we discussed in Section 2. Three standard error estimating methods for the spline-based sieve GEE estimator of the regression parameter are discussed in Chapter 4. Section 4.1 presents an estimating procedure based on the projection of the infinite-dimensional parameter onto the tangent space of the finite parameter spaces. Spline-based sieve method is applied again to approximate a so-called 'least favorable direction' used in the estimation. Instead of using the projection algorithm proposed in Section 4.1, we could heuristically treat spline coefficients as finite dimensional parameters and use the ordinary sandwich estimator of the standard error proposed by Zeger & Liang (1986) in parametric GEE model. Bootstrap method is also explored in Section 4.3. Chapter 5 discusses the algorithms used in computing the spline-based sieve GEE estimates. Solution of the spline-based sieve semiparametric generalized estimating equation subject to the monotone constraint can be solved using a combination of Newton-Raphson iteration and different projection algorithms. Section 5.1.1 discusses a Generalized Rosen (GR) algorithm utilized by Lu et al. (2007) and Lu et al. (2009). It is also implemented in our sieve GEE method. Isotonic regression is another commonly used algorithm in the optimization problems subjecting to the monotone constraint. We propose to combine Newton-Raphson algorithm and the isotonic regression (NR/IR) to compute the spline-based sieve semiparametric GEE estimates in Section 5.1.2. Section 5.2 presents different estimation methods for the overdispersion parameter in the covariance matrix, but not in the mean function. An extensive simulation study is done to compare the performance of the spline-based sieve semiparametric GEE estimator using different covariance matrices. Chapter 6 summarizes the simulation results. The proposed spline-based sieve semiparametric GEE method is applied to the data from a superficial bladder tumor clinical trial. Finally, we give some final remarks of the proposed method and discuss possible future works in Chapter 7.

CHAPTER 2
SPLINE-BASED SIEVE SEMIPARAMETRIC GENERALIZED ESTIMATING EQUATION

In this Chapter, a spline-based sieve semiparametric generalized estimating equation method is proposed to analyze panel count data. As mentioned in section 1.2, the proportional mean model in Equation (1.2) is assumed in the analysis. The baseline mean function is left unspecified. It can be estimated using step functions with jumps at distinct observation times as shown in Wellner & Zhang (2007). However, the dimension of the estimation of Wellner & Zhang's method increases rapidly as sample size increases and hence their method is computationally intensive. In most of applications, the true baseline mean function can be assumed as a smooth function, therefore it is more desirable to have a smooth estimator of the baseline mean function. Section 2.1 presents two common smoothing techniques used in the estimation of infinite-dimensional parameters in statistical literature. See examples in Huang (1996) and Wellner & Zhang (2007), etc. With the proportional mean assumption, regression splines render a simple approximation of the baseline mean function of the panel count data and facilitate an easy-to-implement estimating procedure. Section 2.2 reviews the spline-based sieve semiparametric maximum pseudolikelihood estimator and the spline-based sieve semiparametric maximum likelihood estimator for panel count data studied by Lu et al. (2009). These two estimators are different versions of the semiparametric maximum pseudolikelihood estimator and the semiparametric maximum likelihood estimator studied by Wellner & Zhang (2007). The spline-based sieve estimators have a faster convergence rate than their counterparts. Section 2.3 presents a new method for the semiparametric inference using generalized estimating equation approach. Different working covariance matrices are suggested in the method for different data structure. The newly proposed method could produce more efficient estimates for both the regression parameters and the baseline mean function than the spline-based sieve semiparametric maximum likelihood estimators studied by Lu et al. (2009) if the overdispersion problem is present in the panel count data.

2.1 Smoothing and its application in the analysis of clustered data

Kernel regression and spline smoothing are the two techniques widely used to estimate unknown functions in the nonparametric/semiparametric estimation literature. Both methods have been applied to the analysis of longitudinal data in statistical literature. The kernel smoothing is the simplest smoothing method. It is based on the weighted local average of available data points, e.g.

$$\hat{f}\left(x\right) = \frac{\sum_{i=1}^{n} K_h\left(x_i\right) y_i}{\sum_{i=1}^{n} K_h\left(x_i\right)}$$

where the weights are explicitly determined by the kernel function $K_h\left(x\right) = \frac{1}{h} K_h\left(\frac{x-x_i}{h}\right)$, and the neighborhood is determined by h, so called bandwidth. Selections of the kernel function and the bandwidth are the two main considerations in the kernel smoothing.

The kernel function K can be any unimodal and symmetric function. In the weighted average approximation, the center of the kernel is placed at each data point.

Commonly used kernel functions include Uniform, Gaussian and Epanechnikov functions. The performance of kernel smoothing is often measured by mean integrated square error (MISE) or asymptotic mean integrated square error (AMISE). Epanechnikov kernel often minimizes AMISE and is therefore optimal.

The choice of the shape of the kernel function is less important than the bandwidth, h. The bandwidth controls the level of smoothing. A wider bandwidth tends to over-smooth the estimation in the sense that it is too biased and may not reveal structural features of the data. A narrower bandwidth may result in a wiggly looking estimate. Different methods have been proposed to select the bandwidth, such as *Rule of thumb* discussed in Silverman (1986), *maximal smoothing principle* proposed by Terrell (1990), *least square cross-validation* by Rudemo (1982) and Breslow (1984) and other variants of cross-validation methods.

Many articles applied the kernel smoothing method to the analysis of clustered data; see Severini & Staniswalis (1994), Wild & Yee (1996), Zeger & Diggle (1994) and Lin & Carroll (2000) and Lin & Carroll (2001). Lin & Carroll (2000) and Lin & Carroll (2001) applied the traditional kernel smoothing in the clustered nonparametric and semiparametric regression respectively. In the nonparametric setting, the traditional kernel-based nonparametric estimations are efficient only when ignoring the correlations within a cluster. In the semiparametric setting, even when a working independent covariance matrix is used, the estimate of the parametric regression parameter is still not efficient. These results are rather different from the results of the parametric analysis. Wang (2003) proposed a different kernel function, the seemingly unrelated kernel (SUR), in the clustered nonparametric regression. Correctly specifying the correlation can improve the estimation efficiency.

Polynomial function, $f\left(x\right)=\sum_{i=0}^{K}a_ix^i$, is another method that has long been used to approximate some unknown function due to its linearity of the regression parameters and easy calculation with respect to derivatives and integrations. A main drawback of polynomial approximation is its 'non-localness', namely a slight change of one data point may cause large changes in the regression parameters and polynomial approximations. Substantial improvement can be gained by using piecewise polynomial functions, *splines*. The regions that define the pieces are separated by a series of breakpoints called *knots*. The function within each pair of adjacent knots is approximated by a polynomial function with the same order K. In order to enforce the smoothness of the estimate, the derivatives of the adjacent polynomials at any knot are the same up to the order of $K-1$. For a given set of knots, such constructed piecewise polynomial approximation can be expressed as a linear span of an appropriate set of basis functions. For example, the region of approximation $[L,U]$ can be divided into m_n+1 subintervals by a series of interior knots, $\Xi=\{\xi_i, i=1,2,\cdots,m_n\}$ such that

$$L=\xi_0<\xi_1<\cdots<\xi_{m_n}<\xi_{m_n+1}=U$$

Given these knots, any function $f\left(x\right)$ within this region can be approximated by

$$\tilde{f}\left(x\right)=\sum_{l=1}^{q_n}\alpha_l B_l\left(x\right) \tag{2.1}$$

where $B_l\left(x\right), l=1,2,\cdots,q_n$ are spline basis functions and are themselves a series of piecewise polynomials that are smoothly connected at the knots; q_n is the sum of the number of the interior knots m_n and the order of the basis functions B_l. In order to use this approximation, we need to determine the number and location of the interior knots as well as the basis functions and their order used in the linear span.

The smoothing spline method chooses the order of the spline by minimizing a modified function

$$\sum_{i=1}^{n} \left(Y_i - \tilde{f}\left(x_i\right)\right)^2 + \lambda \int \left(\tilde{f}''\left(x\right)\right)^2 dx$$

where λ is a penalty term that controls the smoothness of the approximated $\tilde{f}\left(x\right)$. A larger λ corresponds to a less smooth spline estimation. Wang et al. (2005) used the smoothing spline in analyzing clustered data. They proved the asymptotic equivalence between the smoothing spline and the SUR proposed by Wang (2003). In both methods, using the true covariance matrix as a working covariance matrix increases the estimating efficiency, and these two estimators outperform the traditional kernel estimators studied by Lin & Carroll (2000) and Lin & Carroll (2001).

Zhu et al. (2008) studied longitudinal data using regression splines. Given a set of interior knots, they studied an estimator based on weighted least square regression splines by minimizing

$$\sum_{i=1}^{n} \left(Y_i - \tilde{f}\left(x_i\right)\right)^T V^{-1} \left(Y_i - \tilde{f}\left(x_i\right)\right).$$

The bias of the estimator does not depend on working correlation matrix, and the mean square error is minimized when the true correlation structure is used. However, this method only deals with situations where subjects are followed at same observation times. It cannot be readily applied to the scenario of panel count data where the number of observations and each observation time vary from subject to subject.

In this dissertation, the regression spline method is used to estimate the baseline mean function of the counting process nonparametrically. As shown in Equation (2.1), regression spline approximates the function in a sieve space made by a linear span of some basis functions $B_l(t), l = 1, 2, \cdots, q_n$. The dimension of the sieve space, q_n, increases as sample size increases. But it could increase, depending on the choice of sieves, much slower than the sample size increases. Asymptotically, the closure of the limiting approximation space contains the true infinite dimensional parameter space. The definition of the spline function and the formulation of the regression splines in panel count data are stated in details in Section 2.2. Lu et al. (2009) applied a similar sieve approximation to the maximum pseudolikelihood estimator and the maximum likelihood estimator of the panel count data by Wellner & Zhang (2007). Instead of using the likelihood of the counting process as in Lu et al. (2009) or a weighted least square estimate as in Zhu et al. (2008), we discuss a generalized estimating equation in Section 2.3. Different working covariance matrices are discussed for improving the efficiency of the estimation and they can be subject-specific.

2.2 Spline-based Sieve Maximum Likelihood Estimation

Suppose, $\mathbb{N} = \{\mathbb{N}(t) : t \geq 0\}$ is a univariate counting process. There are random number K observations of this counting process at $0 \equiv T_0 < T_{K,1} < \cdots < T_{K,K}$. We denote $\underline{T}_K \equiv (T_{K,1}, T_{K,2}, \cdots, T_{K,K})$, and $\mathbb{N} \equiv (\mathbb{N}(T_{K,1}), \mathbb{N}(T_{K,2}), \cdots, \mathbb{N}(T_{K,K}))$, the cumulative event count at these discrete observation times. We assume the number of observations and the observation times, $(K, \underline{T}_K)$, are independent of the

point process $\mathbb{N}$, conditioning on the covariate vector Z. Panel count data are composed of a random sample of $X_1, X_2, \cdots, X_n$, where the observation X_i consists of $\left(K_i, \underline{T}_{K_i}, \mathbb{N}^{(i)}, Z_i\right)$ with $\underline{T}_{K_i} = \left(T^{(i)}_{K_i,1}, T^{(i)}_{K_i,2}, \cdots, T^{(i)}_{K_i,K_i}\right)$ and
$\mathbb{N}^{(i)} = \left(\mathbb{N}^{(i)}\left(T^{(i)}_{K_i,1}\right), \mathbb{N}^{(i)}\left(T^{(i)}_{K_i,2}\right), \cdots, \mathbb{N}^{(i)}\left(T^{(i)}_{K_i,K_i}\right)\right)$.

Assume observation times are restricted in a finite interval $[L, U]$ and the true function $log\Lambda(t)$ is continuous and bounded in this interval. Let a sequence of knots $t = \{L = t_1 = t_2 = \cdots = t_l < t_{l+1} < \cdots < t_{l+m_n} < t_{l+m_n+1} = \cdots = t_{m_n+2l} = U\}$ partition the closed interval $[L, U]$ into $m_n + 1$ subintervals, where $m_n \approx n^{\nu}$ is a positive integer such that $\max_{1 \le k \le m_n} |t_{l+k} - t_{l+k-1}| = O\left(n^{-\nu}\right)$. Denote $\phi_{l,t}$ as a class of polynomial spline functions of order l, $l \ge 1$. $\phi_{l,t}$ is spanned by a series of polynomial spline basis functions $\{B_i, 1 \le i \le q_n\}$ where $q_n = m_n + l$.

The dimension of the sieve space, q_n is determined by sample size and is related to the asymptotic properties of the estimates. The discussion of these asymptotic properties is delegated to Chapter 3. The choice of the knots is suggested by the data. Reducing the number of knots reduces the flexibility of the fitted spline, and increasing the density of knots in different regions of the observation time allows increased flexibility within those regions. Uniform partitions and partitions according to the quantiles of the data are two commonly used convenient choices. In our simulation setup in Chapter 6, the observations scheduled at a later time have a higher probability of missing. Knots allocated with the uniform partition scheme will end up with fewer observations in the intervals in the later time and hence introduce a greater bias to the estimation of the baseline mean function especially when sample size is small. In this case, knots allocated with the quantile-based scheme are preferred.

Different basis functions have been used in the literature, such as truncated first order splines used in Zhang (1997) and piecewise second order splines used in Huang (1996). And it is noteworthy that if the knots $t_i, i = 1, 2, \cdots m_n$ are chosen to be all distinct event times and the order of spline is one, we end up using step functions to approximate the nonparametric function. In this thesis, we consider to use cubic B-spline functions to approximate the logarithm of the baseline mean function, $\log\Lambda_0\left(t\right)$. B-spline is easily interpretable. It is local so the coefficient can be related to the behavior of the estimate at specific locations. And it is widely used in the software packages. Transformations of B-splines to other bases are easy to implement. Cubic spline is chosen since it is flexible and twice differentiable at the knots without being overly complex. When using B-spline basis functions, a subclass of $\phi_{l,t}$, $\psi_{l,t} = \{\sum_{l=1}^{q_n} \alpha_l B_l\left(t\right), \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_{q_n}\}$ is a collection of monotone nondecreasing splines as a consequence of the variation diminishing properties of B-splines (Schumaker 1981). Therefore $\psi_{l,t}$ is a proper space from which the estimates of $\log \Lambda_0\left(t\right)$ can be found.

Using spline-based sieve approximation, the proportional mean function shown in Equation (1.2) is rewritten as

$$E\left(\mathbb{N}\left(T\right)|Z\right) = exp\{\sum_{l=1}^{q_n} \alpha_l B_l\left(t\right) + \beta_0^T Z\}. \tag{2.2}$$

Based on the same approximation, Lu et al. (2009) studied the spline-based sieve semiparametric maximum pseudolikelihood estimator and the spline-based sieve semiparametric maximum likelihood estimator as counterparts of Wellner & Zhang's estimators (Wellner & Zhang 2007) assuming a nonhomogeneous Poisson process.

The pseudolikelihood is given by

$$\begin{aligned} l_n^{ps}\left(\Lambda,\beta|D\right) &= \sum_{i=1}^{n}\sum_{j=1}^{K_i}\Big(\mathbb{N}^{(i)}\left(T_{K_i,j}^{(i)}\right)\beta^T Z_i+\mathbb{N}^{(i)}\left(T_{K_i,j}^{(i)}\right)log\Lambda\left(T_{K_i,j}^{(i)}\right) \\ &\qquad -exp\left\{log\Lambda\left(T_{K_i,j}^{(i)}\right)+\beta^T Z\right\}\Big) \end{aligned}$$

Its spline-based sieve counterpart is

$$\begin{aligned} \tilde{l}_n^{ps}\left(\alpha,\beta|D\right) &= \sum_{i=1}^{n}\sum_{j=1}^{K_i}\Bigg(\mathbb{N}^{(i)}\left(T_{K_i,j}^{(i)}\right)\beta^T Z_i+\mathbb{N}^{(i)}\left(T_{K_i,j}^{(i)}\right)\sum_{l=1}^{q_n}\alpha_l B_l\left(T_{K_i,j}^{(i)}\right) \\ &\qquad -exp\left\{\sum_{l=1}^{q_n}\alpha_l B_l\left(T_{K_i,j}^{(i)}\right)+\beta^T Z\right\}\Bigg). \end{aligned} \tag{2.3}$$

Both likelihood functions are derived based on the assumption that the cumulative counts follow an independent Poisson distribution and neglect the correlations between the cumulative counts within the same subject. Thus the two maximum pseudolikelihood estimators are not efficient.

Using the independence of the count increment based on the nonhomogeneous Poisson process assumption, the likelihood is given by

$$l_n\left(\Lambda,\beta|D\right)=\sum_{i=1}^{n}\sum_{j=1}^{K_i}\left[\Delta\mathbb{N}_{K_i,j}^{(i)}log\Delta\Lambda_{K_i,j}^{(i)}+\Delta\mathbb{N}_{K_i,j}^{(i)}\beta^T Z_i-e^{\beta^T Z_i}\Delta\Lambda_{K_i,j}^{(i)}\right] \tag{2.4}$$

where

$$\Delta\Lambda_{K_i,j}^{(i)}=\Lambda\left(T_{K_i,j}^{(i)}\right)-\Lambda\left(T_{K_i,j-1}^{(i)}\right); \qquad \Delta\mathbb{N}_{K_i,j}^{(i)}=\mathbb{N}_{K_i,j}^{(i)}\left(T_{K_i,j}^{(i)}\right)-\mathbb{N}_{K_i,j}^{(i)}\left(T_{K_i,j-1}^{(i)}\right)$$

Its spline-based sieve counterpart is

$$\tilde{l}_n\left(\alpha,\beta|D\right)=\sum_{i=1}^{n}\sum_{j=1}^{K_i}\left[\Delta\mathbb{N}_{K_i,j}^{(i)}log\Delta\tilde{\Lambda}_{K_i,j}^{(i)}+\Delta\mathbb{N}_{K_i,j}^{(i)}\beta^T Z_i-e^{\beta^T Z_i}\Delta\tilde{\Lambda}_{K_i,j}^{(i)}\right] \tag{2.5}$$

where

$$\Delta\tilde{\Lambda}^{(i)}_{K_i,j} = exp\left(\sum_{l=1}^{q_n}\alpha_l B_l\left(T^{(i)}_{K_i,j}\right)\right) - exp\left(\sum_{l=1}^{q_n}\alpha_l B_l\left(T^{(i)}_{K_i,j-1}\right)\right)$$

$\Delta\mathbb{N}^{(i)}_{K_i,j}$ is defined the same as in Equation (2.4). The (sieve) maximum likelihood estimators incorporate the correlations between cumulative counts and they are more efficient than the (sieve) pseudolikelihood estimators at the cost of more computing time. Both (sieve) maximum pseudolikelihood estimator and (sieve) maximum likelihood estimator are consistent. The maximum pseudolikelihood estimator and the maximum likelihood estimator converges at a rate of $n^{1/3}$, the regular convergence rate of nonparametric estimator. However the spline-based sieve estimators converge at a faster rate than their counterparts, but still slower than $n^{1/2}$. Despite a slower convergence rate of the nonparametric estimator, the estimate of the regression parameter still converges at $n^{1/2}$, and the (sieve) maximum likelihood estimator is more efficient than the (sieve) maximum pseudolikelihood estimator.

The (sieve) maximum likelihood estimator, though more efficient than the (sieve) maximum pseudolikelihood estimator, is still based on the model assuming independent increments. This assumption is often violated in medical settings because a high incidence of a disease in an interval may indicate another high incidence of the disease in the subsequent non-overlapping intervals. When this is the case, the (sieve) maximum likelihood estimator may not be an efficient estimator either. Instead of constructing a likelihood function based on specific distribution assumptions of the underlying counting process, we propose to use a generalized estimating equation (GEE) for the analysis of panel count data.

2.3 Spline-based Sieve Semiparametric Generalized Estimating Equation

Generalized Estimating Equation (GEE) method, originally developed by Liang & Zeger (1986) is widely used in parametric regression settings. It provides a robust inference with only weak assumptions of the underlying distributions. A large amount of literature generalized the same idea to semiparametric settings with a mean response model given by Equation (1.1). Zeger & Diggle (1994), Hoover et al. (1998), Lin & Ying (2001) and Wu & Zhang (2002) among others, used kernel-based estimating equation and ignored the correlation structure. Lin & Carroll (2001), Fan & Li (2004) and Wang et al. (2005) incorporated the correlation structure in their estimating procedures within the kernel framework.

We use a spline-based sieve semiparametric generalized estimating equation (sieve GEE), with the conditional mean function given by Equation (2.2), and estimate (β_0, Λ_0) through finding the roots of the estimating equation

$$U(\theta) = \sum_{i=1}^{n} \left(\frac{\partial \mu^{(i)}}{\partial \theta} \right)^T V^{(i)^{-1}} \left(\mathbb{N}(T_i) - \mu^{(i)} \right) = 0 \tag{2.6}$$

where $\mu^{(i)} = \left(\mu^{(i)}_{K_i,1}, \mu^{(i)}_{K_i,2}, \cdots \mu^{(i)}_{K_i,K_i} \right)^T$ with $\mu^{(i)}_{K_i,j} = exp\left(\beta^T Z_i + \sum_{l=1}^{q_n} \alpha_l B_l \left(T^{(i)}_{K_i,j} \right) \right)$ for $j = 1, 2, \cdots K_i$ and $\theta = (\beta, \alpha)$ with the constraints $\alpha_1 \le \alpha_2 \le \cdots \le \alpha_{q_n}$. The spline-based sieve GEE estimator of Λ_0 is taken as $exp\left(\sum_{l=1}^{q_n} \hat{\alpha}_l B_l(t) \right)$ after the spline coefficient estimates $\hat{\alpha}_l, l = 1, 2, \cdots, q_n$ are obtained from Equation (2.6). $V^{(i)}$ is the working covariance matrix for the panel counts from the i^{th} process. Different choices of this covariance matrix could accommodate the characteristics of different counting processes. We discuss three possible covariance matrices and they correspond to scores of different 'likelihood' functions. The equivalence is shown in Appendix A.

2.3.1 Diagonal Covariance Matrix

The easiest choice of the covariance matrix is to use a diagonal matrix, in which the diagonal element is determined by the variance function of Poisson distribution, e.g., $Var\left(\mathbb{N}\left(T_{K_i,j}\right)\right) = E\left(\mathbb{N}\left(T_{K_i,j}\right)\right)$ and

$$V_1^{(i)} = \begin{pmatrix} \mu_{K_i,1}^{(i)} & 0 & \cdots & 0 \\ 0 & \mu_{K_i,2}^{(i)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mu_{K_i,K_i}^{(i)} \end{pmatrix}_{K_i \times K_i}$$

This is exactly the score equation of the log pseudolikelihood shown in Equation (2.3) (see Appendix A). However, the diagonal matrix implies an independence between cumulative counts in the counting process, although these cumulative counts are obviously positively correlated. The misspecification of the covariance matrix causes a loss of efficiency in the estimation.

2.3.2 Covariance Matrix Based on the Poisson Process Assumption

Instead of using a diagonal matrix that ignores the correlation among the cumulative counts, a covariance matrix that accommodates such correlation will intuitively produce more efficient estimates. The covariance function based on the Poisson counting process, i.e.,

$$Cov\left(\mathbb{N}\left(t_1\right), \mathbb{N}\left(t_2\right)\right) = E\left(\mathbb{N}\left(t_1\right)\right), \quad \text{for } t_1 \leq t_2$$

leads to the selection of the covariance matrix V_i in the form of

$$V_2^{(i)} = \begin{pmatrix} \mu_{K_i,1}^{(i)} & \mu_{K_i,1}^{(i)} & \cdots & \mu_{K_i,1}^{(i)} \\ \mu_{K_i,1}^{(i)} & \mu_{K_i,2}^{(i)} & \cdots & \mu_{K_i,2}^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{K_i,1}^{(i)} & \mu_{K_i,2}^{(i)} & \cdots & \mu_{K_i,K_i}^{(i)} \end{pmatrix}_{K_i \times K_i} .$$

The spline-based sieve semiparametric GEE with this covariance matrix is exactly the score equation of the log likelihood based on Poisson process model shown in Equation (2.5) (see Appendix A). In spite of the improved efficiency of this estimation compared to the one using $V_1^{(i)}$ as the covariance matrix, it still imposes possibly unrealistic assumptions to the covariance structure of the data: first, it assumes the variance of the counts equals to the mean, that is no over-dispersion is allowed in the count data; Secondly, it assumes the independence between the count increments. When either of these two assumptions is violated, the estimator based on $V_2^{(i)}$ will not be very efficient.

2.3.3 Sieve GEE With Over-Dispersion Term

Although the results of Lu et al. (2009) have demonstrated that the sieve GEE estimate of β_0 with $V_2^{(i)}$ is more efficient than that with $V_1^{(i)}$, it is not guaranteed that using $V_2^{(i)}$ would always produce a highly efficient estimate, as it does not account for either the over-dispersion or the correlation among the count increments. In literature, Poisson model with a frailty variable, namely $E\left(\mathbb{N}\left(t\right)|\gamma, Z\right) = \gamma\Lambda_0\left(t\right)e^{\beta^T Z}$, is a common way in parametric regression analysis for count data to account for possible over-dispersions. Chan & Ledolter (1995) and Hay & Pettitt (2001) discussed a log normal frailty model by assuming a log normal distribution of the frailty term γ. However, there is no close form for the marginal distribution of the count and the estimation with this frailty variable is computationally demanding. Another common frailty model assumes a gamma-distributed subject-specific frailty term as studied in Thall (1988) and Diggle et al. (1994) among others. Integrating out the gamma frailty variable results in a negative binomial distribution for the correlated counts. Zhang & Jamshidian (2004) introduced a gamma frailty term to nonparametric estimation of the mean function of the counting process. They constructed a maximum pseudolikelihood estimate with a gamma frailty term and computed the estimate using EM algorithm. Zeger (1988) considered a latent frailty process while assuming only the mean of the frailty term and a covariance function. A similar idea is adopted in the semiparametric sieve GEE setting in this manuscript. The expectation of γ is specified as 1, e.g., $E\left(\gamma\right) = 1$, which guarantees the identifiability of the model and does not violate the proportional mean model specified in Equation (1.2). The variance of γ is denoted as σ^2. The marginal variance function based on such Frailty Poisson process is $Var\left(\mathbb{N}\left(t\right)\right) = \mu_t + \sigma^2\mu_t^2$, where $\mu_t = E\left(\mathbb{N}\left(t\right)\right)$. The correlation between successive counts is explained by the frailty parameter γ, namely $Cov\left(\mathbb{N}\left(t_1\right), \mathbb{N}\left(t_2\right)\right) = \mu_{t_1} + \sigma^2\mu_{t_1}\mu_{t_2}$, for $t_1 \leq t_2$.

This leads to a working covariance matrix $V_3^{(i)}$ of the form

$$\begin{pmatrix} \mu_{K_i,1}^{(i)} + \sigma^2 \mu_{K_i,1}^{(i)} \mu_{K_i,1}^{(i)} & \mu_{K_i,1}^{(i)} + \sigma^2 \mu_{K_i,1}^{(i)} \mu_{K_i,2}^{(i)} & \cdots & \mu_{K_i,1}^{(i)} + \sigma^2 \mu_{K_i,1}^{(i)} \mu_{K_i,K_i}^{(i)} \\ \mu_{K_i,1}^{(i)} + \sigma^2 \mu_{K_i,1}^{(i)} \mu_{K_i,2}^{(i)} & \mu_{K_i,2}^{(i)} + \sigma^2 \mu_{K_i,2}^{(i)} \mu_{K_i,2}^{(i)} & \cdots & \mu_{K_i,2}^{(i)} + \sigma^2 \mu_{K_i,2}^{(i)} \mu_{K_i,K_i}^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{K_i,1}^{(i)} + \sigma^2 \mu_{K_i,1}^{(i)} \mu_{K_i,K_i}^{(i)} & \mu_{K_i,2}^{(i)} + \sigma^2 \mu_{K_i,2}^{(i)} \mu_{K_i,K_i}^{(i)} & \cdots & \mu_{K_i,K_i}^{(i)} + \sigma^2 \mu_{K_i,K_i}^{(i)} \mu_{K_i,K_i}^{(i)} \end{pmatrix}_{K_i \times K_i}$$

and it can be rewritten as,

$$V_3^{(i)} = V_2^{(i)} + \sigma^2 \left(\mu^{(i)}\right) \left(\mu^{(i)}\right)^T$$

The estimating equation using $V_2^{(i)}$ is a special case of $V_3^{(i)}$ with $\sigma^2 = 0$. When over-dispersion exists, the spline-based sieve semiparametric GEE method using this working covariance matrix with σ^2 replaced by its consistent estimate may lead to a more efficient estimate than the spline-based sieve maximum likelihood estimate studied by Lu et al. (2009). The estimating equation using $V_3^{(i)}$ turns out to be the score function of the marginal likelihood of the panel count data under the Gamma-Frailty nonhomogeneous Poisson process model. That is, given the gamma distribution of the frailty term, e.g., $\gamma \sim \Gamma\left(1/\sigma^2, 1/\sigma^2\right)$, the cumulative counts follow a nonhomogeneous Poisson process with mean $\gamma \Lambda(t) e^{\beta^T Z}$. The conditional likelihood of the counts given the frailty term can be written as

$$f\left(\mathbb{N}_1, \mathbb{N}_2, \cdots, \mathbb{N}_K | \gamma\right) = \Pi_{j=1}^{k} \frac{e^{-\gamma \Delta\Lambda_j e^{\beta^T Z}} \left(\gamma \Delta\Lambda_j e^{\beta^T Z}\right)^{\Delta\mathbb{N}_j}}{\Delta\mathbb{N}_j!}$$

where $\mathbb{N}_j = \mathbb{N}\left(T_{K,j}\right), \Delta\mathbb{N}_j = \mathbb{N}_j - \mathbb{N}_{j-1}$ and $\Lambda_j = \Lambda\left(T_{K,j}\right), \Delta\Lambda_j = \Lambda_j - \Lambda_{j-1}$ for $j = 1, 2, \cdots, K$. We let $T_{K,0} \equiv 0$ and assume $\mathbb{N}(0) = \Lambda(0) = 0$.

Integrating out γ, we have

$$\begin{aligned} f\left(\mathbb{N}_1, \mathbb{N}_2, \cdots, \mathbb{N}_K\right) &= \int_\gamma \Pi_{j=1}^K \frac{e^{-\gamma \Delta\Lambda_j e^{\beta^T Z}} \left(\gamma \Delta\Lambda_j e^{\beta^T Z}\right)^{\Delta\mathbb{N}_j}}{\Delta\mathbb{N}_j!} \frac{\left(1/\sigma^2\right)^{1/\sigma^2}}{\Gamma\left(1/\sigma^2\right)} e^{-1/\sigma^2 \gamma} \gamma^{1/\sigma^2 - 1} d\gamma \\ &= \frac{\left(\Delta\Lambda_j e^{\beta^T Z}\right)^{\Delta\mathbb{N}_j} \left(1/\sigma^2\right)^{1/\sigma^2}}{\Pi_{j=1}^K \Delta\mathbb{N}_j! \Gamma\left(1/\sigma^2\right)} \int_\gamma e^{-\left(\Lambda_K e^{\beta^T Z} + 1/\sigma^2\right)\gamma} \gamma^{\mathbb{N}_K + 1/\sigma^2 - 1} d\gamma \\ &= \frac{\left(\Delta\Lambda_j e^{\beta^T Z}\right)^{\Delta\mathbb{N}_j} \left(1/\sigma^2\right)^{1/\sigma^2}}{\Pi_{j=1}^K \Delta\mathbb{N}_j! \Gamma\left(1/\sigma^2\right)} \frac{\Gamma\left(\mathbb{N}_K + 1/\sigma^2\right)}{\left(\Lambda_K e^{\beta^T Z} + 1/\sigma^2\right)^{\mathbb{N}_K + 1/\sigma^2}} \end{aligned}$$

The log likelihood based on this model is,

$$\begin{aligned} l\left(\beta, \Lambda, \sigma^2; X_i\right) = &\sum_{i=1}^n \left\{ \sum_{j=1}^K \Delta\mathbb{N}_{K_i,j}^{(i)} log\left(\Delta\Lambda_{K_i,j}^{(i)} e^{\beta^T Z_i}\right) - \left(\mathbb{N}_{K_i,K_i}^{(i)} + 1/\sigma^2\right) \times \right. \\ &\qquad log\left(\Lambda_{K_i,K_i}^{(i)} e^{\beta^T Z} + 1/\sigma^2\right) + 1/\sigma^2 \times log 1/\sigma^2 + \\ &\qquad \left. log\Gamma\left(\mathbb{N}_{K_i,K_i}^{(i)} + 1/\sigma^2\right) - log\Gamma\left(1/\sigma^2\right) \right\} \end{aligned} \tag{2.7}$$

The score function of this likelihood is the same as the sieve GEE using $V_3^{(i)}$ as the working covariance matrix (see Appendix A).

The sieve semiparametric GEE estimator with $V_1^{(i)}$ as the covariance matrix coincide with the sieve semiparametric maximum pseudolikelihood estimator $\left(\hat{\Lambda}_n^{ps}, \hat{\beta}_n^{ps}\right)$ and the sieve semiparametric GEE estimator using $V_2^{(i)}$ as the working covariance matrix is the same as the sieve semiparametric maximum likelihood estimator $\left(\hat{\Lambda}_n, \hat{\beta}_n\right)$. The consistency and convergence rate of $\left(\hat{\Lambda}_n^{ps}, \hat{\beta}_n^{ps}\right)$ and $\left(\hat{\Lambda}_n, \hat{\beta}_n\right)$ and the asymptotic normality of $\hat{\beta}_n^{ps}$ and $\hat{\beta}_n$ are proved in Lu et al. (2009). The asymptotic properties of this sieve semiparametric GEE estimator using $V_3^{(i)}$ as the covariance matrix and σ^2 replaced by its consistent estimate $\hat{\sigma}_n^2$ are discussed in Chapter 3.

CHAPTER 3
ASYMPTOTIC PROPERTIES OF SPLINE-BASED GEE ACCOUNTING FOR OVERDISPERSION

In this chapter, we apply modern empirical process theory to prove the consistency, convergence rate and the asymptotic normality of our sieve GEE estimator with $V_3^{(i)}$ as the working covariance matrix and σ^2 replaced by its consistent estimate $\hat{\sigma}_n^2$. In Section 3.1 we present some technical terms and lemmas in modern empirical process theory. In Section 3.2 we develop three general theorems for the asymptotic properties of the pseudo GEE (or pseudo MLE) estimator. In Section 3.3 these theorems are further applied to the Gamma-Frailty nonhomogeneous Poisson process model to prove the asymptotic properties of our proposed pseudo spline-based sieve GEE estimators.

3.1 Basic elements of modern empirical process theory

In this section we present some technical terms and lemmas in modern empirical process theory from the book by van der Vaart & Wellner (1996). These results will be used to prove the asymptotic properties of our pseudo GEE (pseudo MLE) estimator in the next two sections.

Let $X_1, X_2, \cdots, X_n$ be a random sample from a probability distribution P on a measurable space $(\Omega, \mathcal{B})$. For a measurable function $f : \mathcal{X} \mapsto \mathbb{R}$, let Pf denote the integral $\int f dP$, equivalently it is the expectation of f under the probability measure P, i.e., $E_P f(X)$. Let $\mathbb{P}_n$ denote the discrete uniform measure, i.e.,

$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$. It is the expectation of f under empirical measure $\mathbb{P}_n$. The empirical process $\mathbb{G}_n f$ is the centered and scaled version of the empirical measure, i.e., $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n f - Pf) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(X_i) - E_p f(X_i))$.

By the law of large numbers and central limit theorem, for a fixed function f, it follows

$$\mathbb{P}_n f \rightarrow_{a.s.} Pf \text{ and } \mathbb{G}_n f \rightarrow_d N\left(0, P(f - Pf)^2\right).$$

provided Pf exists and $Pf^2 < \infty$, respectively.

When dealing with the set to which parameters belong, a uniform version of law of large numbers and central limit theorem is defined in modern empirical process theory. A class $\mathcal{F}$ of measurable functions $f : \mathcal{F} \mapsto \mathbb{R}$ is called *P-Glivenko-Cantelli* if

$$\|\mathbb{P}_n f - Pf\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \|\mathbb{P}_n f - Pf\| \to 0 \quad \text{almost surely.}$$

A class $\mathcal{F}$ of measurable functions $f : \mathcal{F} \mapsto \mathbb{R}$ is called *P-Donsker* if the sequence of processes $\{\mathbb{G}_n f : f \in \mathcal{F}\}$ converges in distribution to a tight limit process in the space $l^{\infty}(\mathcal{F})$.

Whether a class of functions $\mathcal{F}$ is a Glivenko-Cantelli or Donsker class depends on the size of the class. A relatively simply way to measure the size of a class $\mathcal{F}$ is in terms of *entropy.* For any probability measure P, define $L_r(P) = \left\{f : \int f^r dP < \infty\right\}$. For any element of $\mathcal{F}$, f, define a metric as

$$\|f\|_{L_r(P)} = \left(P\left(|f|^r\right)\right)^{1/r} = \left(\int_{\Omega} |f(x)|^r \, dP(x)\right)^{1/r}$$

The *covering number* $N(\varepsilon, \mathcal{F}, \|\cdot\|)$ is the minimal number of balls $\{g : \|g - f\| < \varepsilon\}$ of radius ε needed to cover the set $\mathcal{F}$. The *entropy (without bracketing)* is the logarithm of the covering number. Given two functions l and u, the *bracket* $[l,u]$ is the set of all functions f with $l \leq f \leq u$. An $\varepsilon-$ *bracket* in $L_r(P)$ is a bracket $[l,u]$ with $\|u-l\|_{L_r(P)} < \varepsilon$. The *bracketing number* $N_{[]}(\varepsilon, \mathcal{F}, L_r(P))$ is the minimum number of $\varepsilon-$brackets needed to cover $\mathcal{F}$. The *entropy with bracketing* is the logarithm of the bracketing number.

Remark: If f is in the 2ε-bracket $[l,u]$, then it is in the ball of radius ε around $(l+u)/2$. So the covering and bracketing number are related by

$$N(\varepsilon, \mathcal{F}, \|\cdot\|) \leq N_{[]}(2\varepsilon, \mathcal{F}, \|\cdot\|)$$

The Glivenko-Cantelli Theorem as stated in Theorem 2.4.1 in van der Vaart & Wellner (1996) relates Glivenko-Cantelli class with the bracketing number .

Lemma 3.1. *(Glivenko-Cantelli Theorem) Let $\mathcal{F}$ be a class of measurable functions such that $N_{[]}(\varepsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\varepsilon > 0$. Then $\mathcal{F}$ is P-Glivenko-Cantelli.*

Theorem 2.4.3 in van der Vaart & Wellner (1996) as stated in the next lemma relate the Glivenko-Cantelli with a random entropy condition.

Lemma 3.2. *Let $\mathcal{F}$ be a P-measurable class of measurable functions with envelope F such that $PF < \infty$. Let $\mathcal{F}_M$ be the class of functions $f1\{F \leq M\}$ when f ranges over $\mathcal{F}$. if $logN(\varepsilon, \mathcal{F}_M, L_1(\mathbb{P}_n)) = o_p(n)$ for every ε and $M > 0$, then $\|\mathbb{P}_n - P\|_{\mathcal{F}} \to 0$ both almost surely and in mean. In particular, $\mathcal{F}$ is Glivenko-Cantelli.*

The bracketing number $N_{[]}(\varepsilon, \mathcal{F}, L_r(P))$ grows to infinity as $\varepsilon \downarrow 0$. A sufficient condition for a class to be Donsker is that they do not grow too fast. The following theorem from Section 2.5 in van der Vaart & Wellner (1996) relates P-Donsker with the *bracketing integral* defined therein.

Lemma 3.3. *(Donsker Theorem) Let $\mathcal{F}$ be a class of measurable functions such that its bracketing integral defined as*

$$J_{[]}\left(\delta, \mathcal{F}, L_2\left(P\right)\right) = \int_0^\delta \sqrt{logN_{[]}\left(\varepsilon, \mathcal{F}, L_2\left(P\right)\right)}d\varepsilon < \infty$$

for every $\varepsilon > 0$. Then $\mathcal{F}$ is P-Donsker.

Remark: The integrand is a decreasing function of ε. So the convergence of the integral depends on the size of the bracketing number for $\varepsilon \downarrow 0$. Because $\int_0^1 \varepsilon^{-r} d\varepsilon$ converges for $r < 1$, the integral condition requires the entropy grows no faster than the order of $(1/\varepsilon)^2$ for a Donsker class.

In our analysis of panel count data, the baseline mean function in Equation (1.2) is a monotone nondecreasing function. Theorem 2.7.5 in van der Vaart & Wellner (1996), as stated below, indicates that the class of uniformly bounded, monotone functions on the real line is Donsker, the bracketing entropy of this class is of the order $1/\varepsilon$.

Lemma 3.4. *The class $\mathcal{F}$ of monotone functions $f : \mathbb{R} \mapsto [0, 1]$ satisfies*

$$logN_{[]}\left(\varepsilon, \mathcal{F}, L_r\left(Q\right)\right) \leq K\left(\frac{1}{\varepsilon}\right),$$

for every probability measure Q, every $r \geq 1$, and a constant K that depends on r only.

There is a nice relationship between Donsker class and the asymptotic equicontinuity. We now state Corollary 2.3.13 from van der Vaart & Wellner (1996) about this relationship in the following lemma.

Lemma 3.5. *(Semi-equicontinuity Theorem) Let $\mathcal{F}$ be a class of measurable functions. Define a seminorm ρ_P on $\mathcal{F}$ by*

$$\rho_P(f) = \left(P(f - Pf)^2\right)^{1/2}, \ \textit{for } f \in \mathcal{F}$$

Let

$$\mathcal{F}_\delta = \{f - g : f, g \in \mathcal{F}, \rho_P(f - g) < \delta\}$$

Then the following are equivalent:

1. *$\mathcal{F}$ is P-Donsker;*
2. *$(\mathcal{F}, \rho_P)$ is totally bounded and*
$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} P\left(\sup_{\rho_P(f-g)<\delta} |\mathbb{G}_n(f - g)| > \varepsilon\right) = 0.$$
3. *$(\mathcal{F}, \rho_P)$ is totally bounded and*
$$E\sqrt{n}\,\|\mathbb{P}_n - P\|_{\mathcal{F}_{\delta_n}} \to 0, \ \textit{for every } \delta_n \to 0.$$

Lemma 3.4.3 in van der Vaart & Wellner (1996) is used in the proof of the convergence rate. It involves a specific norm and a different entropy integral from that defined above. We define ‘Bernstein norm’ as

$$\|f\|_{P,B} = \left(2P\left(e^{|f|} - 1 - |f|\right)\right)^{1/2}$$

A different bracketing integral using Bernstein norm is defined as

$$\tilde{J}_{[]}\left(\delta,\mathcal{F},\|\cdot\|_{P,B}\right)=\int_0^\delta\sqrt{1+logN_{[]}\left(\varepsilon,\mathcal{F},\|\cdot\|_{P,B}\right)}d\varepsilon$$

Lemma 3.6. *Let $\mathcal{F}$ be a class of measurable functions such that $\|f\|_{P,B}\leq\delta$ for every $f\in\mathcal{F}$. Then*

$$E_P\|\mathbb{G}_n\|_{\mathcal{F}}\leq C\tilde{J}_{[]}\left(\delta,\mathcal{F},\|\cdot\|_{P,B}\right)\left(1+\frac{\tilde{J}_{[]}\left(\delta,\mathcal{F},\|\cdot\|_{P,B}\right)}{\delta^2\sqrt{n}}M\right).$$

In this manuscript, we work with a spline-based sieve space. Given the number of knots, q_n, a set of knots is denoted by

$$t=\{t_1=t_2=\cdots=t_l<t_{l+1}<\cdots<t_{m_n+l}=t_{m_n+l+1}=\cdots=t_{m_n+2l}\}$$

We approximate the compact original function space by $\phi_{l,t}$ with order l and knots t, where

$$\phi_{l,t}=\{\sum_{i=1}^{q_n}a_iB_i:B_i,i=1,2,\cdots,q_n \text{ are the basis functions with knots } t$$
$$\text{and }\sum_{i=1}^{q_n}a_i^2\leq\delta^2\text{ for some constant }\delta\}$$

To apply either the Glivenko-Cantelli theorem or the Donsker theorem, we need to calculate the entropy numbers with bracketing of this sieve space using different norms. This can be done by applying Lemma 5 in Shen & Wong (1994) to the spline-based sieve space $\phi_{l,t}$ with different norms. We first stated Shen & Wong (1994)'s lemma below.

Lemma 3.7. *(A metric entropy calculation). Let* S *be a* $\delta-$*sphere in* $\mathbb{R}^n$*, that is,* $S = \{x = (x_1, \cdots, x_n) \in \mathbb{R}^n : \sum_{i=1}^n x_i^2 \leq \delta^2\}$*. Let* $\|\cdot\|_{L_1}$ *be the usual* L_1-*metric in* $\mathbb{R}^n$*. Then* $logN_{[]}\left(\varepsilon, S, \|\cdot\|_{L_1}\right) \leq cnlog\left(n^{1/2}\delta/\varepsilon\right)$ *for* $\varepsilon < \delta$ *and some constant* $c > 0$*.*

Following the same line of Shen & Wong (1994)'s proof, we calculate the bracketing entropies of the $\delta-$ sphere defined in Lemma 3.7 using L_2- and $L_\infty-$norm. That is

$$logN_{[]}\left(\varepsilon, S, \|\cdot\|_{L_2}\right) \leq cnlog\left(\delta/\varepsilon\right)$$

$$logN_{[]}\left(\varepsilon, S, \|\cdot\|_{L_\infty}\right) \leq cnlog\left(n^{-1/2}\delta/\varepsilon\right)$$

To calculate the entropy number of $\phi_{l,t}$, we apply Lemma 3.7 replacing the $\delta-$sphere with the sphere defined using the spline coefficients.

Lemma 3.8. *The entropy numbers of* $\phi_{l,t}$ *with* L_1-*,* L_2- *and* $L_\infty-$ *norms are bounded by* $Cq_nlog(q_n^{1/2} \times \frac{\delta}{\varepsilon}), Cq_nlog(\frac{\delta}{\varepsilon})$ *and* $Cq_nlog\left(\frac{\delta}{q_n^{1/2}\varepsilon}\right)$ *respectively.*

For the estimation of the mean function of the panel count data specified in Equation (1.2), we approximate the space of the log of the baseline mean function, $log\Lambda$ by a subspace of $\phi_{l,t}$, $\psi_{l,t}$, defined as

$$\psi_{l,t} = \{\sum_{i=1}^{q_n} a_iB_i : B_i, i = 1, 2, \cdots, q_n \text{ are the basis functions defined at } t$$

$$a_1 \leq a_2 \leq \cdots, \leq a_{q_n} \text{ and } \sum_{i=1}^{q_n} a_i^2 \leq \delta^2 \text{ for some constant } \delta\}$$

Obviously, the $\varepsilon-$entropy numbers of $\psi_{l,t}$ with L_1-, L_2- and $L_\infty-$ norms are also bounded by $Cq_nlog(q_n^{1/2} \times \frac{\delta}{\varepsilon}), Cq_nlog(\frac{\delta}{\varepsilon})$ and $Cq_nlog\left(\frac{\delta}{q_n^{1/2}\varepsilon}\right)$ respectively.

3.2 General theorems

3.2.1 Consistency

We generalize Theorem 5.7 in van der Vaart (1998) by including a nuisance parameter η.

Theorem 3.9. *Let $\mathbb{M}_n(\theta,\eta)$ and $\mathbb{M}(\theta,\eta)$ be a random function and a fixed function in an index set $\Theta \times H$ respectively. Denote H° the neighborhood of a fixed value η_0 in H. If*

$$\sup_{\theta:d(\theta,\theta_0)>\varepsilon} \mathbb{M}(\theta,\eta) < \mathbb{M}(\theta_0,\eta) \quad \forall\varepsilon>0, \eta\in H^\circ \tag{3.1}$$

$$\sup_{\{(\theta,\eta):\theta\in\Theta,\eta\in H^\circ\}} |\mathbb{M}_n(\theta,\eta)-\mathbb{M}(\theta,\eta)| \to_p 0, \tag{3.2}$$

then any sequence of estimator $\hat{\theta}_n$ with

$$\inf_{\eta\in H^\circ}\left(\mathbb{M}_n\left(\hat{\theta}_n,\eta\right)-\mathbb{M}_n(\theta_0,\eta)\right) \geq -o_p(1) \tag{3.3}$$

converges in probability to θ_0.

Proof. By condition (3.2), we have $\mathbb{M}_n(\theta_0,\eta) = \mathbb{M}(\theta_0,\eta)+o_p(1)$. Together with condition (3.3), this further implies $\mathbb{M}_n\left(\hat{\theta}_n,\eta\right) \geq \mathbb{M}_n(\theta_0,\eta)-o_p(1) = \mathbb{M}(\theta_0,\eta) - o_p(1)$. So we have

$$\begin{aligned}\mathbb{M}(\theta_0,\eta)-\mathbb{M}\left(\hat{\theta}_n,\eta\right) &\leq \mathbb{M}_n\left(\hat{\theta}_n,\eta\right)-\mathbb{M}\left(\hat{\theta}_n,\eta\right)+o_p(1)\\ &\leq \sup_{\{(\theta,\eta):\theta\in\Theta,\eta\in H^\circ\}}|\mathbb{M}_n(\theta,\eta)-\mathbb{M}(\theta,\eta)|+o_p(1)\to_p 0 \end{aligned} \tag{3.4}$$

By condition (3.1), for any $\delta>0$, we can find $\varepsilon>0$ such that $\mathbb{M}(\theta_0,\eta)-\mathbb{M}(\theta,\eta)\geq\delta$ for every θ that satisfies $d(\theta,\theta_0)>\varepsilon$. So the event $[d(\theta,\theta_0)\geq\varepsilon]$ is a subset of

$[\mathbb{M}(\theta_0, \eta) - \mathbb{M}(\theta, \eta) \geq \delta]$. In view of the inequality (3.4), we have

$$d\left(\hat{\theta}_n, \theta_0\right) \rightarrow_p 0.$$

$\square$

Remark: This theorem is a direct generalization of the Theorem 5.7 in van der Vaart (1998) by including a nuisance parameter η. Condition (3.1) indicates that θ_0 maximizes $\mathbb{M}(\theta, \eta)$ for any given nuisance parameter η.

In applications, with the extra condition specified in condition (3.1) a two-stage estimating procedure could be implemented. Instead of estimating (θ, η) simultaneously by maximizing the original likelihood $\mathbb{M}_n(\theta, \eta)$, we could estimate the nuisance parameter η first and then estimate θ by maximizing a pseudo likelihood $\mathbb{M}_n(\theta, \hat{\eta}_n)$. The estimator, $\hat{\theta}_n$ based on such a two-stage estimating procedure still converges to the true parameter θ_0. The estimation of the sieve GEE using $V_3^{(i)}$ as the working covariance matrix can be implemented in such a two-stage procedure. The consistency of the estimate is established in Section 3.3.

3.2.2 Convergence Rate

Theorem 3.10 is a generalization of the Theorem 3.4.1 in van der Vaart & Wellner (1996) with an extra nuisance parameter, η.

Theorem 3.10. *Let* $\mathbb{M}_n(\theta, \eta)$ *and* $M_n(\theta, \eta)$ *be stochastic processes indexed by* $\Theta_n \times H$. *Denote* H° *the neighborhood of a fixed value* η_0 *in* H. *Let* $\theta_n \in \Theta_n$ *and* $0 < \delta_n < \zeta$ *be arbitrary, and let* $\theta \mapsto d_n(\theta, \theta_n)$ *be an arbitrary map from* Θ_n *to* $[0, \infty)$. *Suppose that* *for every n and $\delta_n < \delta \le \zeta$*

$$\sup_{\delta/2<d_n(\theta,\theta_n)\le\delta,\theta\in\Theta_n,\eta\in H^\circ} \left(M_n(\theta,\eta) - M_n(\theta_n,\eta)\right) \le -C\delta^2 \tag{3.5}$$

$$E\left\{\sup_{\delta/2<d_n(\theta,\theta_n)\le\delta,\theta\in\Theta_n,\eta\in H^\circ} \left|\left(\mathbb{M}_n - M_n\right)(\theta,\eta) - \left(\mathbb{M}_n - M_n\right)(\theta_n,\eta)\right|\right\} \le C\frac{\phi_n(\delta)}{\sqrt{n}} \tag{3.6}$$

for functions ϕ_n such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing on (δ_n,ζ) for some $\alpha < 2$ (not depending on n). If there is a $r_n = C/\delta_n$ such that

$$r_n^2\phi_n(1/r_n) \le C\sqrt{n} \text{ for every } n \tag{3.7}$$

and the sequence $\hat{\theta}_n$ takes its values in Θ_n and satisfies

$$\inf_{\eta\in H^\circ}\left(\mathbb{M}_n\left(\hat{\theta}_n,\eta\right) - \mathbb{M}_n(\theta_n,\eta)\right) \ge -O_p\left(r_n^{-2}\right) \tag{3.8}$$

and $d_n\left(\hat{\theta}_n,\theta_n\right) \to_p 0$; then

$$r_n d_n\left(\hat{\theta}_n,\theta_n\right) = O_p(1)$$

Proof. We first partition the parameter space Θ_n into different 'shells' defined by $S_{j,n} = \{\theta : 2^{j-1} < r_n d_n(\theta,\theta_n) \le 2^j\}$ with $j = 1, 2, \cdots$. The event $\left\{r_n d_n\left(\hat{\theta}_n,\theta_n\right) > 2^M\right\}$ for some M is a subset of the event $\left\{\hat{\theta}_n \in S_{j,n} : \text{ for some } j > M\right\}$. So for any $\eta \in H^\circ$ we have

$$\begin{aligned} P\left(r_n d\left(\hat{\theta}_n,\theta_n\right) > 2^M\right) &\le P\left(\hat{\theta}_n \in S_{j,n} \text{ for some } j > M\right) \\ &\le \sum_{j>M,2^j<r_n\varepsilon} P\left(\sup_{\theta\in S_{j,n}}\left(\mathbb{M}_n(\theta,\eta) - \mathbb{M}_n(\theta_n,\eta)\right) \ge -Cr_n^{-2}\right) \\ &\quad + P\left(2d_n\left(\hat{\theta}_n,\theta_n\right) > \varepsilon\right) \end{aligned}$$

By the consistency condition, the last probability goes to zero as $n \to \infty$ for any $\varepsilon > 0$.

We choose small enough ε and $\delta < \varepsilon$ such that both conditions (3.5) and (3.6) hold for $\theta \in S_{j,n}$. By condition (3.5), $\sup_{\eta \in H^\circ} \left(M_n(\theta, \eta) - M_n(\theta_n, \eta)\right) \leq -C2^{2j} r_n^{-2}$. Also,

$$\begin{aligned}
&\sup_{\theta \in S_{j,n}} \left\{\mathbb{M}_n(\theta, \eta) - \mathbb{M}_n(\theta_n, \eta)\right\} \\
&= \sup_{\theta \in S_{j,n}} \left\{(\mathbb{M}_n - M_n)(\theta, \eta) - (\mathbb{M}_n - M_n)(\theta_n, \eta) + M_n(\theta, \eta) - M_n(\theta_n, \eta)\right\} \\
&\leq \sup_{\theta \in S_{j,n}} \left\{(\mathbb{M}_n - M_n)(\theta, \eta) - (\mathbb{M}_n - M_n)(\theta_n, \eta)\right\} + \sup_{\theta \in S_{j,n}} \left\{M_n(\theta, \eta) - M_n(\theta_n, \eta)\right\}
\end{aligned}$$

So

$$\begin{aligned}
&P\left\{\sup_{\theta \in S_{j,n}} \left[\mathbb{M}_n(\theta, \eta) - \mathbb{M}_n(\theta_n, \eta)\right] \geq -Cr_n^{-2}\right\} \\
\leq &P\left\{\sup_{\theta \in S_{j,n}} \left[(\mathbb{M}_n - M_n)(\theta, \eta) - (\mathbb{M}_n - M_n)(\theta_n, \eta)\right]\right. \\
&\qquad\qquad \left. + \sup_{\theta \in S_{j,n}} \left[M_n(\theta, \eta) - M_n(\theta_n, \eta)\right] \geq -Cr_n^{-2}\right\} \\
= &P\left\{\sup_{\theta \in S_{j,n}} \left[(\mathbb{M}_n - M_n)(\theta, \eta) - (\mathbb{M}_n - M_n)(\theta_n, \eta)\right]\right. \\
&\qquad\qquad \left. \geq - \sup_{\theta \in S_{j,n}} \left[M_n(\theta, \eta) - M_n(\theta_n, \eta)\right] - Cr_n^{-2}\right\} \\
= &P\left\{\sup_{\theta \in S_{j,n}} \left((\mathbb{M}_n - M_n)(\theta, \eta) - (\mathbb{M}_n - M_n)(\theta_n, \eta)\right) \geq C2^{2j} r_n^{-2}\right\} \\
\leq &C\frac{E\| (\mathbb{M}_n - M_n)(\theta, \eta) - (\mathbb{M}_n - M_n)(\theta_n, \eta) \|_{S_{j,n}}}{2^{2j}/r_n^2} \qquad \text{(by Markov's Inequality)} \\
\leq &C\frac{\phi_n(2^j/r_n)}{\sqrt{n}2^{2j}/r_n^2} \qquad \text{(by Condition 3.6)} \qquad (**)
\end{aligned}$$

As $\phi_n(\delta)/\delta^\alpha$ is decreasing, $\frac{\phi_n(2^j/r_n)}{(2^j/r_n)^\alpha} \leq \frac{\phi_n(1/r_n)}{1/r_n^\alpha}$. This further implies $\phi_n(2^j/r_n) \leq 2^{\alpha j}\phi_n(1/r_n)$. Thus by (**) and Condition (3.8),

$$P\left(\sup_{\theta \in S_{j,n}} \left(\mathbb{M}_n(\theta, \eta) - \mathbb{M}_n(\theta_n, \eta)\right) \geq -Cr_n^{-2}\right) \leq C\frac{2^{\alpha j}\phi_n(1/r_n)}{\sqrt{n}2^{2j}/r_n^2} \leq C2^{(\alpha-2)j}$$

Therefore,

$$\sum_{j>M,2^j<r_n\varepsilon} P\left(\sup_{\theta\in S_{j,n}}\left(\mathbb{M}_n(\theta,\eta)-\mathbb{M}_n(\theta_n,\eta)\right)\geq -Cr_n^{-2}\right)\leq \qquad C\sum_{j>M}2^{(\alpha-2)j}$$

The last quantity converges to zero as $M\to\infty$, □

3.2.3 Asymptotic Normality

We generalize Theorem 6.1 in Wellner & Zhang (2007) by including an extra nuisance parameter. Given i.i.d. observations $X_1, X_2, \cdots, X_n$ and the extra nuisance parameter σ^2, we estimate (β,Λ) by maximizing an objective function $\frac{1}{n}\sum_{i=1}^n m\left(\beta,\Lambda,\sigma^2;X_i\right)=\mathbb{P}m\left(\beta,\Lambda,\sigma^2;X\right)$. We follow similar notations as those in Huang (1996) and Wellner & Zhang (2007).

Let (β,Λ) be the parameter of our primary interest. Suppose that Λ_η is a parametric path in the monotone nondecreasing function space $\mathcal{F}$ through Λ, i.e. $\Lambda_\eta\in\mathcal{F}$, and $\Lambda_\eta|_{\eta=0}=\Lambda$.

Let $\mathcal{H}=\left\{h: h=\frac{\partial\Lambda_\eta}{\partial\eta}|_{\eta=0}\right\}$ and for any $h\in\mathcal{H}$, we define

$$\begin{aligned}
m_1\left(\beta,\Lambda,\sigma^2;x\right) &= \bigtriangledown_\beta m\left(\beta,\Lambda,\sigma^2;x\right)\\
&\equiv\left(\frac{\partial m\left(\beta,\Lambda,\sigma^2;x\right)}{\partial\beta_1},\cdots,\frac{\partial m\left(\beta,\Lambda,\sigma^2;x\right)}{\partial\beta_d}\right)^T,\\
m_2\left(\beta,\Lambda,\sigma^2;x\right)[h] &= \frac{\partial m\left(\beta,\Lambda_\eta,\sigma^2;x\right)}{\partial\eta}|_{\eta=0},\\
m_{11}\left(\beta,\Lambda,\sigma^2;x\right) &= \bigtriangledown_\beta^2 m\left(\beta,\Lambda,\sigma^2;x\right),\\
m_{12}\left(\beta,\Lambda,\sigma^2;x\right)[h] &= \frac{\partial m_1\left(\beta,\Lambda_\eta,\sigma^2;x\right)}{\partial\eta}|_{\eta=0},\\
m_{21}\left(\beta,\Lambda,\sigma^2;x\right)[h] &= \bigtriangledown_\beta m_2\left(\beta,\Lambda,\sigma^2;x\right)[h],\\
m_{22}\left(\beta,\Lambda,\sigma^2;x\right)[h_1,h_2] &= \frac{\partial^2 m\left(\beta,\Lambda_{\eta_j},\sigma^2;x\right)}{\partial\eta^2}|_{\eta_j=0,j=1,2}\equiv\frac{\partial m_2\left(\beta,\Lambda_{\eta_2},\sigma^2;x\right)[h_1]}{\partial\eta_2}
\end{aligned}$$

To establish the asymptotic distribution of the pseudo-MLE of $\hat{\beta}_n$, we need the following assumptions:

A1: $|\hat{\beta}_n - \beta_0| = o_p(1)$, and $\|\hat{\Lambda}_n - \Lambda_0\| = O_p(n^{-\gamma})$ for some $\gamma > 0$.

A2: $Pm_1(\beta_0, \Lambda_0, \sigma^2; X) = 0$ and $Pm_2(\beta_0, \Lambda_0, \sigma^2; X)[h] = 0, \quad \forall h \in \mathcal{H}, \sigma^2 \in \mathcal{R}^+$. where $\mathcal{R}^+$ is a compact set in the neighborhood of a fixed point σ_0^2 in $\mathbb{R}^+$.

A3: For any $\sigma^2 \in \mathcal{R}^+$, there exists a $h^*_{\sigma^2} = \left(h^*_{1,\sigma^2}, \cdots, h^*_{d,\sigma^2}\right)^T$ such that

$$P\left(m_{12}\left(\beta_0, \Lambda_0, \sigma^2\right)[h] - m_{22}\left(\beta_0, \Lambda_0, \sigma^2\right)[h^*_{\sigma^2}, h]\right) = 0 \quad \forall h \in \mathcal{H}$$

Let

$$A\left(\beta_0, \Lambda_0, \sigma^2\right) = -P\left(m_{11}\left(\beta_0, \Lambda_0, \sigma^2\right) - m_{21}\left(\beta_0, \Lambda_0, \sigma^2\right)[h^*_{\sigma^2}]\right)$$

A4: Estimators $\left(\hat{\beta}_n, \hat{\Lambda}_n\right)$ satisfies

$$\sup_{\sigma^2 \in \mathcal{R}^+} \mathbb{P}_n m_1\left(\hat{\beta}_n, \hat{\Lambda}_n, \sigma^2; X\right) = o_p\left(n^{-1/2}\right); \tag{1}$$

$$\sup_{\sigma^2 \in \mathcal{R}^+} \mathbb{P}_n m_2\left(\hat{\beta}_n, \hat{\Lambda}_n, \sigma^2; X\right)[h^*_{\sigma^2}] = o_p\left(n^{-1/2}\right) \tag{2}$$

A5: For any $\delta_n \downarrow 0$ and $C > 0$,

$$\sup_{|\beta-\beta_0| \le \delta_n, \|\Lambda - \Lambda_0\| \le Cn^{-\gamma}, \sigma^2 \in \mathcal{R}^+} \Big|\sqrt{n}\left(\mathbb{P}_n - \mathbb{P}\right) m_1\left(\beta, \Lambda, \sigma^2\right) - \sqrt{n}\left(\mathbb{P}_n - \mathbb{P}\right) m_1\left(\beta_0, \Lambda_0, \sigma^2\right)\Big| = o_p(1)$$

$$\sup_{|\beta-\beta_0| \le \delta_n, \|\Lambda - \Lambda_0\| \le Cn^{-\gamma}, \sigma^2 \in \mathcal{R}^+} \Big|\sqrt{n}\left(\mathbb{P}_n - \mathbb{P}\right) m_2\left(\beta, \Lambda, \sigma^2\right)[h^*_{\sigma^2}] - \sqrt{n}\left(\mathbb{P}_n - \mathbb{P}\right) m_2\left(\beta_0, \Lambda_0, \sigma^2\right)[h^*_{\sigma^2}]\Big| = o_p(1)$$

A6: For (β, Λ) at the neighborhood of (β_0, Λ_0):

$$\left\{(\beta, \Lambda) : |\beta - \beta_0| \leq \delta_n, \|\Lambda - \Lambda_0\| \leq Cn^{-\gamma}; \alpha > 1, \alpha\gamma > 1/2\right\},$$

$$\sup_{\sigma^2 \in \mathcal{R}^+} \Big| P\left(m_1\left(\beta, \Lambda, \sigma^2\right) - m_1\left(\beta_0, \Lambda_0, \sigma^2\right) - m_{11}\left(\beta_0, \Lambda_0, \sigma^2\right)(\beta - \beta_0) \right.$$
$$\left. - m_{12}\left(\beta_0, \Lambda_0, \sigma^2\right)[\hat{\Lambda}_n - \Lambda_0]\right)\Big| = o\left(|\beta - \beta_0|\right) + O\left(\|\Lambda - \Lambda_0\|^\alpha\right)$$

$$\sup_{\sigma^2 \in \mathcal{R}^+} \Big| P\left(m_2\left(\beta, \Lambda, \sigma^2\right)[h^*_{\sigma^2}] - m_2\left(\beta_0, \Lambda_0, \sigma^2\right)[h^*_{\sigma^2}] \right.$$
$$\left. - m_{21}\left(\beta_0, \Lambda_0, \sigma^2\right)[h^*_{\sigma^2}](\beta - \beta_0) - m_{22}\left(\beta_0, \Lambda_0, \sigma^2\right)[h^*_{\sigma^2}, \hat{\Lambda}_n - \Lambda_0]\right)\Big|$$
$$= o\left(|\beta - \beta_0|\right) + O\left(\|\Lambda - \Lambda_0\|^\alpha\right)$$

A7: There exist $m_{1\sigma}$ and $m_{2\sigma}$ such that for any $\sigma_1^2, \sigma_2^2 \in \mathcal{R}^+$,

$$|m_1\left(\beta_0, \Lambda_0, \sigma_1^2\right) - m_1\left(\beta_0, \Lambda_0, \sigma_2^2\right)| \leq m_{1\sigma}\left|\sigma_1^2 - \sigma_2^2\right|$$

$$|m_2\left(\beta_0, \Lambda_0, \sigma_1^2\right)[h^*_{\sigma_1}] - m_2\left(\beta_0, \Lambda_0, \sigma_2^2\right)[h^*_{\sigma_2}]| \leq m_{2\sigma}\left|\sigma_1^2 - \sigma_2^2\right|$$

and

$$\left\{Pm_{1\sigma}^4(X)\right\}^{1/4} < \infty; \quad \left\{Pm_{2\sigma}^4(X)\right\}^{1/4} < \infty$$

Theorem 3.11. *(Asymptotic Normality of the regression parameter with extra nuisance parameter σ^2). Suppose that Assumptions A1-A6 hold. The nuisance parameter σ^2 is replaced by its estimate $\hat{\sigma}_n^2$. Then*

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) = -A^{-1}\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right)\mathbb{G}_n\left(m_1\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right) - m_2\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right)[h^*_{\hat{\sigma}_n^2}]\right) + o_p(1)$$

Furthermore, if $\hat{\sigma}_n^2 \longrightarrow_p \sigma_0^2$, the above asymptotic expansion and condition A7 lead to

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) = -A_0^{-1}\mathbb{G}_n\left(m_1\left(\beta_0, \Lambda_0, \sigma_0^2\right) - m_2\left(\beta_0, \Lambda_0, \sigma_0^2\right)[h^*_{\sigma_0^2}]\right) + o_p(1)$$
$$\longrightarrow_d N\left(0, A_0^{-1}B_0A_0^{-1}\right)$$

where

$$\begin{aligned} A_0 &= A\left(\beta_0, \Lambda_0, \sigma_0^2\right) = -P\left(m_{11}\left(\beta_0, \Lambda_0, \sigma_0^2\right) - m_{21}\left(\beta_0, \Lambda_0, \sigma_0^2\right)[h^*_{\sigma_0^2}]\right) \\ B_0 &= B\left(\beta_0, \Lambda_0, \sigma_0^2\right) = P\left(m_1\left(\beta_0, \Lambda_0, \sigma_0^2\right) - m_2\left(\beta_0, \Lambda_0, \sigma_0^2\right)[h^*_{\sigma_0^2}]\right)^{\otimes 2} \end{aligned}$$

Proof. By Condition A1 and A5

$$\begin{cases} \sqrt{n}\left(\mathbb{P}_n - P\right)\left(m_1\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2\right) - m_1\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right)\right) = o_p(1) \\ \sqrt{n}\left(\mathbb{P}_n - P\right)\left(m_2\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2\right)[h^*_{\hat{\sigma}_n^2}] - m_2\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right)\right)[h^*_{\hat{\sigma}_n^2}] = o_p(1) \end{cases}$$

Together with Condition A2 and A4, this implies

$$\begin{cases} \sqrt{n}Pm_1\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2\right) + \sqrt{n}\mathbb{P}_n m_1\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right) = o_p(1) \\ \sqrt{n}Pm_2\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2\right)[h^*_{\hat{\sigma}_n^2}] + \sqrt{n}\mathbb{P}_n m_2\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right)[h^*_{\hat{\sigma}_n^2}] = o_p(1) \end{cases}$$

So by condition A6,

$$\begin{cases} P\left\{m_{11}\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right)\left(\hat{\beta}_n - \beta_0\right) + m_{12}\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right)[\hat{\Lambda}_n - \Lambda_0]\right\} \\ \qquad + o\left(|\hat{\beta}_n - \beta_0|\right) + O\left(\|\Lambda - \Lambda_0\|^{\alpha}\right) = -\mathbb{P}_n m_1\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right) + o_p\left(n^{-1/2}\right) \\ P\left\{m_{21}\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right)[h^*_{\hat{\sigma}_n^2}]\left(\hat{\beta}_n - \beta_0\right) + m_{22}\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right)[h^*_{\hat{\sigma}_n^2}, \hat{\Lambda}_n - \Lambda_0]\right\} \\ \qquad + o\left(|\hat{\beta}_n - \beta_0|\right) + O\left(\|\Lambda - \Lambda_0\|^{\alpha}\right) = -\mathbb{P}_n m_2\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right)[h^*_{\hat{\sigma}_n^2}] + o_p\left(n^{-1/2}\right) \end{cases}$$

By condition A1, $|\hat{\beta}_n - \beta_0| = o_p(1)$, $\|\hat{\Lambda}_n - \Lambda\|^{\alpha} = O_p\left(n^{-\alpha\gamma}\right)$ and $\alpha\gamma > 1/2$, so

$$\begin{cases} \sqrt{n}\left(Pm_{11}\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right) + o(1)\right)\left(\hat{\beta}_n - \beta_0\right) + \\ \quad \sqrt{n}Pm_{12}\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right)[\hat{\Lambda}_n - \Lambda_0] = -\sqrt{n}\mathbb{P}_n m_1\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right) + o_p(1) \qquad (1) \\ \sqrt{n}\left(Pm_{21}\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right)[h^*_{\hat{\sigma}_n^2}] + o(1)\right)\left(\hat{\beta}_n - \beta_0\right) + \\ \quad \sqrt{n}Pm_{22}\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right)[h^*_{\hat{\sigma}_n^2}, \hat{\Lambda}_n - \Lambda_0] = -\sqrt{n}\mathbb{P}_n m_2\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right)[h^*_{\hat{\sigma}_n^2}] + o_p(1) \qquad (2) \end{cases}$$

Using (1)-(2), we have

$$\begin{aligned}
&\sqrt{n}P\left(m_{11}\left(\beta_0,\Lambda_0,\hat{\sigma}_n^2\right)-m_{21}\left(\beta_0,\Lambda_0,\hat{\sigma}_n^2\right)[h^*_{\hat{\sigma}_n^2}]+o\left(1\right)\right)\left(\hat{\beta}_n-\beta_0\right)\\
&\qquad+\sqrt{n}P\left(m_{12}\left(\beta_0,\Lambda_0,\hat{\sigma}_n^2\right)[\hat{\Lambda}_n-\Lambda_0]-m_{22}\left(\beta_0,\Lambda_0,\hat{\sigma}_n^2\right)[h^*_{\hat{\sigma}_n^2},\hat{\Lambda}_n-\Lambda_0]\right)\\
&=-\sqrt{n}\mathbb{P}_n\left(m_1\left(\beta_0,\Lambda_0,\hat{\sigma}_n^2\right)-m_2\left(\beta_0,\Lambda_0,\hat{\sigma}_n^2\right)[h^*_{\hat{\sigma}_n^2}]\right)+o_p(1)
\end{aligned}$$

By condition A3,

$$P\left(m_{12}\left(\beta_0,\Lambda_0,\hat{\sigma}_n^2\right)[\hat{\Lambda}_n-\Lambda_0]-m_{22}\left(\beta_0,\Lambda_0,\hat{\sigma}_n^2\right)[h^*_{\hat{\sigma}_n^2},\hat{\Lambda}_n-\Lambda_0]\right)=0$$

So

$$\begin{aligned}
&\sqrt{n}P\left(m_{11}\left(\beta_0,\Lambda_0,\hat{\sigma}_n^2\right)-m_{21}\left(\beta_0,\Lambda_0,\hat{\sigma}_n^2\right)[h^*_{\hat{\sigma}_n^2}]+o\left(1\right)\right)\left(\hat{\beta}_n-\beta_0\right)\\
&=-\sqrt{n}\mathbb{P}_n\left(m_1\left(\beta_0,\Lambda_0,\hat{\sigma}_n^2\right)-m_2\left(\beta_0,\Lambda_0,\hat{\sigma}_n^2\right)[h^*_{\hat{\sigma}_n^2}]\right)+o_p(1)
\end{aligned}$$

And

$$\begin{aligned}
&\sqrt{n}\left(\hat{\beta}_n-\beta_0\right)\\
&=-A^{-1}\left(\beta_0,\Lambda_0,\hat{\sigma}_n^2\right)\sqrt{n}\mathbb{P}_n\left(m_1\left(\beta_0,\Lambda_0,\hat{\sigma}_n^2\right)-m_2\left(\beta_0,\Lambda_0,\hat{\sigma}_n^2\right)[h^*_{\hat{\sigma}_n^2}]\right)+o_p\left(1\right)\\
&=-A^{-1}\left(\beta_0,\Lambda_0,\hat{\sigma}_n^2\right)\mathbb{G}_n\left(m_1\left(\beta_0,\Lambda_0,\hat{\sigma}_n^2\right)-m_2\left(\beta_0,\Lambda_0,\hat{\sigma}_n^2\right)[h^*_{\hat{\sigma}_n^2}]\right)+o_p\left(1\right)
\end{aligned}$$

(This is true by condition A2)

Furthermore, we can rewrite the above expansion as

$$\begin{aligned}
\sqrt{n}\left(\hat{\beta}_n-\beta_0\right)=&A^{-1}\left(\beta_0,\Lambda_0,\hat{\sigma}_n^2\right)\Big\{\mathbb{G}_n\left(m_1\left(\beta_0,\Lambda_0,\sigma_0^2\right)-m_2\left(\beta_0,\Lambda_0,\sigma_0^2\right)[h^*_{\sigma_0^2}]\right)\\
&+\mathbb{G}_n\left(m_1\left(\beta_0,\Lambda_0,\hat{\sigma}_n^2\right)-m_1\left(\beta_0,\Lambda_0,\sigma_0^2\right)\right)\\
&-\mathbb{G}_n\left(m_2\left(\beta_0,\Lambda_0,\hat{\sigma}_n^2\right)[h^*_{\hat{\sigma}_n^2}]-m_2\left(\beta_0,\Lambda_0,\sigma_0^2\right)[h^*_{\sigma_0^2}]\right)\Big\}+o_p\left(1\right)
\end{aligned}$$

By the consistency, condition A7 and the dominate convergence theorem it is easily seen that

$$A\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right) \to_p A\left(\beta_0, \Lambda_0, \sigma_0^2\right)$$

Define two classes,

$$\begin{aligned}
\tilde{\mathcal{M}}_1 &= \left\{m_1\left(\beta_0, \Lambda_0, \sigma^2\right) - m_1\left(\beta_0, \Lambda_0, \sigma_0^2\right) : |\sigma^2 - \sigma_0^2| \leq \eta\right\} \\
\tilde{\mathcal{M}}_2 &= \Big\{m_2\left(\beta_0, \Lambda_0, \sigma^2\right)[h^*_{\sigma^2}] - m_2\left(\beta_0, \Lambda_0, \sigma_0^2\right)[h^*_{\sigma_0^2}] : |\sigma^2 - \sigma_0^2| \leq \eta\Big\}
\end{aligned}$$

With the compactness of $\mathcal{R}^+$, similar to the proof of convergence, we can construct an ε-net, $\left\{\sigma_1^2, \sigma_2^2, \cdots, \sigma_q^2\right\}, q = O(1/\varepsilon)$ over $\mathcal{R}^+$. By the Lipschitz condition specified in Condition A7, both $\tilde{\mathcal{M}}_1$ and $\tilde{\mathcal{M}}_2$ are indexed by σ^2, so their bracket numbers are both $O(1/\varepsilon)$ and hence both $\tilde{\mathcal{M}}_1$ and $\tilde{\mathcal{M}}_2$ are P-Donsker. And by condition A7

$$\begin{aligned}
&\Big(P\left(m_1\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right) - m_1\left(\beta_0, \Lambda_0, \sigma_0^2\right)\right)^2\Big)^{1/2} \\
\leq &\Big(P m_{1\sigma}^2 \left|\hat{\sigma}_n^2 - \sigma_0^2\right|^2\Big)^{1/2} \qquad \text{(by condition A7)} \\
\leq &\left(P m_{1\sigma}^4\right)^{1/4} \Big(P\left(\hat{\sigma}_n^2 - \sigma_0^2\right)^4\Big)^{1/4} \longrightarrow 0 \qquad \text{(by Hölder's inequality)}
\end{aligned}$$

Similarly

$$\left(P\Big(m_2\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right)[h^*_{\hat{\sigma}_n^2}] - m_2\left(\beta_0, \Lambda_0, \sigma_0^2\right)[h^*_{\sigma_0^2}]\Big)^2\right)^{1/2} \longrightarrow 0$$

By the semi-equicontinuity of Theorem 2.8.2 in van der Vaart & Wellner (1996) (Lemma 3.5), this implies that both classes are Donsker class. Therefore,

$$\begin{aligned}
&\mathbb{G}_n\left(m_1\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right) - m_1\left(\beta_0, \Lambda_0, \sigma_0^2\right)\right) = o_p(1) \\
&\mathbb{G}_n\Big(m_2\left(\beta_0, \Lambda_0, \hat{\sigma}_n^2\right)[h^*_{\hat{\sigma}_n^2}] - m_1\left(\beta_0, \Lambda_0, \sigma_0^2\right)[h^*_{\sigma_0^2}]\Big) = o_p(1)
\end{aligned}$$

So

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}_n-\beta_0\right)=&A_0^{-1}\mathbb{G}_n\left(m_1\left(\beta_0,\Lambda_0,\sigma_0^2\right)-m_2\left(\beta_0,\Lambda_0,\sigma_0^2\right)\left[h^*_{\sigma_0^2}\right]\right)+o_p\left(1\right)\\ &\longrightarrow_d N\left(0,A_0^{-1}B_0A_0^{-1}\right)\end{aligned}$$

□

3.3 Asymptotic properties of the estimates based on Gamma-Frailty Poisson Model

In this section, we first provide the regularity conditions and some preliminary results that are used in the proof of the asymptotic properties of our estimator and then apply the theorems in Section 3.2 to the Gamma-Frailty Poisson model.

3.3.1 Preliminary Results

The following regularity conditions are sufficient to guarantee the asymptotic properties, including their consistency, convergence rate and asymptotic normality of the regression parameter, of the spline-based sieve GEE estimate with $V_3^{(i)}$ as a covariance matrix.

Condition 1. The true parameter $(\beta_0,\Lambda_0,\sigma_0^2)\in\overset{\circ}{\mathcal{R}^d}\times\mathcal{F}\times\overset{\circ}{\mathcal{R}^+}$, where $\overset{\circ}{\mathcal{R}^d}$ and $\overset{\circ}{\mathcal{R}^+}$ are the interior of some compact set of $\mathcal{R}^d$ and $\mathcal{R}^+$ in $\mathbb{R}^d$ and $\mathbb{R}^+$, respectively. $\mathcal{F}$ is the monotone nondecreasing function space.

Condition 2. The observation time $T_{K,j}:j=1,2,\cdots,K, K=1,2,\cdots$ are bounded in interval $[0,\tau]$ for some $\tau\in(0,\infty)$ and $P\left(T_{K,j}-T_{K,j-1}\geq s_0\right)=1$ for some constant s_0. $P\left(K\leq k_0\right)=1$ for some constant k_0.

Condition 3. The true baseline mean function Λ_0 is p^{th} differentiable and bounded. The derivative has a positive and finite lower and upper bounds in the observation interval $[0, \tau]$.

Condition 4. For some $\eta \in (0,1), a^T Var(Z|U,V)a \geq \eta a^T E(ZZ^T|U,V)a$ a.s. for all $a \in \mathbb{R}^d$ where (U,V,Z) follows distribution $\mu/\mu(\mathbb{R}^{+2} \times \mathcal{Z})$.

Condition 5. The covariate Z is bounded, i.e., $P(|Z| \leq z_0) = 1$ for some constant z_0. And $P(aZ \neq c) > 0$ for all compatible vectors a and c.

Condition 6. $E\left\{e^{C\mathbb{N}(t)}\right\}$ is uniformly bounded for $t \in S[T] = \{t : 0 < t < \tau\}$ for some $\tau > 0$.

Condition 7. The number of knots $q_n = O(n^{\nu})$ for $\frac{1}{2p+1} < \nu < \frac{1}{2}$.

Given the frailty parameter, σ^2, the log likelihood of Gamma-Frailty Poisson process in Equation (2.7) can be rewritten as

$$l\left(\beta, \Lambda, \sigma^2; X_i\right) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{K_i} \Delta\mathbb{N}_{K_i,j}^{(i)} log\left(\Delta\Lambda_{K_i,j}^{(i)} e^{\beta^T Z_i}\right) - \right.$$
$$\left. \left(\mathbb{N}_{K_i,K_i}^{(i)} + 1/\sigma^2\right) log\left(\Lambda_{K_i,K_i}^{(i)} e^{\beta^T Z_i} + 1/\sigma^2\right) \right\}$$

up to a constant. Let $\mathbb{M}(\beta, \Lambda, \sigma^2) = Pm_{\beta,\Lambda,\sigma^2}(X)$ and $\mathbb{M}_n(\beta, \Lambda, \sigma^2) = \mathbb{P}_n m_{\beta,\Lambda,\sigma^2}(X)$, where

$$m_{\beta,\Lambda,\sigma^2}(X) = \sum_{j=1}^{K} \Delta\mathbb{N}_j log\left(\Delta\Lambda_j e^{\beta^T Z}\right) - \left(\mathbb{N}_K + 1/\sigma^2\right) log\left(\Lambda_K e^{\beta^T Z} + 1/\sigma^2\right) \quad (3.9)$$

We define probability measures μ, γ and the corresponding metrics d and d_K in a similar manner as those used to study the asymptotic property of the maximum likelihood estimators in Wellner & Zhang (2007), e.g.,

$$\begin{aligned}
&\mu\left(B_1 \times B_2\right) \\
&= \int_{\mathbb{R}^d} \sum_{k=1}^{\infty} P\left(K=k|Z=z\right) \sum_{j=1}^{k} P\left(T_{k,j-1} \in B_1, T_{k,j} \in B_2 | K=k, Z=z\right) dH\left(z\right) \\
&\gamma\left(B\right) = \int_{\mathbb{R}^d} \sum_{k=1}^{\infty} P\left(K=k|Z=z\right) P\left(T_{k,k} \in B | K=k, Z=z\right) dH\left(z\right)
\end{aligned}$$

Based on the measure μ and γ, define the metrics

$$\begin{aligned}
d\left(\theta_1, \theta_2\right) &= \left\{|\beta_1 - \beta_2|^2 + \|\Lambda_1 - \Lambda_2\|^2_{L_2(\mu)}\right\}^{1/2} \\
&= \left\{|\beta_1 - \beta_2|^2 + \int \left(\left(\Lambda_1\left(u\right) - \Lambda_1\left(v\right)\right) - \left(\Lambda_2\left(u\right) - \Lambda_2\left(v\right)\right)\right)^2 d\mu\left(u,v\right)\right\}^{1/2} \\
d_K\left(\theta_1, \theta_2\right) &= \left\{|\beta_1 - \beta_2|^2 + \|\Lambda_1 - \Lambda_2\|^2_{L_2(\gamma)}\right\}^{1/2} \\
&= \left\{|\beta_1 - \beta_2|^2 + \int \left(\Lambda_1\left(u\right) - \Lambda_2\left(u\right)\right)^2 d\gamma\left(u\right)\right\}^{1/2}
\end{aligned}$$

Lemma 3.12. *Suppose Conditions 1, 3-5 hold, then*

(i) $\mathbb{M}\left(\beta_0, \Lambda_0, \sigma^2\right) \geq \mathbb{M}\left(\beta, \Lambda, \sigma^2\right)$ *for any* $(\beta, \Lambda) \in \mathcal{R}^d \times \mathcal{F}, \sigma^2 \in \mathcal{R}^+$ *and the equality hold iff* $\beta = \beta_0$ *and* $\Lambda = \Lambda_0$ *a.e with respect to* μ.

(ii) There exists a constant C, such that

$$\mathbb{M}\left(\beta_0, \Lambda_0, \sigma^2\right) - \mathbb{M}\left(\beta, \Lambda, \sigma^2\right) \geq C d^2\left(\left(\beta_0, \Lambda_0\right), \left(\beta, \Lambda\right)\right)$$

for any (β, Λ) *in a neighborhood of* (β_0, Λ_0) *and* $\sigma^2 \in \mathcal{R}^+$.

Proof. First, we prove the uniqueness of the maximum.

$$
\begin{aligned}
&\mathbb{M}\left(\beta_0, \Lambda_0, \sigma^2\right) - \mathbb{M}\left(\beta, \Lambda, \sigma^2\right) \\
&= P\left(\sum_{j=1}^{K}\left(\triangle\Lambda_{0,j}e^{\beta_0^T Z}log\frac{\triangle\Lambda_{0,j}e^{\beta_0^T Z}}{\triangle\Lambda_j e^{\beta^T Z}}\right) - \left(\Lambda_{0,K}e^{\beta_0^T Z} + 1/\sigma^2\right)log\frac{\Lambda_{0,K}e^{\beta_0^T Z} + 1/\sigma^2}{\Lambda_K e^{\beta^T Z} + 1/\sigma^2}\right) \\
&= P\left\{\sum_{j=1}^{K}\left(\triangle\Lambda_{0,j}e^{\beta_0^T Z}log\frac{\triangle\Lambda_{0,j}e^{\beta_0^T Z}}{\triangle\Lambda_j e^{\beta^T Z}}\right) - \left(\Lambda_{0,K}e^{\beta_0^T Z}\right)log\frac{\Lambda_{0,K}e^{\beta_0^T Z}}{\Lambda_K e^{\beta^T Z}}\right\} \\
&\quad + P\left\{\left(\Lambda_{0,K}e^{\beta_0^T Z}\right)log\frac{\Lambda_{0,K}e^{\beta_0^T Z}}{\Lambda_K e^{\beta^T Z}} - \left(\Lambda_{0,K}e^{\beta_0^T Z} + 1/\sigma^2\right)log\frac{\Lambda_{0,K}e^{\beta_0^T Z} + 1/\sigma^2}{\Lambda_K e^{\beta^T Z} + 1/\sigma^2}\right\} \\
&= PI_1 + PI_2
\end{aligned}
$$

$$
\begin{aligned}
I_1 &= \sum_{j=1}^{K}\left(\triangle\Lambda_{0,j}e^{\beta_0^T Z}log\frac{\triangle\Lambda_{0,j}e^{\beta_0^T Z}}{\triangle\Lambda_j e^{\beta^T Z}}\right) - \left(\Lambda_{0,K}e^{\beta_0^T Z}\right)log\frac{\Lambda_{0,K}e^{\beta_0^T Z}}{\Lambda_K e^{\beta^T Z}} \\
&= \sum_{j=1}^{K}\left(\triangle\Lambda_{0,j}e^{\beta_0^T Z}\left(log\frac{\triangle\Lambda_{0,j}e^{\beta_0^T Z}}{\triangle\Lambda_j e^{\beta^T Z}} - log\frac{\Lambda_{0,K}e^{\beta_0^T Z}}{\Lambda_K e^{\beta^T Z}}\right)\right) \\
&= \sum_{j=1}^{K}\left(\triangle\Lambda_{0,j}e^{\beta_0^T Z}log\frac{\triangle\Lambda_{0,j}/\Lambda_{0,K}}{\triangle\Lambda_j/\Lambda_K}\right) \\
&= \Lambda_{0,K}e^{\beta_0^T Z}\sum_{j=1}^{K}\left(\frac{\triangle\Lambda_{0,j}}{\Lambda_{0,K}}log\frac{\triangle\Lambda_{0,j}/\Lambda_{0,K}}{\triangle\Lambda_j/\Lambda_K}\right)
\end{aligned}
$$

$\sum_{j=1}^{K}\left(\frac{\triangle\Lambda_{0,j}}{\Lambda_{0,K}}log\frac{\triangle\Lambda_{0,j}/\Lambda_{0,K}}{\triangle\Lambda_j/\Lambda_K}\right)$ is the Kullback-Leibler's information $K_{p_0}\left(p_0, p\right)$ with $p_{0,j} = \frac{\triangle\Lambda_{0,j}}{\Lambda_{0,K}}$ and $p_j = \frac{\triangle\Lambda_j}{\Lambda_K}$ for $j = 1, 2, \cdots, K$. So, it is nonnegative and the equality hold when $\frac{\triangle\Lambda_{0,j}}{\Lambda_{0,K}} = \frac{\triangle\Lambda_j}{\Lambda_K}, j = 1, 2, \cdots, K$. Therefore, $PI_1 \geq 0$ and $PI_1 = 0$ iff

$$\Lambda = C\Lambda_0 \text{ a.e. w.r.t } \mu. \text{ for some constant } C \tag{3.10}$$

$$I_2 = \left(\Lambda_{0,K}e^{\beta_0^T Z}\right)log\frac{\Lambda_{0,K}e^{\beta_0^T Z}}{\Lambda_K e^{\beta^T Z}} - \left(\Lambda_{0,K}e^{\beta_0^T Z} + 1/\sigma^2\right)log\frac{\Lambda_{0,K}e^{\beta_0^T Z} + 1/\sigma^2}{\Lambda_K e^{\beta^T Z} + 1/\sigma^2}$$

For the simplicity, denote $x = \Lambda_{0,K}e^{\beta_0^T Z} > 0, b = \Lambda_K e^{\beta^T Z} - \Lambda_{0,K}e^{\beta_0^T Z}$, So

$$I_2 = xlog\frac{x}{x+b} - \left(x + 1/\sigma^2\right)log\frac{x + 1/\sigma^2}{x + 1/\sigma^2 + b}, \quad x > 0, x + b > 0$$

Let $f\left(b\right)=xlog\frac{x}{x+b}-\left(x+1/\sigma^{2}\right)log\frac{x+1/\sigma^{2}}{x+1/\sigma^{2}+b}$, then

$$\frac{\partial}{\partial b}f\left(b\right)=-\frac{x}{x+b}+\frac{x+1/\sigma^{2}}{x+1/\sigma^{2}+b}=\frac{b\times1/\sigma^{2}}{\left(x+b\right)\left(x+1/\sigma^{2}+b\right)}$$

This equals to zero only when $b=0$ and

$$\frac{\partial^{2}}{\partial b^{2}}f\left(b\right)=\frac{1/\sigma^{2}\left(x\left(x+1/\sigma^{2}\right)-b^{2}\right)}{\left(x+b\right)^{2}\left(x+1/\sigma^{2}+b\right)^{2}}$$

When $-x<b<\sqrt{x\left(x+1/\sigma^{2}\right)}$, $\frac{\partial^{2}}{\partial b^{2}}f\left(b\right)>0$ and when $b>\sqrt{x\left(x+1/\sigma^{2}\right)}$, $\frac{\partial^{2}}{\partial b^{2}}f\left(x\right)<0$. Thus $f\left(b\right)$ reaches its minimum at $b=0$ and $f\left(0\right)=0$. So $PI_{2}\geq0$ and the equality hold when

$$\Lambda e^{\beta^{T}Z}=\Lambda_{0}e^{\beta_{0}^{T}Z}\text{ a.e. w.r.t. }\gamma.$$

By the argument given by Wellner & Zhang (2007), this implies that

$$\beta=\beta_{0}\text{ and }\Lambda=\Lambda_{0}\text{ a.e. w.r.t. }\gamma$$

and furthermore by Equation (3.10), it implies that

$$\beta=\beta_{0}\text{ and }\Lambda=\Lambda_{0}\text{ a.e. w.r.t }\mu$$

Now we prove the second part of the lemma. I_{1} can be rewritten as following,

$$\begin{aligned}
I_{1}&=\Lambda_{0,K}e^{\beta_{0}^{T}Z}\sum_{j=1}^{K}\left(\frac{\triangle\Lambda_{0,j}}{\Lambda_{0,K}}log\frac{\triangle\Lambda_{0,j}/\Lambda_{0,K}}{\triangle\Lambda_{j}/\Lambda_{K}}\right)\\
&=\Lambda_{0,K}e^{\beta_{0}^{T}Z}\sum_{j=1}^{K}\left[\frac{\triangle\Lambda_{j}}{\Lambda_{K}}\left(\frac{\triangle\Lambda_{0,j}/\Lambda_{0,K}}{\triangle\Lambda_{j}/\Lambda_{K}}log\frac{\triangle\Lambda_{0,j}/\Lambda_{0,K}}{\triangle\Lambda_{j}/\Lambda_{K}}-\frac{\triangle\Lambda_{0,j}/\Lambda_{0,K}}{\triangle\Lambda_{j}/\Lambda_{K}}+1\right)\right]\\
&\geq\frac{1}{4}\Lambda_{0,K}e^{\beta_{0}^{T}Z}\sum_{j=1}^{K}\frac{\triangle\Lambda_{j}}{\Lambda_{K}}\left(\frac{\triangle\Lambda_{0,j}/\Lambda_{0,K}}{\triangle\Lambda_{j}/\Lambda_{K}}-1\right)^{2}\\
&=\frac{1}{4}\Lambda_{0,K}e^{\beta_{0}^{T}Z}\sum_{j=1}^{K}\frac{1}{\triangle\Lambda_{j}/\Lambda_{K}}\left(\frac{\triangle\Lambda_{0,j}}{\Lambda_{0,K}}-\frac{\triangle\Lambda_{j}}{\Lambda_{K}}\right)^{2}\\
&\geq\frac{1}{4}\Lambda_{0,K}e^{\beta_{0}^{T}Z}\sum_{j=1}^{K}\left(\frac{\triangle\Lambda_{0,j}}{\Lambda_{0,K}}-\frac{\triangle\Lambda_{j}}{\Lambda_{K}}\right)^{2}
\end{aligned}$$

The first inequality is due to the fact that $xlogx - x + 1 \geq \frac{1}{4}(x-1)^2$ for x in a neighborhood of $x = 1$, the equality hold only when $x = 1$.

I_2 can be expanded by Taylor expansion as

$$\begin{aligned}I_2 = f(b) &= f(0) + f'(0)b + \frac{1}{2}f^{''}(\xi)b^2\\ &= \frac{1}{2}f^{''}(\xi)b^2 = \frac{1/\sigma^2\left[x\left(x+1/\sigma^2\right)-\xi^2\right]}{2\left(x+\xi\right)^2\left(x+1/\sigma^2+\xi\right)^2}b^2 \text{ where } |\xi| < |b|\end{aligned}$$

When b is at the neighborhood of zero, e.g. $|b| < |x|$ at almost everywhere in ϕ, the numerator

$$1/\sigma^2\left[x\left(x+1/\sigma^2\right)-\xi^2\right] \geq 1/\sigma^2\left[x\left(x+1/\sigma^2\right)-x^2\right] = \left(1/\sigma^2\right)^2 x;$$

And the denominator

$$2\left(x+\xi\right)^2\left(x+1/\sigma^2+\xi\right)^2 \leq 2(2x)^2(x+1/\sigma^2+x)^2 = 8x^2(2x+1/\sigma^2)^2.$$

Therefore $f(b) \geq \frac{\left(1/\sigma^2\right)^2 x}{8x^2(2x+1/\sigma^2)^2}b^2$. And

$$I_2 = f(b) \geq \frac{\left(1/\sigma^2\right)^2}{8\Lambda_{0,K}e^{\beta_0^T Z}\left(2\Lambda_{0,K}e^{\beta_0^T Z}+1/\sigma^2\right)^2}\left(\Lambda_{0K}e^{\beta_0^T Z} - \Lambda_K e^{\beta^T Z}\right)^2$$

Combine the results from I_1 and I_2, we have,

$$\begin{aligned}&I_1 + I_2\\ \geq&\frac{1}{4}\Lambda_{0,K}e^{\beta_0^T Z}\sum_{j=1}^{K}\left(\frac{\triangle\Lambda_{0,j}}{\Lambda_{0,K}} - \frac{\triangle\Lambda_j}{\Lambda_K}\right)^2 + \frac{\left(1/\sigma^2\right)^2}{8\Lambda_{0,K}e^{\beta_0^T Z}\left(2\Lambda_{0,K}e^{\beta_0^T Z}+1/\sigma^2\right)^2}\times\\ &\qquad\left(\Lambda_{0K}e^{\beta_0^T Z} - \Lambda_K e^{\beta^T Z}\right)^2\\ =&\frac{1}{4}\Lambda_{0,K}e^{\beta_0^T Z} \times \frac{1}{k^2}\sum_{j=1}^{K}\left[k^2\left(\theta_{j1}-\theta_{j2}\right)^2 + \left(l_1 - l_2\right)^2\right]\end{aligned}$$

Where we denote $\theta_{j1} = \frac{\triangle\Lambda_{0,j}}{\Lambda_{0,K}}, \theta_{j2} = \frac{\triangle\Lambda_j}{\Lambda_K}, l_1 = \Lambda_{0,K}e^{\beta_0^T Z}, l_2 = \Lambda_K e^{\beta^T Z}$ and $k = \frac{\sqrt{2K}\Lambda_{0,K}e^{\beta_0^T Z}\left(2\Lambda_{0,K}e^{\beta_0^T Z}+1/\sigma^2\right)}{1/\sigma^2}$.

When $l_1 = l_2$, $I_1 + I_2 \geq \frac{1}{4}\Lambda_{0,K}e^{\beta_0^T Z} \times \sum_{j=1}^{K}\left(\theta_{j1} - \theta_{j2}\right)^2$. Therefore $P\left(I_1 + I_2\right) \geq CP\sum_{j=1}^{K}\left(\triangle\Lambda_{0,j}e^{\beta_0^T Z} - \Delta\Lambda_j e^{\beta^T Z}\right)^2$. We now show that this inequality is also true when $l_1 \neq l_2$. We claim that for $C = \frac{1}{2} \wedge \frac{k^2}{(l_1 \wedge l_2)^2}$, we have

$$k^2\left(\theta_1 - \theta_2\right)^2 + \left(l_1 - l_2\right)^2 \geq C\left(l_2\theta_2 - l_1\theta_1\right)^2 \quad \forall 0 \leq \theta_1 \leq 1, 0 \leq \theta_2 \leq 1, l_1 \geq \gamma_1, l_2 \geq \gamma_2$$

for some $\gamma_1 > 0$ and $\gamma_2 > 0$. First we discuss the case when l_1, l_2 and θ_1, θ_2 are concordant, e.g. $\left(l_1 - l_2\right)\left(\theta_1 - \theta_2\right) \geq 0$. Without a lost of generality, we assume $l_1 > l_2$ and $\theta_1 \geq \theta_2$.

$$\begin{aligned} k^2\left(\theta_1 - \theta_2\right)^2 + \left(l_1 - l_2\right)^2 &\geq \frac{1}{2}\left(k\left(\theta_1 - \theta_2\right) + \left(l_1 - l_2\right)\right)^2 \\ &\geq \frac{1}{2}\left(k\left(\theta_1 - \theta_2\right) + \left(l_1 - l_2\right)\theta_1\right)^2 \\ &= \frac{1}{2}\left(l_1\theta_1 - l_2\theta_2 + \left(k - l_2\right)\left(\theta_1 - \theta_2\right)\right)^2 \qquad (*) \end{aligned}$$

Since

$$\begin{aligned} k = &\frac{\sqrt{2K}\Lambda_{0,K}e^{\beta_0^T Z}\left(2\Lambda_{0,K}e^{\beta_0^T Z} + 1/\sigma^2\right)}{1/\sigma^2} \geq \sqrt{2K}\Lambda_{0,K}e^{\beta_0^T Z} \geq \Lambda_{0,K}e^{\beta_0^T Z} \\ &\geq min\left(\Lambda_{0,K}e^{\beta_0^T Z}, \Lambda_K e^{\beta^T Z}\right) = l_2. \end{aligned}$$

By (*), $k^2\left(\theta_1 - \theta_2\right)^2 + \left(l_1 - l_2\right)^2 \geq \frac{1}{2}\left(l_1\theta_1 - l_2\theta_2\right)^2$.

For discordant pair, say, $l_1 < l_2, \theta_1 \geq \theta_2$, we further discuss the claim in two cases:

(i) When $l_1\theta_1 \geq l_2\theta_2$ we have

$$\theta_1 - \theta_2 = \frac{1}{l_1}\left(l_1\theta_1 - l_1\theta_2\right) > \frac{1}{l_1}\left(l_1\theta_1 - l_2\theta_2\right) \geq 0$$

So $(\theta_1-\theta_2)^2 > \frac{1}{l_1^2}(l_1\theta_1-l_2\theta_2)^2$.

(ii) When $l_1\theta_1 < l_2\theta_2$ we have

$$l_2-l_1 \geq l_2\theta_2-l_1\theta_2 \geq l_2\theta_2-l_1\theta_1 > 0$$

So $(l_2-l_1)^2 > (l_1\theta_1-l_2\theta_2)^2$.

Therefore, $k^2(\theta_1-\theta_2)^2+(l_1-l_2)^2 \geq C(l_1\theta_1-l_2\theta_2)^2$ where $C=\frac{1}{2}\wedge\frac{k^2}{(l_1\wedge l_2)^2}$.

So,

$$\begin{aligned}P(I_1+I_2) \geq & P\left\{\frac{1}{4}\Lambda_{0,K}e^{\beta_0^TZ}\times\left(\frac{1}{2k^2}\wedge\frac{1}{(\Lambda_{0,K}e^{\beta_0^TZ}\wedge\Lambda_Ke^{\beta^TZ})^2}\right)\sum_{j=1}^K(l_2\theta_{j2}-l_1\theta_{j1})^2\right\}\\ & (k \text{ is specified as before. })\\ \geq & CP\sum_{j=1}^K\left(\triangle\Lambda_{0,j}e^{\beta_0^TZ}-\Delta\Lambda_je^{\beta^TZ}\right)^2\end{aligned}$$

The last inequality is due to the compactness of the parameter space of β,Λ and the boundness of the covariates (Z,K,T) specified in conditions 1,2 and 5.

Following the same proof as in Wellner & Zhang (2007), with condition 4, the above inequality further implies

$$\mathbb{M}\left(\beta_0,\Lambda_0,\hat{\sigma}_n^2\right)-\mathbb{M}\left(\beta,\Lambda,\hat{\sigma}_n^2\right) \geq C\left\{|\beta-\beta_0|^2+\|\Lambda-\Lambda_0\|_{L_2(\mu)}^2\right\}$$

□

3.3.2 Asymptotic Properties of the pseudo-MLE

Theorem 3.13. *(Consistency). Suppose that conditions 1-3,5 and 7 hold and the counting process* $\mathbb{N}$ *satisfies the proportional mean regression model. Then given* $\hat{\sigma}_n^2$*,* *a consistent estimate of the overdispersion parameter* σ_0^2*,*

$$d\left(\left(\hat{\beta}_n,\hat{\Lambda}_n\right),(\beta_0,\Lambda_0)\right)\rightarrow_p 0$$

Proof. The proof of the consistency is done by checking the three conditions specified in Theorem 3.9 with $\theta = (\beta,\Lambda)$ and $\eta = \sigma^2$. Condition (3.1) is automatically true by the result of Lemma 3.12. Now we prove the uniform convergence condition specified in condition (3.2). Let $\mathcal{L}_1 = \{m(\beta,\Lambda,\sigma^2), \beta \in \mathcal{R}^d, log\Lambda \in \mathcal{F}, \sigma^2 \in \mathcal{R}^+\}$. Since $\mathcal{F}$ is a class of monotone nondecreasing functions, by Theorem 2.7.5 of van der Vaart & Wellner (1996) (Lemma 3.4), $\mathcal{F}$ is covered by $\left\{[\Lambda_i^L,\Lambda_i^R] : i = 1,2,\cdots,l\right\}$, $l = O\left(exp(1/\varepsilon)\right)$ and $\|\Lambda_i^R - \Lambda_i^L\|_{L_1(\mu)} = \int\left(\Lambda_i^R(t) - \Lambda_i^L(t)\right)d\mu(t) < \varepsilon$. Let $\Lambda_{i,j}^R = \Lambda_i^R(T_{K,j})$ and $\Lambda_{i,j}^L = \Lambda_i^L(T_{K,j})$. We further define

$$\Delta\Lambda_{i,j}^L = \Lambda_{i,j}^L - \Lambda_{i,j-1}^L; \qquad \Delta\Lambda_{i,j}^R = \Lambda_{i,j}^R - \Lambda_{i,j-1}^R;$$

$$\Delta\Lambda_{i,j}^{RL} = \Lambda_{i,j}^R - \Lambda_{i,j-1}^L; \qquad \Delta\Lambda_{i,j}^{LR} = \Lambda_{i,j}^L - \Lambda_{i,j-1}^R;$$

We can make these bracketing functions satisfy $\Lambda_i^R - \Lambda_i^L \leq \gamma_1$ and $\Lambda_i^L \geq \gamma_2$ with $\gamma_1,\gamma_2 > 0$ for all $t \in [0,\tau]$ and $1 \leq i \leq l$. And $\Delta\Lambda_{i,j}^{LR} \geq \gamma_3 > 0$. The proof of this claim follows the same lines as given by Wellner & Zhang (1995). Then

$$\begin{aligned}
&\sum_{j=1}^K\left(\Delta\Lambda_{i,j}^{RL} - \Delta\Lambda_{i,j}^{LR}\right)\\
=&\sum_{j=1}^K\left\{\Lambda_i^R(T_{K,j}) - \Lambda_i^L(T_{K,j}) + \Lambda_i^R(T_{K,j-1}) - \Lambda_i^L(T_{K,j-1})\right\}\\
\leq& C\sum_{j=1}^K\left(\Lambda_{i,j}^R - \Lambda_{i,j}^L\right) = C\sum_{j=1}^K\sum_{l=1}^j\left(\Delta\Lambda_{i,l}^R - \Delta\Lambda_{i,l}^L\right)\\
=& C\sum_{l=1}^K(K-l+1)\left(\Delta\Lambda_{i,l}^R - \Delta\Lambda_{i,l}^L\right) \leq CK\sum_{l=1}^K\left(\Delta\Lambda_{i,l}^R - \Delta\Lambda_{i,l}^L\right) \qquad (3.11)
\end{aligned}$$

Since $\mathcal{R}^d$ is compact, there exists a ε-net, $\{\beta_1, \beta_2, \cdots, \beta_p\}$, $p = O\left(1/\varepsilon^d\right)$ such that $\forall \beta \in \mathcal{R}^d, \exists s \in \{1, 2, \cdots, p\}$ such that $|\beta^T Z - \beta_s^T Z| \leq \varepsilon$ and $|exp\left(\beta^T Z\right) - exp\left(\beta_s^T Z\right)| \leq C\varepsilon$. Similarly by the compactness of $\mathcal{R}^+$, there exists another ε-net, $\left\{\sigma_1^2, \sigma_2^2, \cdots, \sigma_q^2\right\}$, $q = O\left(1/\varepsilon\right)$ such that $\forall \sigma^2 \in \mathcal{R}^+, \exists t \in \{1, 2, \cdots, q\}$ such that $|\frac{1}{\sigma^2} - \frac{1}{\sigma_t^2}| \leq \varepsilon$.

Let

$$\begin{aligned}
m_{i,s,t}^L = &\sum_{j=1}^{K} \Delta\mathbb{N}_j \left(log\Delta\Lambda_{i,j}^{LR} + \left(\beta_s^T Z - \varepsilon\right)\right) \\
&\qquad - \left(\mathbb{N}_K + 1/\sigma_t^2 + \varepsilon\right) log\left(\Lambda_{i,K}^R\left(e^{\beta_s^T Z} + C\varepsilon\right) + 1/\sigma_t^2 + \varepsilon\right) \\
m_{i,s,t}^R = &\sum_{j=1}^{K} \Delta\mathbb{N}_j \left(log\Delta\Lambda_{i,j}^{RL} + \left(\beta_s^T Z + \varepsilon\right)\right) \\
&\qquad - \left(\mathbb{N}_K + 1/\sigma_t^2 - \varepsilon\right) log\left(\Lambda_{i,K}^L\left(e^{\beta_s^T Z} - C\varepsilon\right) + 1/\sigma_t^2 - \varepsilon\right)
\end{aligned}$$

So, $\mathcal{L}_1$ is covered by $\left\{[m_{i,s,t}^L, m_{i,s,t}^R], i = 1, 2, \cdots, l, s = 1, 2, \cdots, p, t = 1, 2, \cdots, q\right\}$. And

$$\begin{aligned}
f_{i,s,t} =& m_{i,s,t}^R - m_{i,s,t}^L \\
=& \sum_{j=1}^{K} \Delta\mathbb{N}_j \left(log\Delta\Lambda_{i,j}^{RL} - log\Delta\Lambda_{i,j}^{LR} + 2\varepsilon\right) \\
&+ \left(\mathbb{N}_K + 1/\sigma_t^2 + \varepsilon\right) log\left(\Lambda_K^R\left(e^{\beta_s^T Z} + \varepsilon\right) + 1/\sigma_t^2 + C\varepsilon\right) \\
&- \left(\mathbb{N}_K + 1/\sigma_t^2 - \varepsilon\right) log\left(\Lambda_K^L\left(e^{\beta_s^T Z} - \varepsilon\right) + 1/\sigma_t^2 - C\varepsilon\right) \\
=& \sum_{j=1}^{K} \Delta\mathbb{N}_j \left(log\Delta\Lambda_{i,j}^{RL} - log\Delta\Lambda_{i,j}^{LR} + 2\varepsilon\right) \\
&+ \left(\mathbb{N}_K + 1/\sigma_t^2\right) log\frac{\left(\Lambda_K^R\left(e^{\beta_s^T Z} + C\varepsilon\right) + 1/\sigma_t^2 + \varepsilon\right)}{\left(\Lambda_K^L\left(e^{\beta_s^T Z} - C\varepsilon\right) + 1/\sigma_t^2 - \varepsilon\right)} \\
&+ \varepsilon log\left\{\left(\Lambda_K^R\left(e^{\beta_s^T Z} + C\varepsilon\right) + 1/\sigma_t^2 + \varepsilon\right)\left(\Lambda_K^R\left(e^{\beta_s^T Z} - C\varepsilon\right) + 1/\sigma_t^2 - \varepsilon\right)\right\}
\end{aligned}$$

By Taylor expansion,

$$
\begin{aligned}
log\Delta\Lambda_{ij}^{RL} - log\Delta\Lambda_{ij}^{LR} &= \frac{1}{\xi_{ij}}\left\{\Delta\Lambda_{ij}^{RL} - \Delta\Delta\Lambda_{ij}^{LR}\right\} \quad (\text{Where } \gamma_3 \leq \Delta\Lambda_{ij}^{LR} \leq \xi_{ij} \leq \Delta\Lambda_{ij}^{RL}) \\
&\leq CK\sum_{l=1}^{K}\left(\Delta\Lambda_{il}^{R} - \Delta\Lambda_{il}^{L}\right) \quad (\text{ by Inequality in (3.11)})
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&log\left(\Lambda_{i,K}^{R}\left(e^{\beta_s^T Z} + C\varepsilon\right) + 1/\sigma_t^2 + \varepsilon\right) - log\left(\Lambda_{i,K}^{L}\left(e^{\beta_s^T Z} - C\varepsilon\right) + 1/\sigma_t^2 - \varepsilon\right) \\
=&\frac{1}{\xi_{iK}}\left\{e^{\beta_s^T Z}\left(\Lambda_{i,K}^{R} - \Lambda_{i,K}^{L}\right) + \left(C\left(\Lambda_{i,K}^{R} + \Lambda_{i,K}^{L}\right) + 2\right)\varepsilon\right\} \\
&(\text{Where } \Lambda_{i,K}^{L}\left(e^{\beta_s^T Z} - \varepsilon\right) + 1/\sigma_t^2 - \varepsilon \leq \xi_{iK} \leq \Lambda_{i,K}^{R}\left(e^{\beta_s^T Z} + \varepsilon\right) + 1/\sigma_t^2 + \varepsilon) \\
\leq&C_1\left(\Lambda_{i,K}^{R} - \Lambda_{K}^{L}\right) + C_2\varepsilon \qquad (\text{by the boundness of } \Lambda_{i,K}^{L}, \Lambda_{i,K}^{L} \text{ and } Z) \\
=&C_1\left(\sum_{j=1}^{K}\left(\Delta\Lambda_{i,j}^{R} - \Delta\Lambda_{i,j}^{L}\right)\right) + C_2\varepsilon
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
|f_{i,s,t}| \leq\ & C_1\mathbb{N}_K K\left(\sum_{j=1}^{K}\left(\Delta\Lambda_{i,j}^{R} - \Delta\Lambda_{i,j}^{L}\right) + 2\varepsilon\right) \\
&+ \left(\mathbb{N}_K + 1/\sigma_t^2\right)\left(C_2\sum_{j=1}^{K}\left(\Delta\Lambda_{i,j}^{R} - \Delta\Lambda_{i,j}^{L}\right) + C_3\varepsilon\right) \\
&+ C_4\varepsilon log\left\{\left(\Lambda_{K}^{R}\left(e^{\beta_s^T Z} + C\varepsilon\right) + 1/\sigma_t^2 + C\varepsilon\right)\left(\Lambda_{K}^{R}\left(e^{\beta_s^T Z} - C\varepsilon\right) + 1/\sigma_t^2 - C\varepsilon\right)\right\} \\
\leq\ & C_1\mathbb{N}_K\sum_{j=1}^{K}\left(\Delta\Lambda_{i,j}^{R} - \Delta\Lambda_{i,j}^{L}\right) + C_2\varepsilon
\end{aligned}
$$

Then $P|f_{i,s}| \leq P\left(C_1\mathbb{N}_K\sum_{j=1}^{K}\left(\Delta\Lambda_{j}^{R} - \Delta\Lambda_{j}^{L}\right) + C_2\varepsilon\right) \leq C\varepsilon$. The total number of brackets of $\mathcal{L}_1$ is $Cexp(1/\varepsilon)\cdot(1/\varepsilon)^{d+1}$. By Theorem 2.4.1 of van der Vaart & Wellner (1996) (Lemma 3.1), $\mathcal{L}_1$ is a Glivenko-Cantelli. This guarantees the uniform convergence condition (3.2) in Theorem 3.9.

Now we prove the nearly maximization condition in 3.3 in Theorem 3.9. According to page 148 in de Boor (2001), there exist a $\Lambda_{0,n} \in \psi_{l,t}$ of order $m \geq p+2$ such that $\|\Lambda_{0,n} - \Lambda_0\|_\infty \leq C q_n^{-p} = O(n^{-p\nu})$. Since,

$$\begin{aligned}
&\mathbb{M}_n\left(\hat{\beta}_n, \hat{\Lambda}_n, \sigma^2\right) - \mathbb{M}_n\left(\beta_0, \Lambda_0, \sigma^2\right) \\
=&\mathbb{M}_n\left(\hat{\beta}_n, \hat{\Lambda}_n, \sigma^2\right) - \mathbb{M}_n\left(\beta_0, \Lambda_{0,n}, \sigma^2\right) + \mathbb{M}_n\left(\beta_0, \Lambda_{0,n}, \sigma^2\right) - \mathbb{M}_n\left(\beta_0, \Lambda_0, \sigma^2\right) \\
\geq&\mathbb{M}_n\left(\beta_0, \Lambda_{0,n}, \sigma^2\right) - \mathbb{M}_n\left(\beta_0, \Lambda_0, \sigma^2\right) \\
=&\left(\mathbb{P}_n - P\right)\left\{m\left(\beta_0, \Lambda_{0,n}, \sigma^2\right) - m\left(\beta_0, \Lambda_0, \sigma^2\right)\right\} + \\
&\qquad P\left\{m\left(\beta_0, \Lambda_{0,n}, \sigma^2\right) - m\left(\beta_0, \Lambda_0, \sigma^2\right)\right\} \qquad (3.12)
\end{aligned}$$

Let $\mathcal{L}_2 = \{m\left(\beta_0, \Lambda_0, \sigma^2\right) : \sigma^2 \in \mathcal{R}^+\}$. Similar to the proof shown before, there exists an ε-net, $\left\{\sigma_1^2, \sigma_2^2, \cdots, \sigma_q^2\right\}$, $q = O\left(1/\varepsilon\right)$ such that $\forall \sigma^2 \in \mathcal{R}^+, \exists t \in \{1, 2, \cdots, q\}$ such that $|\frac{1}{\sigma^2} - \frac{1}{\sigma_t^2}| \leq \varepsilon$. $\mathcal{L}_2$ is bracketed by $\left[m_t^L, m_t^R\right]$ with $t = 1, 2, \cdots, q$, where

$$\begin{aligned}
m_t^L &= \sum_{j=1}^{K} \Delta\mathbb{N}_j \left(log\triangle\Lambda_{0,j} + \beta_0^T Z\right) - \left(\mathbb{N}_K + 1/\sigma_t^2 + \varepsilon\right) log\left(\Lambda_{0,K} + 1/\sigma_t^2 + \varepsilon\right) \\
m_t^R &= \sum_{j=1}^{K} \Delta\mathbb{N}_j \left(log\triangle\Lambda_{0,j} + \beta_0^T Z\right) - \left(\mathbb{N}_K + 1/\sigma_t^2 - \varepsilon\right) log\left(\Lambda_{0,K} + 1/\sigma_t^2 - \varepsilon\right)
\end{aligned}$$

So

$$
\begin{aligned}
& m_t^R - m_t^L \\
& = \left(\mathbb{N}_K + 1/\sigma_t^2 + \varepsilon\right) log\left(\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma_t^2 + \varepsilon\right) \\
& \qquad - \left(\mathbb{N}_K + 1/\sigma_t^2 - \varepsilon\right) log\left(\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma_t^2 - \varepsilon\right) \\
& = \left(\mathbb{N}_K + 1/\sigma_t^2 + \varepsilon\right)\left\{log\left(\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma_t^2 + \varepsilon\right) - log\left(\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma_t^2 - \varepsilon\right)\right\} \\
& \qquad + \varepsilon\left\{log\left(\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma_t^2 + \varepsilon\right) + log\left(\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma_t^2 - \varepsilon\right)\right\} \\
& = \left(\mathbb{N}_K + 1/\sigma_t^2 + \varepsilon\right)\frac{1}{\xi_t} 2\varepsilon + C\varepsilon
\end{aligned}
$$

Where $\Lambda_{0,K} + 1/\sigma_t^2 - \varepsilon \leq \xi_t \leq \Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma_t^2 + \varepsilon$. Then $P\left(m_t^R - m_t^L\right) \leq C\varepsilon$. So the bracket number of $\mathcal{L}_2$ is $C(1/\varepsilon)$. $\mathcal{L}_2$ is a Glivenko-Cantelli by the Glivenko-Cantelli Theorem (Lemma 3.1). Since $\Lambda_{0,n} \in \mathcal{F}$ and $\mathcal{L}_1$ is Glivenko-Cantelli, the derivation in (3.12) further implies

$$
\mathbb{M}_n\left(\hat{\beta}_n, \hat{\Lambda}_n, \sigma^2\right) - \mathbb{M}_n\left(\beta_0, \Lambda_0, \sigma^2\right) \geq P\left(m\left(\beta_0, \Lambda_{0,n}, \sigma^2\right) - m\left(\beta_0, \Lambda_0, \sigma^2\right)\right) - o_p(1)
$$

For Λ at the neighborhood of Λ_0 and any σ^2,

$$\begin{aligned}
&P\left(m\left(\beta_0,\Lambda_0,\sigma^2\right)-m\left(\beta_0,\Lambda,\sigma^2\right)\right)\\
=&P\left\{\sum_{j=1}^{K}\Delta\Lambda_{0j}e^{\beta_0^TZ}log\frac{\Delta\Lambda_{0j}}{\Delta\Lambda_j}-\left(\Lambda_{0K}e^{\beta_0^TZ}+1/\sigma^2\right)log\frac{\Lambda_{0K}e^{\beta_0^TZ}+1/\sigma^2}{\Lambda_Ke^{\beta_0^TZ}+1/\sigma^2}\right\}\\
=&P\left\{\sum_{j=1}^{K}\Delta\Lambda_{0j}e^{\beta_0^TZ}\left[-\frac{1}{\Delta\Lambda_{0j}}\left(\Delta\Lambda_j-\Delta\Lambda_{0j}\right)+\frac{1}{\xi_{1j}^2}\left(\Delta\Lambda_i-\Delta\Lambda_{0i}\right)^2\right]-\right.\\
&\quad\left(\Lambda_{0K}e^{\beta_0^TZ}+1/\sigma^2\right)\left[-\frac{1}{\Lambda_{0K}e^{\beta_0^TZ}+1/\sigma^2}\left(\Lambda_Ke^{\beta_0^TZ}-\Lambda_{0K}e^{\beta_0^TZ}\right)\right.\\
&\quad\left.\left.+\frac{1}{\xi_2}\left(\Lambda_Ke^{\beta_0^TZ}-\Lambda_{0K}e^{\beta_0^TZ}\right)^2\right]\right\}\\
&\quad(\text{ where }\xi_{1j}\text{ is between }\Delta\Lambda_{0j}\text{ and }\Delta\Lambda_j;\xi_2\text{ is between }\Delta\Lambda_{0K}e^{\beta_0^TZ}\text{ and }\Delta\Lambda_je^{\beta^TZ})\\
=&P\left\{\sum_{j=1}^{K}\Delta\Lambda_{0j}e^{\beta_0^TZ}\frac{1}{\xi_{1j}^2}\left(\Delta\Lambda_j-\Delta\Lambda_{0j}\right)^2-\left(\Lambda_{0K}e^{\beta_0^TZ}+1/\sigma^2\right)\right.\\
&\quad\left.\times\frac{1}{\xi_2^2}\left(\Lambda_Ke^{\beta_0^TZ}-\Lambda_{0K}e^{\beta_0^TZ}\right)^2\right\}\\
\leq&CP\left\{\sum_{j=1}^{K}\left(\Delta\Lambda_j-\Delta\Lambda_{0j}\right)^2\right\}=Cd^2\left(\left(\beta_0,\Lambda\right),\left(\beta_0,\Lambda_0\right)\right).
\end{aligned}$$

The inequality is due to the boundness of Λ_0 and Z by condition 3 and 5. Therefore,

$$Pm\left(\beta_0,\Lambda_{0,n},\sigma^2\right)-Pm\left(\beta_0,\Lambda_0,\sigma^2\right)\geq -Cd^2\left(\Lambda_{0,n},\Lambda_0\right)=-O(n^{-2p\nu})$$

hence for any given σ^2,

$$\mathbb{M}_n\left(\hat{\beta}_n,\hat{\Lambda}_n,\sigma^2\right)-\mathbb{M}_n\left(\beta_0,\Lambda_0,\sigma^2\right)\geq -o_p\left(1\right)\text{ for any }\sigma^2$$

By the compactness of $\mathcal{R}^+$, this further implies

$$\inf_{\sigma^2\in\mathcal{R}^+}\left(\mathbb{M}_n\left(\hat{\beta}_n,\hat{\Lambda}_n,\sigma^2\right)-\mathbb{M}_n\left(\beta_0,\Lambda_0,\sigma^2\right)\right)\geq -o_p\left(1\right)$$

Therefore by applying Theorem 3.9, we have $d\left(\left(\hat{\beta}_n,\hat{\Lambda}_n\right),\left(\beta_0,\Lambda_0\right)\right)\rightarrow_p 0$. □

Theorem 3.14. *(Rate of Convergence). Suppose that Conditions 1-7 hold and the counting process* $\mathbb{N}$ *satisfies the proportional mean regression model. Then given* $\hat{\sigma}_n^2$*, a consistent estimate of the overdispersion parameter* σ_0^2*,*

$$d\left(\left(\hat{\beta}_n, \hat{\Lambda}_n\right), (\beta_0, \Lambda_0)\right) = O_p\left(n^{-min(p\nu,(1-\nu)/2)}\right).$$

Proof. The convergence rate of the estimator is derived by checking the conditions in Theorem 3.10. We set $\Theta_n = \Theta \equiv \mathcal{R}^d \times \mathcal{F}$. Let $\theta_n = \theta_0 = (\beta_0, \Lambda_0)$ and $d_n(\theta, \theta_n) = d(\theta, \theta_0)$ as previously defined. And let $\eta = \sigma^2$.

First, in Lemma 3.12 we have shown that when (β, Λ) is in a neighborhood of (β_0, Λ_0), $\mathbb{M}(\beta_0, \Lambda_0, \sigma^2) - \mathbb{M}(\beta, \Lambda, \sigma^2) \geq Cd^2((\beta, \Lambda), (\beta_0, \Lambda_0))$ for any $\sigma^2 > 0$ where C is a constant related to σ^2. By the compactness of $\mathcal{R}^+$, this further implies

$$\inf_{\sigma^2 \in \mathcal{R}^+} \left(\mathbb{M}\left(\beta_0, \Lambda_0, \sigma^2\right) - \mathbb{M}\left(\beta, \Lambda, \sigma^2\right)\right) \geq Cd^2((\beta, \Lambda), (\beta_0, \Lambda_0))$$

with C being a constant independent of σ^2. And

$$\inf_{\substack{\delta/2 < d((\beta,\beta_0),(\Lambda,\Lambda_0)) \leq \delta, \\ (\beta,\Lambda) \in \mathcal{R}^d \times \mathcal{F}, \sigma^2 \in \mathcal{R}^+}} \left(\mathbb{M}\left(\beta_0, \Lambda_0, \sigma^2\right) - \mathbb{M}\left(\beta, \Lambda, \sigma^2\right)\right) \geq C\delta^2$$

Second, we need to find $\phi_n(\eta)$ such that

$$E \sup_{\substack{\delta/2 < d((\beta,\beta_0),(\Lambda,\Lambda_0)) \leq \delta, \\ (\beta,\Lambda) \in \mathcal{R}^d \times \mathcal{F}, \sigma^2 \in \mathcal{R}^+}} |\left(\mathbb{P}_n - P\right)\left[m\left(\beta, \Lambda, \sigma^2\right) - m\left(\beta_0, \Lambda_0, \sigma^2\right)\right]| \leq C\frac{\phi_n(\delta)}{\sqrt{n}}$$

Define a class $\mathcal{L}_3$ as following,

$$\mathcal{L}_3 = \{m\left(\beta, \Lambda, \sigma^2\right) - m\left(\beta_0, \Lambda_0, \sigma^2\right) : \beta \in \mathcal{R}^d, log\Lambda \in \mathcal{F}, \sigma^2 \in \mathcal{R}^+,$$
$$d((\beta, \Lambda), (\beta_0, \Lambda_0)) \leq \delta\}$$

Again, due to its monotonicity and by Theorem 2.7.5 in van der Vaart & Wellner (1996) (Lemma 3.4) $\mathcal{F}$ is covered by $\left\{[\Lambda_i^L, \Lambda_i^R] : i = 1, 2, \cdots, l\right\}, l = O\left(exp(\delta/\varepsilon)\right)$ and

$$\|\Lambda_i^R - \Lambda_i^L\|_{L_2(\mu)} = \int \left(\Lambda_i^R\left(t\right) - \Lambda_i^L\left(t\right)\right)^2 d\mu\left(t\right) < \varepsilon^2.$$

Since $\mathcal{R}^d$ is compact, there exists a ε-net, $\{\beta_1, \beta_2, \cdots, \beta_p\}$, $p = [M/\varepsilon^d]$ such that $\forall \beta \in \mathcal{R}^d, \exists s \in \{1, 2, \cdots, p\}$ such that $|\beta^T Z - \beta_s^T Z| \leq \varepsilon$ and $|exp\left(\beta^T Z\right) - exp\left(\beta_s^T Z\right)| \leq C\varepsilon$. Similarly by the compactness of $\mathcal{R}^+$, there exists another ε-net, $\left\{\sigma_1^2, \sigma_2^2, \cdots, \sigma_q^2\right\}$, $q = O\left(1/\varepsilon\right)$ such that $\forall \sigma^2 \in \mathcal{R}^+, \exists t \in \{1, 2, \cdots, q\}$ such that $|\frac{1}{\sigma^2} - \frac{1}{\sigma_t^2}| \leq \varepsilon$.

Let $\Delta\Lambda_{ij}^{RL}, \Delta\Lambda_{ij}^{LR}, \Lambda_{iK}^R, \Lambda_{iK}^L, m_{i,s,t}^L, m_{i,s,t}^R$ and m_t^L, m_t^R defined same as those in the proof of consistency. So, $\mathcal{L}_3$ is covered by

$$\left\{[m_{i,s,t}^L - m_t^R, m_{i,s,t}^R - m_t^L], i = 1, \cdots, l, s = 1, \cdots, p, t = 1, \cdots, q\right\}.$$

Denote $\tilde{f}_{i,s,t} = \left(m_{i,s,t}^R - m_t^L\right) - \left(m_{i,s,t}^L - m_t^R\right) = \left(m_{i,s,t}^R - m_{i,s,t}^L\right) + \left(m_t^R - m_t^L\right)$, we have

$$\begin{aligned}
|\tilde{f}_{i,s,t}| &\leq \left|m_{i,s,t}^R - m_{i,s,t}^L\right| + \left|m_t^R - m_t^L\right| \\
&\leq C_1 \mathbb{N}_K \sum_{j=1}^{K} \left(\Delta\Lambda_{i,j}^R - \Delta\Lambda_{i,j}^L\right) + C_2\varepsilon \leq C\mathbb{N}_K \\
|\tilde{f}_{i,s,t}|^2 &\leq \left(C_1 \mathbb{N}_K \sum_{j=1}^{K} \left(\Delta\Lambda_{i,j}^R - \Delta\Lambda_{i,j}^L\right) + C_2\varepsilon\right)^2 \\
&\leq C\left\{\mathbb{N}_K^2 \sum_{j=1}^{K} \left(\Delta\Lambda_{i,j}^R - \Delta\Lambda_{i,j}^L\right)^2 + \varepsilon^2\right\}
\end{aligned}$$

So,

$$\begin{aligned}
\|\tilde{f}_{i,s,t}\|_{P,B}^2 &\leq P\left(\left|\tilde{f}_{i,s,t}\right|^2 e^{\left|\tilde{f}_{i,s,t}\right|}\right) \leq C_1 P\left\{e^{C_2\mathbb{N}_K}\left(\mathbb{N}_K^2 \sum_{j=1}^{K} \left(\Delta\Lambda_{K,j}^R - \Delta\Lambda_{K,j}^L\right)^2 + \varepsilon^2\right)\right\} \\
&\leq CP\left(\sum_{j=1}^{K} \left(\Delta\Lambda_{K,j}^R - \Delta\Lambda_{K,j}^L\right)^2 + \varepsilon^2\right) \leq C\varepsilon^2
\end{aligned}$$

This shows the number of ε-brackets for $\mathcal{L}_3$ is $Cexp\left(\delta/\varepsilon\right)\cdot\left(1/\varepsilon\right)^{d+1}$.

By definition,

$$\begin{aligned}\tilde{J}_{[]}\left(\delta,\mathcal{L}_3,\|\cdot\|_{P,B}\right)&=\int_0^\delta\sqrt{1+logN_{[]}\left(\varepsilon,\mathcal{L}_3,\|\cdot\|_{P,B}\right)}d\varepsilon\\&\leq\int_0^\delta\sqrt{1+C\left(\delta/\varepsilon\right)}d\varepsilon\\&\leq C\int_0^\delta\sqrt{\left(\delta/\varepsilon\right)}d\varepsilon\leq C\delta\end{aligned}$$

Then by Lemma 3.4.3 in van der Vaart & Wellner (1996) (Lemma 3.6),we have $E_P\|\mathbb{G}_n\|_{\mathcal{L}_3}\leq C\phi_n\left(\delta\right)$ where $\phi_n\left(\delta\right)=\delta+\frac{1}{\sqrt{n}}$. Therefore,

$$E\left|\sqrt{n}\left(\mathbb{P}_n-P\right)\left(m\left(\beta,\Lambda,\sigma^2\right)-m\left(\beta_0,\Lambda_0,\sigma^2\right)\right)\right|_{\mathcal{L}_3}\leq C\phi_n\left(\delta\right)$$

Third, we prove $\inf_{\eta\in H}\left(\mathbb{M}_n\left(\hat{\theta}_n,\eta\right)-\mathbb{M}_n\left(\theta_0,\eta\right)\right)\geq -O_p\left(\delta_n^2\right)$. As shown in the proof of the consistency,

$$\begin{aligned}&\mathbb{M}_n\left(\hat{\beta}_n,\hat{\Lambda}_n,\sigma^2\right)-\mathbb{M}_n\left(\beta_0,\Lambda_0,\sigma^2\right)\\&\geq\left(\mathbb{P}_n-P\right)\left(m\left(\beta_0,\Lambda_{0,n},\sigma^2\right)-m\left(\beta_0,\Lambda_0,\sigma^2\right)\right)+P\left(m\left(\beta_0,\Lambda_{0,n},\sigma^2\right)-m\left(\beta_0,\Lambda_0,\sigma^2\right)\right)\\&=I_{1,n}+I_{2,n}\end{aligned}$$

$$I_{1,n}=\left(\mathbb{P}_n-P\right)\left(m\left(\beta_0,\Lambda_{0,n},\sigma^2\right)-m\left(\beta_0,\Lambda_0,\sigma^2\right)\right)$$

By Taylor expansion,

$$m\left(\beta_0,\Lambda_{0,n},\sigma^2\right)-m\left(\beta_0,\Lambda_0,\sigma^2\right)=\dot{m}\left(\beta_0,\left(1-\xi\right)\Lambda_0+\xi\Lambda_{0,n},\sigma^2\right)\left(\Lambda_{0,n}-\Lambda_0\right),0<\xi<1$$

Define a class $\mathcal{L}_4$ as following,

$$\mathcal{L}_4=\left\{\dot{m}\left(\beta_0,\left(1-\xi\right)\Lambda_0+\xi\Lambda,\sigma^2\right)\left(\Lambda-\Lambda_0\right):\Lambda\in\mathcal{F},\sigma^2\in\mathcal{R}^+,0<\xi<1\right\}$$

It can be similarly shown that the bracketing number of $\mathcal{L}_4$ is bounded by $Cexp\,(1/\varepsilon)\cdot(1/\varepsilon)^2$. By the Donsker Theorem (Lemma 3.3), $\mathcal{L}_4$ is a Donsker. Because $\|\Lambda_{0,n}-\Lambda_0\|_\infty = O\left(n^{-p\nu}\right)$ and $\dot{m}\left(\beta_0, (1-\xi)\,\Lambda_0 + \xi\Lambda_{0,n}, \sigma^2\right)$ is bounded by conditions 2-4, $P\left\{\dot{m}\left(\beta_0, (1-\xi)\,\Lambda_0 + \xi\Lambda_{0,n}, \sigma^2\right)\frac{\Lambda_{0,n}-\Lambda_0}{n^{-p\nu+\varepsilon}}\right\}^2 \to 0$. By the asymptotic equicontinuity (Lemma 3.5), this implies

$$\sup_{\sigma^2\in\mathcal{R}^+} |\left(\mathbb{P}_n - P\right)\left\{\dot{m}\left(\beta_0, \tilde{\Lambda}, \sigma^2\right)\frac{\Lambda_{0,n}-\Lambda_0}{n^{-p\nu+\varepsilon}}\right\}| = o_p\left(n^{-1/2}\right)$$

Hence,

$$\sup_{\sigma^2\in\mathcal{R}^+} |\left(\mathbb{P}_n - P\right)\left(m\left(\beta_0, \Lambda_{0,n}, \sigma^2\right) - m\left(\beta_0, \Lambda_0, \sigma^2\right)\right) = o_p\left(n^{-p\nu+\varepsilon}n^{-1/2}\right)| = o_p\left(n^{-2p\nu}\right)$$

And as shown in the proof of consistency, for any $\sigma^2 > 0$,

$$I_{2,n} = P\left(m\left(\beta_0, \Lambda_{0,n}, \sigma^2\right) - m\left(\beta_0, \Lambda_0, \sigma^2\right)\right) \geq -Cd^2\left(\Lambda_{0,n}, \Lambda_0\right) = O\left(n^{-2p\nu}\right)$$

Thus by the compactness of $\mathcal{R}^+$,

$$\begin{aligned}\inf_{\sigma^2\in\mathcal{R}^+} \mathbb{M}_n\left(\hat{\beta}_n, \hat{\Lambda}_n, \sigma^2\right) - \mathbb{M}_n\left(\beta_0, \Lambda_0, \sigma^2\right) &\geq o_p\left(n^{-2p\nu}\right) - O_p\left(n^{-2p\nu}\right) = -O_p\left(n^{-2p\nu}\right)\\ &= -O_p\left(n^{-2min(p\nu,(1-\nu)/2)}\right)\end{aligned}$$

Let $r_n = n^{min(p\nu,(1-\nu)/2)} \leq n^{(1-\nu)/2}$. Then

$$r_n^2\phi_n\left(1/r_n\right) = r_n^2\left(r_n^{-1} + n^{-1/2}\right) = r_n + r_n^2 n^{-1/2} \leq n^{1/2-\nu/2} + n^{1-\nu-1/2} < Cn^{1/2}$$

So $r_n d\left(\left(\beta, \Lambda\right), \left(\beta_0, \Lambda_0\right)\right) = O_p\left(1\right)$. □

The Gamma-Frailty Poisson likelihood for one observation is specified in Equation (3.9). Correspondingly,

$$\begin{aligned}
m_1\left(\beta,\Lambda,\sigma^2;X\right) =& \frac{\mathbb{N}_K-\Lambda_K e^{\beta^TZ}}{\Lambda_K e^{\beta^TZ}+1/\sigma^2}\times\frac{1}{\sigma^2}\times Z\\
m_2\left(\beta,\Lambda,\sigma^2;X\right)[h] =& \sum_{j=1}^{K}\frac{\Delta\mathbb{N}_j}{\Delta\Lambda_j}\Delta h_j-\frac{\mathbb{N}_K+1/\sigma^2}{\Lambda_K e^{\beta^TZ}+1/\sigma^2}h_K e^{\beta^TZ}\\
m_{11}\left(\beta,\Lambda,\sigma^2;X\right) =& -\frac{\mathbb{N}_K+1/\sigma^2}{\left(\Lambda_K e^{\beta^TZ}+1/\sigma^2\right)^2}\times\Lambda_K e^{\beta^TZ}\times\frac{1}{\sigma^2}ZZ^T\\
m_{12}\left(\beta,\Lambda,\sigma^2;X\right)[h] =& -\frac{\mathbb{N}_K+1/\sigma^2}{\left(\Lambda_K e^{\beta^TZ}+1/\sigma^2\right)^2}\times\frac{1}{\sigma^2}Zh_K e^{\beta^TZ}\\
m_{22}\left(\beta,\Lambda,\sigma^2;X\right)[h^*_{\sigma^2},h] =& -\sum_{j=1}^{K}\frac{\Delta\mathbb{N}_j}{\triangle\Lambda_j^2}\Delta h^*_{\sigma^2,j}\Delta h_j^T+\\
& \frac{\mathbb{N}_K+1/\sigma^2}{\left(\Lambda_K e^{\beta^TZ}+1/\sigma^2\right)^2}\left(h^*_{\sigma^2,K}e^{\beta^TZ}\right)\left(h_k e^{\beta^TZ}\right)
\end{aligned}$$

In order to apply Theorem 3.11 to the Gamma-frailty Poisson model, we first find $h^*_{\sigma^2}$ such that

$$P\left(m_{12}\left(\beta_0,\Lambda_0,\sigma^2;X\right)[h]-m_{22}\left(\beta_0,\Lambda_0,\sigma^2;X\right)[h^*_{\sigma^2},h]\right)=0\quad\forall h\in\mathcal{H}.$$

$$\begin{aligned}
&P\left(m_{12}\left(\beta_0,\Lambda_0,\sigma^2;X\right)[h]-m_{22}\left(\beta_0,\Lambda_0,\sigma^2;X\right)[h^*_{\sigma^2},h]\right)\\
&=P\left\{-\frac{\mathbb{N}_K+1/\sigma^2}{\left(\Lambda_{0,K}e^{\beta_0^TZ}+1/\sigma^2\right)^2}\times\frac{1}{\sigma^2}Zh_K e^{\beta_0^TZ}+\right.\\
&\qquad\left.\sum_{j=1}^{K}\frac{\Delta\mathbb{N}_j}{\triangle\Lambda_{0,j}^2}\Delta h^*_{\sigma^2,j}\Delta h_j-\frac{\mathbb{N}_K+1/\sigma^2}{\left(\Lambda_{0,K}e^{\beta_0^TZ}+1/\sigma^2\right)^2}\left(h^*_{\sigma^2,K}e^{\beta_0^TZ}\right)\left(h_K e^{\beta^TZ}\right)\right\}\\
&=P\left\{\sum_{j=1}^{K}\frac{\Delta\mathbb{N}_j}{\triangle\Lambda_{0,j}^2}\Delta h^*_{\sigma^2,j}\Delta h_j-\frac{\mathbb{N}_K+1/\sigma^2}{\left(\Lambda_{0,K}e^{\beta_0^TZ}+1/\sigma^2\right)^2}\left(Z\times\frac{1}{\sigma^2}+h^*_{\sigma^2,K}e^{\beta_0^TZ}\right)\left(h_K e^{\beta_0^TZ}\right)\right\}\\
&=P\left\{\sum_{j=1}^{K}\left[\frac{\Delta\mathbb{N}_j}{\triangle\Lambda_{0,j}^2}\Delta h^*_{\sigma^2,j}-\frac{\mathbb{N}_K+1/\sigma^2}{\left(\Lambda_{0,K}e^{\beta_0^TZ}+1/\sigma^2\right)^2}\left(Z\times\frac{1}{\sigma^2}+h^*_{\sigma^2,K}e^{\beta_0^TZ}\right)e^{\beta_0^TZ}\right]\Delta h_j\right\}\\
&=P\left\{\sum_{j=1}^{K}\left[\frac{e^{\beta_0^TZ}}{\triangle\Lambda_{0,j}}\Delta h^*_{\sigma^2,j}-\frac{1}{\left(\Lambda_{0,K}e^{\beta_0^TZ}+1/\sigma^2\right)}\left(Z\times\frac{1}{\sigma^2}+h^*_{\sigma^2,K}e^{\beta_0^TZ}\right)e^{\beta_0^TZ}\right]\Delta h_j\right\}
\end{aligned}$$

An obvious choice of h^* is the one that satisfies,

$$E\left(\frac{e^{\beta_0^T Z}\Delta h^*_{\sigma^2,j}}{\triangle\Lambda_{0,j}}|K,T\right) = E\left(\frac{Z \times 1/\sigma^2 + h^*_{\sigma^2,K}e^{\beta_0^T Z}}{\Lambda_{0,K}e^{\beta_0^T Z} + 1/\sigma^2} \times e^{\beta_0^T Z}|K,T\right)$$

where $T = (T_{K,1}, T_{K,2}, \cdots, T_{K,K})$. Set $h^*_j = \frac{\Lambda_{0,j}}{E\left(e^{\beta_0^T Z}|K,T\right)}a$ for $j = 1, 2, \cdots, K$ and set

$$E\left(\frac{e^{\beta_0^T Z}}{\triangle\Lambda_{0,j}}\Delta h^*_{\sigma^2,j} - \frac{Z \times 1/\sigma^2 + h^*_{\sigma^2,K}e^{\beta_0^T Z}}{\Lambda_{0,K}e^{\beta_0^T Z} + 1/\sigma^2} \times e^{\beta_0^T Z}|K,T\right) \equiv 0$$

we have

$$a = E\left(\frac{Z \times 1/\sigma^2 + \frac{\Lambda_{0,K}e^{\beta_0^T Z}}{E\left(e^{\beta_0^T Z}|K,T\right)}a}{\Lambda_{0,K}e^{\beta_0^T Z} + 1/\sigma^2} \times e^{\beta_0^T Z}|K,T\right)$$

$$\Rightarrow a = \frac{E\left(\frac{Z\times 1/\sigma^2}{\Lambda_{0,K}e^{\beta_0^T Z}+1/\sigma^2} \times e^{\beta_0^T Z}|K,T\right)}{E\left(e^{\beta_0^T Z}|K,T\right) - E\left(\frac{\Lambda_{0,K}e^{2\beta_0^T Z}}{\Lambda_{0,K}e^{\beta_0^T Z}+1/\sigma^2}|K,T\right)}E\left(e^{\beta_0^T Z}|K,T\right)$$

So

$$h^*_{\sigma^2,j} = \Lambda_{0,j} \times S; \quad S = \frac{E\left(\frac{Z\times 1/\sigma^2}{\Lambda_{0,K}e^{\beta_0^T Z}+1/\sigma^2} \times e^{\beta_0^T Z}|K,T\right)}{E\left(e^{\beta_0^T Z}|K,T\right) - E\left(\frac{\Lambda_{0,K}e^{2\beta_0^T Z}}{\Lambda_{0,K}e^{\beta_0^T Z}+1/\sigma^2}|K,T\right)} \tag{3.13}$$

$$\begin{aligned}
&m^*\left(\beta_0, \Lambda_0, \sigma^2\right) \\
&= m_1\left(\beta_0, \Lambda_0, \sigma^2; X\right) - m_2\left(\beta_0, \Lambda_0, \sigma^2; X\right)\left[h^*_{\sigma^2}\right] \\
&= \frac{\mathbb{N}_K - \Lambda_{0,K}e^{\beta_0^T Z}}{\Lambda_{0,K}e^{\beta_0^T Z} + 1/\sigma^2} \times \frac{1}{\sigma^2} \times Z - \sum_{j=1}^{K}\frac{\Delta\mathbb{N}_j}{\Delta\Lambda_{0,j}}\Delta h^*_{\sigma^2,j} + \frac{\mathbb{N}_K + 1/\sigma^2}{\Lambda_{0,K}e^{\beta_0^T Z} + 1/\sigma^2}h^*_{\sigma^2,K}e^{\beta_0^T Z} \\
&= \frac{\mathbb{N}_K - \Lambda_{0,K}e^{\beta_0^T Z}}{\Lambda_{0,K}e^{\beta_0^T Z} + 1/\sigma^2} \times \frac{1}{\sigma^2}\left(Z - S\right)
\end{aligned}$$

And,

$$\begin{aligned}
A\left(\beta_0, \Lambda_0, \sigma^2\right) =& P\left\{-m_{11}\left(\beta_0, \Lambda_0, \sigma^2; X\right) + m_{12}\left(\beta_0, \Lambda_0, \sigma^2; X\right)\left[h^*_{\sigma^2}\right]\right\} \\
=& P\left\{\frac{\mathbb{N}_K + 1/\sigma^2}{\left(\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^2\right)^2} \times \Lambda_{0,K} e^{\beta_0^T Z} \times \frac{1}{\sigma^2} Z Z^T \right. \\
& \left. - \frac{\mathbb{N}_K + 1/\sigma^2}{\left(\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^2\right)^2} \times \frac{1}{\sigma^2} Z h^{*T}_{\sigma^2, K} e^{\beta_0^T Z}\right\} \\
=& P\left\{\frac{\Lambda_{0,K} e^{\beta_0^T Z} \times 1/\sigma^2}{\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^2} \times Z\left(Z - S\right)^T\right\} \\
B\left(\beta_0, \Lambda_0, \sigma^2\right) =& P m^*\left(\beta_0, \Lambda_0, \sigma^2; X\right)^{\otimes 2} = P\left\{\left(\frac{\mathbb{N}_K - \Lambda_{0,K} e^{\beta_0^T Z}}{\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^2} \times \frac{1}{\sigma^2}\right)^2 \left(Z - S\right)^{\otimes 2}\right\}
\end{aligned}$$

Since S can be rewritten as following,

$$S = \frac{E\left(\frac{Z \times 1/\sigma^2}{\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^2} \times e^{\beta_0^T Z} | K, T\right)}{E\left(e^{\beta_0^T Z} | K, T\right) - E\left(\frac{\Lambda_{0,K} e^{2\beta_0^T Z}}{\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^2} | K, T\right)} = \frac{E\left(\frac{e^{\beta_0^T Z} \times 1/\sigma^2}{\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^2} \times Z | K, T\right)}{E\left(\frac{e^{\beta_0^T Z} \times 1/\sigma^2}{\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^2} | K, T\right)}$$

We have

$$\begin{aligned}
& P\left\{\frac{\Lambda_{0,K} e^{\beta_0^T Z} \times 1/\sigma^2}{\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^2} \times S\left(Z - S\right)^T\right\} \\
=& P\left\{\frac{\Lambda_{0,K} e^{\beta_0^T Z} \times 1/\sigma^2}{\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^2} \times S Z^T\right\} - P\left\{\frac{\Lambda_{0,K} e^{\beta_0^T Z} \times 1/\sigma^2}{\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^2} \times S^{\otimes 2}\right\} \\
=& P\left\{\Lambda_{0,K} S E\left(\frac{e^{\beta_0^T Z} \times 1/\sigma^2}{\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^2} \times Z | K, T\right)^T\right\} \\
& - P\left\{\Lambda_{0,K} E\left(\frac{e^{\beta_0^T Z} \times 1/\sigma^2}{\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^{\otimes 2}} | K, T\right) \times S^{\otimes 2}\right\} = 0
\end{aligned}$$

And

$$A\left(\beta_0, \Lambda_0, \sigma^2\right) = P\left\{\frac{\Lambda_{0,K} e^{\beta_0^T Z} \times 1/\sigma^2}{\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^2}\left(Z - S\right)^{\otimes 2}\right\}$$

When the variance of $\mathbb{N}_K$ is correctly specified as displayed in $V_3^{(i)}$, i.e., $Var\left(\mathbb{N}_K\right) = \Lambda_{0,K} e^{\beta_0^T Z}\left(\sigma^2 \Lambda_{0,K} e^{\beta_0^T Z} + 1\right)$, then $A\left(\beta_0, \Lambda_0, \sigma^2\right) = B\left(\beta_0, \Lambda_0, \sigma^2\right)$. They are the information matrix. Otherwise, the sandwich form

$$A\left(\beta_0, \Lambda_0, \sigma^2\right)^{-1} B\left(\beta_0, \Lambda_0, \sigma^2\right) A\left(\beta_0, \Lambda_0, \sigma^2\right)^{-1}$$

gives the robust variance estimate of the regression parameter.

Theorem 3.15. *(Asymptotic Normality). Suppose that Condition 1-7 hold and the counting process* $\mathbb{N}$ *satisfies the proportional mean regression model. Then given* $\hat{\sigma}_n^2$*, a consistent estimate of* σ_0^2*, it follows that*

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) &= - A_0^{-1}\mathbb{G}_n\left(m_1\left(\beta_0, \Lambda_0, \sigma_0^2\right) - m_2\left(\beta_0, \Lambda_0, \sigma_0^2\right)\left[h^*_{\sigma_0^2}\right]\right) + o_p(1)\\ &\to_d N\left(0, A_0^{-1} B_0 A_0^{-1}\right)\end{aligned}$$

where

$$\begin{aligned}A_0 = A\left(\beta_0, \Lambda_0, \sigma_0^2\right) &= -P\left(m_{11}\left(\beta_0, \Lambda_0, \sigma_0^2\right) - m_{21}\left(\beta_0, \Lambda_0, \sigma_0^2\right)\left[h^*_{\sigma_0^2}\right]\right)\\ B_0 = B\left(\beta_0, \Lambda_0, \sigma_0^2\right) &= P\left(m_1\left(\beta_0, \Lambda_0, \sigma_0^2\right) - m_2\left(\beta_0, \Lambda_0, \sigma_0^2\right)\left[h^*_{\sigma_0^2}\right]\right)^{\otimes 2}\end{aligned}$$

Proof. We prove the asymptotic normality of $\hat{\beta}_n$ by checking the assumptions in Theorem 3.11.

1. A1 is satisfied with the consistency and convergence rate of $\left(\hat{\beta}_n, \hat{\Lambda}_n\right)$.

2. $Pm_1\left(\beta_0, \Lambda_0, \sigma^2\right) = 0$ and $Pm_2\left(\beta_0, \Lambda_0, \sigma^2\right)[h] = 0$ as long as the proportional mean model in Equation (1.2) hold.

3. $h^*_{\sigma^2}$ is specified as shown in Equation (3.13).

4. Since $\left(\hat{\beta}_n, \hat{\Lambda}_n\right)$ is estimated by solving the estimating equations, we have

$$\mathbb{P}_n m_1 \left(\hat{\beta}_n, \hat{\Lambda}_n, \sigma^2; X\right) = 0 \text{ and } \mathbb{P}_n m_2 \left(\hat{\beta}_n, \hat{\Lambda}_n, \sigma^2; X\right)[h] = 0 \qquad \forall h \in \mathcal{H}$$

The first part of condition 4 is automatically true. To prove the second part, it suffices to show that

$$I = \mathbb{P}_n \left\{ m_2 \left(\hat{\beta}_n, \hat{\Lambda}_n, \sigma^2; X\right)[h^*_{\sigma^2}] - m_2 \left(\hat{\beta}_n, \hat{\Lambda}_n, \sigma^2; X\right)[\hat{\Lambda}_n S] \right\} = o_p\left(n^{-1/2}\right)$$

With $h^*_{\sigma^2}$ specified as in Equation (3.13) we have

$$\begin{aligned} I =& \mathbb{P}_n \left\{ m_2 \left(\hat{\beta}_n, \hat{\Lambda}_n, \sigma^2; X\right)[\Lambda_0 S] - m_2 \left(\hat{\beta}_n, \hat{\Lambda}_n, \sigma^2; X\right)[\hat{\Lambda}_n S] \right\} \\ =& \mathbb{P}_n \left\{ m_2 \left(\hat{\beta}_n, \hat{\Lambda}_n, \sigma^2; X\right)[\Lambda_0 S - \hat{\Lambda}_n S] \right\} \end{aligned}$$

Since $Pm_2\left(\beta_0, \Lambda_0, \sigma^2; X\right)[h] = 0$ for any $h \in H$. I can be decomposed as $I = I_1 + I_2$, where

$$\begin{aligned} I_1 &= \left(\mathbb{P}_n - P\right) \left\{ m_2 \left(\hat{\beta}_n, \hat{\Lambda}_n, \sigma^2; X\right)[\Lambda_0 S - \hat{\Lambda}_n S] \right\} \\ I_2 &= P \left\{ m_2 \left(\hat{\beta}_n, \hat{\Lambda}_n, \sigma^2; X\right)[\Lambda_0 S - \hat{\Lambda}_n S] - m_2 \left(\beta_0, \Lambda_0, \sigma^2; X\right)[\Lambda_0 S - \hat{\Lambda}_n S] \right\}. \end{aligned}$$

We show that I_1 and I_2 are both $o_p\left(n^{-1/2}\right)$. Let

$$\begin{aligned} \phi\left(X; \beta, \Lambda\right) =& m_2\left(\beta, \Lambda, \sigma^2; X\right)[\Lambda_0 S - \Lambda S] \\ =& \sum_{j=1}^{K} \frac{\Delta \mathbb{N}_j}{\Delta \Lambda_j} \left(\Delta \Lambda_{0,j} - \Delta \Lambda_j\right) \cdot S - \frac{\mathbb{N}_K + 1/\sigma^2}{\Lambda_K e^{\beta^T Z} + 1/\sigma^2} e^{\beta^T Z} \left(\Lambda_{0,K} - \Lambda_K\right) \cdot S \end{aligned}$$

And define a class $\Phi\left(\eta\right)$ as

$$\Phi\left(\eta\right) = \left\{ \phi : \left(\beta, \Lambda\right) \in \mathcal{R}^d \times \mathcal{F}, \sigma^2 \in \mathcal{R}^+ \text{ and } d\left(\left(\beta, \Lambda\right), \left(\beta_0, \Lambda_0\right)\right) \leq \eta \right\}.$$

Due to its monotonicity and by Theorem 2.7.5 in van der Vaart & Wellner (1996) $\mathcal{F}$ is covered by $\left\{[\Lambda_i^L, \Lambda_i^R] : i = 1, 2, \cdots, l\right\}$, $l = O\left(exp(\eta/\varepsilon)\right)$ and

$$\|\Lambda_i^R - \Lambda_i^L\|_{L_2(\mu)} = \int \left(\Lambda_i^R\left(t\right) - \Lambda_i^L\left(t\right)\right)^2 d\mu\left(t\right) < \varepsilon^2.$$

And we can construct an ε-net, $\{\beta_1, \beta_2, \cdots, \beta_p\}$, $p = O\left(1/\varepsilon^d\right)$ such that $\forall \beta \in \mathcal{R}^d, \exists s \in \{1, 2, \cdots, p\}$ such that $|\beta^T Z - \beta_s^T Z| \leq \varepsilon$ and $|exp\left(\beta^T Z\right) - exp\left(\beta_s^T Z\right)| \leq C\varepsilon$. For a fixed σ^2, $\Phi\left(\eta\right)$ is covered by $[m_{i,s}^L, m_{i,s}^R], i = 1, \cdots, l, s = 1, \cdots p$, where

$$\begin{aligned} m_{i,s}^L =& S \sum_{j=1}^{K} \left(\frac{\Delta \mathbb{N}_j}{\Delta \Lambda_{i,j}^{RL}} \Delta \Lambda_{0,j} - \Delta \mathbb{N}_j \right) - \frac{\mathbb{N}_K + 1/\sigma^2}{\Lambda_{i,K}^L \left(e^{\beta_s^T Z} - C\varepsilon\right) + 1/\sigma^2} \times \\ & \Lambda_{0,K} S \left(e^{\beta_s^T Z} + C\varepsilon\right) + \left(\mathbb{N}_K + 1/\sigma^2\right) S - \frac{\left(\mathbb{N}_K + 1/\sigma^2\right) S \cdot 1/\sigma^2}{\Lambda_K^L \left(e^{\beta_s^T Z} - C\varepsilon\right) + 1/\sigma^2} \\ m_{i,s}^R =& S \sum_{j=1}^{K} \left(\frac{\Delta \mathbb{N}_j}{\Delta \Lambda_{i,j}^{LR}} \Delta \Lambda_{0,j} - \Delta \mathbb{N}_j \right) - \frac{\mathbb{N}_K + 1/\sigma^2}{\Lambda_{i,K}^R \left(e^{\beta_s^T Z} + C\varepsilon\right) + 1/\sigma^2} \times \\ & \Lambda_{0,K} S \left(e^{\beta_s^T Z} - C\varepsilon\right) + \left(\mathbb{N}_K + 1/\sigma^2\right) S - \frac{\left(\mathbb{N}_K + 1/\sigma^2\right) S \cdot 1/\sigma^2}{\Lambda_K^R \left(e^{\beta_s^T Z} + C\varepsilon\right) + 1/\sigma^2}. \end{aligned}$$

$\Delta\Lambda_{i,j}^L, \Delta\Lambda_{i,j}^R, \Delta\Lambda_{i,j}^{LR}$ and $\Delta\Lambda_{i,j}^{RL}$ are defined the same as those in the proof of consistency. And

$$\begin{aligned}
& f_{i,s} = m_{i,s}^R - m_{i,s}^L \\
=& S\sum_{j=1}^{K} \Delta\mathbb{N}_j \Delta\Lambda_{0,j}\left(\frac{1}{\Delta\Lambda_j^{LR}} - \frac{1}{\Delta\Lambda_j^{RL}}\right) \\
&\quad + \left(\mathbb{N}_K + 1/\sigma^2\right)\Lambda_{0,K} S\left(\frac{e^{\beta_s^T Z} + C\varepsilon}{\Lambda_{i,K}^L\left(e^{\beta_s^T Z} - C\varepsilon\right) + 1/\sigma^2} - \frac{e^{\beta_s^T Z} - C\varepsilon}{\Lambda_{i,K}^R\left(e^{\beta_s^T Z} + C\varepsilon\right) + 1/\sigma^2}\right) \\
&\quad + \left(\mathbb{N}_K + 1/\sigma^2\right) S \cdot 1/\sigma^2 \left(\frac{1}{\Lambda_{i,K}^L\left(e^{\beta_s^T Z} - C\varepsilon\right) + 1/\sigma^2} - \frac{1}{\Lambda_{i,K}^R\left(e^{\beta_s^T Z} + C\varepsilon\right) + 1/\sigma^2}\right) \\
=& S\sum_{j=1}^{K} \Delta\mathbb{N}_j \Delta\Lambda_{0,j}\frac{\Delta\Lambda_{i,j}^{RL} - \Delta\Lambda_{i,j}^{LR}}{\Delta\Lambda_{i,j}^{RL}\Delta\Lambda_{i,j}^{LR}} + \left(\mathbb{N}_K + 1/\sigma^2\right)\Lambda_{0,K} S \times \\
&\quad \frac{e^{2\beta_s^T Z}\left(\Lambda_{i,K}^R - \Lambda_{i,K}^L\right) + \left(\Lambda_{i,K}^R + \Lambda_{i,K}^L\right)e^{\beta_s^T Z} C\varepsilon + \left(\Lambda_{i,K}^R - \Lambda_{i,K}^L\right)C\varepsilon^2 + C\varepsilon 1/\sigma^2}{\left(\Lambda_{i,K}^L\left(e^{\beta_s^T Z} - C\varepsilon\right) + 1/\sigma^2\right)\left(\Lambda_{i,K}^R\left(e^{\beta_s^T Z} + C\varepsilon\right) + 1/\sigma^2\right)} \\
&\quad + \left(\mathbb{N}_K + 1/\sigma^2\right) S \cdot 1/\sigma^2 \frac{\left(\Lambda_{i,K}^R - \Lambda_{i,K}^L\right)e^{\beta_s^T Z} + \left(\Lambda_{i,K}^R + \Lambda_{i,K}^L\right)C\varepsilon}{\left(\Lambda_{i,K}^L\left(e^{\beta_s^T Z} - C\varepsilon\right) + 1/\sigma^2\right)\left(\Lambda_{i,K}^R\left(e^{\beta_s^T Z} + C\varepsilon\right) + 1/\sigma^2\right)} \\
& \le C_1 \sum_{j=1}^{K}\left(\Delta\Lambda_{i,j}^R - \Lambda_{i,j}^L\right) + C_2\varepsilon
\end{aligned}$$

The last inequality is due to the boundness of $\Delta\Lambda_{i,j}^{LR}$, $\Lambda_{K,i}^L$ from 0 as stated in the proof of consistency and conditions 1,2 and 5. By Cauchy-Schwartz inequality,

$$P\left|f_{i,s}\right|^2 \le P\left(C_1\sum_{j=1}^{K}\left(\Delta\Lambda_{i,j}^R - \Lambda_{i,j}^L\right)^2 + C_2\varepsilon^2\right) \le C\varepsilon^2.$$

Therefore with a fixed σ^2, $\Phi\left(\eta\right)$ has a finite ε-bracketing number using $L_2(P)-$norm, $C\left(exp\left(\eta/\varepsilon\right)\right)\cdot\left(1/\varepsilon\right)^d$. Now we allow σ^2 to vary freely. By Condition 2-4,

$$\frac{\partial}{\partial\sigma^2}\left\{m_2\left(\beta, \Lambda, \sigma^2\right)\left[\Lambda_0 S - \Lambda S\right]\right\}$$

is uniformly bounded. If $m_2\left(\beta, \Lambda, \sigma^2\right)\left[\Lambda_0 S - \Lambda S\right] - m_2\left(\beta_0, \Lambda_0, \sigma^2\right)\left[h^* - h\right]$ is contained in a bracket $[l, u]$, then $m_2\left(\beta, \Lambda, \tilde\sigma^2\right)\left[\Lambda_0 S - \Lambda S\right]$ is contained in the bracket $[l - C\varepsilon, u + C\varepsilon]$ for $\tilde\sigma^2$ with $|\tilde\sigma^2 - \sigma^2| \le \varepsilon$. With the compactness of the parameter space of σ^2, we can select an ε-net, $\left\{\sigma_1^2, \sigma_2^2, \cdots, \sigma_q^2\right\}, q = O(1/\varepsilon)$ over $\mathcal{R}^+$ and construct brackets for each σ_i^2 with enlarged bracket size. So the total number of brackets of $\Phi(\eta)$ is $Cexp(\eta/\varepsilon)(1/\varepsilon)^{d+1}$, So $\Phi(\eta)$ is a P-Donsker.

By conditions 2, 5 and 6 and Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&P\left\{\sum_{j=1}^{K}\frac{\Delta\mathbb{N}_j}{\Delta\Lambda_j}\left(\Delta\Lambda_{0,j}-\Delta\Lambda_j\right)S-\frac{\mathbb{N}_K+1/\sigma^2}{\Lambda_K e^{\beta^T Z}+1/\sigma^2}e^{\beta^T Z}\left(\Lambda_{0,K}-\Lambda_K\right)S\right\}^2\\
\leq&P\left\{\sum_{j=1}^{K}\frac{\Delta\mathbb{N}_j^2}{\Delta\Lambda_j^2}\left(\Delta\Lambda_{0,j}-\Delta\Lambda_j\right)^2S^2+\right.\\
&\qquad\left.\left(\frac{\mathbb{N}_K+1/\sigma^2}{\Lambda_K e^{\beta^T Z}+1/\sigma^2}\right)^2e^{2\beta^T Z}\left(\Lambda_{0,K}-\Lambda_K\right)^2S^2\right\}\\
\leq&CP\sum_{j=1}^{k}\left(\Delta\Lambda_{0,j}-\Delta\Lambda_j\right)^2\leq C\eta^2
\end{aligned}$$

Then

$$\sup_{f\in\Phi(\eta)}\rho_P(f)\leq\sup_{f\in\Phi(\eta)}\left\{Pf^2\right\}^{1/2}\leq C\eta\to 0\text{ as }\eta\to 0$$

Due to the relationship between P-Donsker and equicontinuity, $I_1 = o_p\left(n^{-1/2}\right)$.

$$\begin{aligned}
I_2=&P\left\{\sum_{j=1}^{K}\Delta\mathbb{N}_jS\left(\Delta\Lambda_{0,j}-\Delta\hat{\Lambda}_{n,j}\right)\frac{\Delta\Lambda_{0,j}-\Delta\hat{\Lambda}_{n,j}}{\Delta\Lambda_{0,j}\Delta\hat{\Lambda}_{n,j}}-\left(\mathbb{N}_K+1/\sigma^2\right)\times\right.\\
&\left.S\left(\Lambda_{0,K}-\hat{\Lambda}_{n,K}\right)\frac{\left(\Lambda_{0,K}-\hat{\Lambda}_{n,K}\right)e^{\left(\hat{\beta}_n+\beta_0\right)^T Z}+\left(e^{\hat{\beta}_n^T Z}-e^{\beta_0^T Z}\right)1/\sigma^2}{\left(\hat{\Lambda}_{n,K}e^{\hat{\beta}_n^T Z}+1/\sigma^2\right)\left(\Lambda_{0,K}e^{\beta_0^T Z}+1/\sigma^2\right)}\right\}\\
\leq&P\left\{C_1\sum_{j=1}^{K}\left(\Delta\Lambda_{0,j}-\Delta\hat{\Lambda}_{n,j}\right)^2+C_2\left(\Lambda_{0,K}-\hat{\Lambda}_{n,K}\right)\left(e^{\beta_0^T Z}-e^{\hat{\beta}_n^T Z}\right)\right\}
\end{aligned}$$

(by conditions 1, 2 and 5)

$$\begin{aligned}
\leq&P\left\{C_1\sum_{j=1}^{K}\left(\Delta\Lambda_{0,j}-\Delta\hat{\Lambda}_{n,j}\right)^2+C_2\sum_{j=1}^{K}\left(\beta_0-\hat{\beta}_n\right)^T ZZ^T\left(\beta_0-\hat{\beta}_n\right)\right\}\\
\leq&Cd^2\left(\left(\beta,\Lambda\right),\left(\beta_0,\Lambda_0\right)\right)=O_p\left(n^{-min(2p\nu,1-\nu)}\right)=o_p\left(n^{-1/2}\right)
\end{aligned}$$

5. Define a class,

$$\mathcal{M}_1 = \left\{ m_1\left(\beta, \Lambda, \sigma^2\right) - m_1\left(\beta_0, \Lambda_0, \sigma^2\right) : \|\beta - \beta_0\| < \delta, \|\Lambda - \Lambda_0\| < \delta, \right.$$
$$\left. \beta \in \mathcal{R}^d, \Lambda \in \mathcal{F}, \sigma^2 \in \mathcal{R}^+ \right\}$$

where

$$\begin{aligned} & m_1\left(\beta, \Lambda, \sigma^2\right) - m_1\left(\beta_0, \Lambda_0, \sigma^2\right) \\ = & \frac{\left(\mathbb{N}_K + 1/\sigma^2\right) Z \times 1/\sigma^2}{\left(\Lambda_K e^{\beta^T Z} + 1/\sigma^2\right)\left(\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^2\right)} \left(\Lambda_{0,K} e^{\beta_0^T Z} - \Lambda_K e^{\beta^T Z}\right) \end{aligned}$$

Let $\Lambda_{i,K}^L, \Lambda_{i,K}^R, i = 1, 2, \cdots, l, l = Cexp\left(1/\varepsilon\right)$ and $\beta_s, s = 1, 2, \cdots, p, p = C(1/\varepsilon)^d$ defined same as that in the proof of the consistency. So for a fixed σ^2, $\mathcal{M}_1$ is covered by $\left[m_{1,i,s}^L, m_{1,i,s}^R\right]$ where

$$\begin{aligned} m_{1,i,s}^L = & \frac{\left[\mathbb{N}_K - \Lambda_{i,K}^R\left(e^{\beta_s^T Z} + C\varepsilon\right)\right]}{\Lambda_{i,K}^R\left(e^{\beta_s^T Z} + C\varepsilon\right) + 1/\sigma^2} \times 1/\sigma^2 \times Z - m_1\left(\beta_0, \Lambda_0, \sigma^2\right) \\ m_{1,i,s}^R = & \frac{\left[\mathbb{N}_K - \Lambda_{i,K}^L\left(e^{\beta_s^T Z} - C\varepsilon\right)\right]}{\Lambda_{i,K}^L\left(e^{\beta_s^T Z} - C\varepsilon\right) + 1/\sigma^2} \times 1/\sigma^2 \times Z - m_1\left(\beta_0, \Lambda_0, \sigma^2\right) \\ f_{1,i,s} = & m_{1,i,s}^R - m_{1,i,s}^L = \frac{\left(\mathbb{N}_K + 1/\sigma^2\right) \times 1/\sigma^2 \times Z}{\left(\Lambda_{i,K}^R\left(e^{\beta_s^T Z} + C\varepsilon\right) + 1/\sigma^2\right)\left(\Lambda_{i,K}^L\left(e^{\beta_s^T Z} - C\varepsilon\right) + 1/\sigma^2\right)} \\ & \times \left[\left(\Lambda_{i,K}^R - \Lambda_{i,K}^L\right) e^{\beta_s^T Z} + C\left(\Lambda_{i,K}^R + \Lambda_{i,K}^L\right)\varepsilon\right] \end{aligned}$$

And

$$\begin{aligned} P\left|f_{1,i,s}\right|^2 = & CP\left[\left(\Lambda_{K,i}^R - \Lambda_{K,i}^L\right) e^{\beta_s^T Z} + \left(\Lambda_{K,i}^R + \Lambda_{K,i}^L\right) C\varepsilon\right]^2 \\ \leq & CP\left[\left(\Lambda_{K,i}^R - \Lambda_{K,i}^L\right)^2 e^{2\beta_s^T Z} + C\left(\Lambda_{K,i}^R + \Lambda_{K,i}^L\right)^2 \varepsilon^2\right] \\ \leq & CP\left[C_1 \sum_{j=1}^{K}\left(\Delta\Lambda_{j,i}^R - \Delta\Lambda_{j,i}^L\right)^2\right] + C_2\varepsilon^2 \\ \leq & C\varepsilon^2 \end{aligned}$$

Therefore with fixed σ^2, $\mathcal{M}_1$ has a bracket number $Cexp\,(1/\varepsilon)\cdot(1/\varepsilon)^d$. Now we allow σ^2 to vary freely. By Condition 2-4, $\frac{\partial}{\partial\sigma^2}\left[m_1\left(\beta,\Lambda,\sigma^2\right)-m_1\left(\beta_0,\Lambda_0,\sigma^2\right)\right]$ is uniformly bounded. If $m_1\left(\beta,\Lambda,\sigma^2\right)-m_1\left(\beta_0,\Lambda_0,\sigma^2\right)$ is contained in a bracket $[l,u]$, then $m_1\left(\beta,\Lambda,\tilde\sigma^2\right)-m_1\left(\beta_0,\Lambda_0,\tilde\sigma^2\right)$ is contained in the bracket $[l-\varepsilon,u+\varepsilon]$ for $\tilde\sigma^2$ with $|\tilde\sigma^2-\sigma^2|\le\varepsilon$. With the compactness of the parameter space of σ^2, we can select an ε-net, $\left\{\sigma_1^2,\sigma_2^2,\cdots,\sigma_q^2\right\}, q=O(1/\varepsilon)$ over $\mathcal{R}$ and construct brackets for each σ_i^2 with enlarged bracket size. So the total number of brackets of $\mathcal{M}_1$ is $Cexp\,(1/\varepsilon)\cdot(1/\varepsilon)^d$ and it is Donsker.

Also for any $d_1\left(\beta,\Lambda,\sigma^2\right)\in\mathcal{M}_1$

$$d_1\left(\beta,\Lambda,\sigma^2\right)=\frac{\left(\mathbb{N}_K+1/\sigma^2\right)\times 1/\sigma^2\times Z}{\left(\Lambda_K e^{\beta^TZ}+1/\sigma^2\right)\left(\Lambda_{0,K}e^{\beta_0^TZ}+1/\sigma^2\right)}\times\left(\Lambda_{0,K}e^{\beta_0^TZ}-\Lambda_K e^{\beta^TZ}\right)$$

$$\begin{aligned}P\left|d_1\left(\beta,\Lambda,\sigma^2\right)\right|^2 &\le CP\left(\Lambda_{0,K}e^{\beta_0^TZ}-\Lambda_K e^{\beta^TZ}\right)^2\\ &=CP\left(\sum_{j=1}^K\Delta\Lambda_{0,j}e^{\beta_0^TZ}-\sum_{j=1}^K\Delta\Lambda_j e^{\beta^TZ}\right)^2\\ &=CP\left(\sum_{j=1}^K\left(\Delta\Lambda_{0,j}-\Delta\Lambda_j\right)e^{\beta_0^TZ}+\sum_{j=1}^K\Delta\Lambda_j\left(e^{\beta_0^TZ}-e^{\beta^TZ}\right)\right)^2\\ &\le C_1P\sum_{j=1}^K\left(\Delta\Lambda_{0,j}-\Delta\Lambda_j\right)^2+C_2\delta^2=C\delta^2\to 0 \text{ as } \delta\to 0\end{aligned}$$

By Corollary 2.3.12 of van der Vaart & Wellner (1996) (Lemma 3.3), this implies

$$\sup_{|\beta-\beta_0|\le\delta_n,\|\Lambda-\Lambda_0\|\le Cn^{\nu},\sigma^2\in\mathcal{R}^+}\left|\mathbb{G}_n d_1\left(\beta,\Lambda,\sigma^2;X\right)\right|=o_p(1)$$

where ν can be chosen as the convergence rate shown in Section 3.3.

Similarly, define a class,

$$\mathcal{M}_2 = \left\{ m_2\left(\beta, \Lambda, \sigma^2\right)\left[h^*_{\sigma^2}\right] - m_2\left(\beta_0, \Lambda_0, \sigma^2\right)\left[h^*_{\sigma^2}\right] : \|\beta - \beta_0\| < \delta, \|\Lambda - \Lambda_0\| < \delta, \right.$$

$$\left. \beta \in \mathcal{R}^d, \Lambda \in \mathcal{F}, \sigma^2 \in \mathcal{R}^+ \right\}$$

For our specific likelihood we have

$$\begin{aligned} & m_2\left(\beta, \Lambda, \sigma^2\right)\left[h^*_{\sigma^2}\right] - m_2\left(\beta_0, \Lambda_0, \sigma^2\right)\left[h^*_{\sigma^2}\right] \\ = & \left(\sum_{j=1}^{K} \frac{\Delta \mathbb{N}_j}{\Delta \Lambda_j} \triangle\Lambda_{0,j} - \frac{\mathbb{N}_K + 1/\sigma^2}{\Lambda_K e^{\beta^T Z} + 1/\sigma^2} \Lambda_{0,K} e^{\beta^T Z} \right) \cdot S - \frac{\mathbb{N}_K - \Lambda_{0,K} e^{\beta_0^T Z}}{\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^2} 1/\sigma^2 \cdot S \end{aligned}$$

We first discuss the case when σ^2 is fixed. Using a similar argument, $\mathcal{M}_2$ is covered by $\left[m^L_{2,i,s}, m^R_{2,i,s}\right]$ where

$$\begin{aligned} m^L_{2,i,s} = & \left[\sum_{j=1}^{K} \frac{\Delta \mathbb{N}_j}{\Delta \Lambda^R_{i,j}} \triangle\Lambda_{0,j} - \frac{\mathbb{N}_K + 1/\sigma^2}{\Lambda^L_{i,K}\left(e^{\beta_s^T Z} - C\varepsilon\right) + 1/\sigma^2} \Lambda_{0,K}\left(e^{\beta_s^T Z} + C\varepsilon\right) \right] \times S - \\ & m_2\left(\beta_0, \Lambda_0, \sigma^2\right)\left[h^*_{\sigma^2}\right] \\ m^R_{2,i,s} = & \left[\sum_{j=1}^{K} \frac{\Delta \mathbb{N}_j}{\Delta \Lambda^L_{i,j}} \triangle\Lambda_{0,j} - \frac{\mathbb{N}_K + 1/\sigma^2}{\Lambda^L_{i,K}\left(e^{\beta_s^T Z} + C\varepsilon\right) + 1/\sigma^2} \Lambda_{0,K}\left(e^{\beta_s^T Z} - C\varepsilon\right) \right] \times S - \\ & m_2\left(\beta_0, \Lambda_0, \sigma^2\right)\left[h^*_{\sigma^2}\right] \end{aligned}$$

And

$$\begin{aligned} f_{2,i,s} = & \, m^R_{2,i,s} - m^L_{2,i,s} \\ = & \, S\left\{ \sum_{j=1}^{K} \Delta\mathbb{N}_j \triangle\Lambda_{0,j}\left(\frac{1}{\Delta\Lambda^L_{i,j}} - \frac{1}{\Delta\Lambda^R_{i,j}}\right) + \left(\mathbb{N}_K + 1/\sigma^2\right)\Lambda_{0,K} e^{\beta_s^T Z} \times \right. \\ & \left(\frac{1}{\Lambda^L_{i,K}\left(e^{\beta_s^T Z} - C\varepsilon\right) + 1/\sigma^2} - \frac{1}{\Lambda^L_{i,K}\left(e^{\beta_s^T Z} + C\varepsilon\right) + 1/\sigma^2} \right) + \\ & \left. \left(\mathbb{N}_K + 1/\sigma^2\right)\Lambda_{0,K}\left(\frac{1}{\Lambda^L_{i,K}\left(e^{\beta_s^T Z} - C\varepsilon\right) + 1/\sigma^2} + \frac{1}{\Lambda^L_{i,K}\left(e^{\beta_s^T Z} + C\varepsilon\right) + 1/\sigma^2} \right) C\varepsilon \right\} \\ \leq & \, C_1 \mathbb{N}_K \sum_{j=1}^{K}\left(\Delta\Lambda^R_{i,j} - \Delta\Lambda^L_{i,j}\right) + C_2\left(\Lambda^R_{i,K} - \Lambda^L_{i,K}\right) + C_3\varepsilon \end{aligned}$$

So we have

$$
\begin{aligned}
|f_{2,i,s}|^2 &\leq C_1 \mathbb{N}_K^2 \sum_{j=1}^{K} \left(\Delta\Lambda_{i,j}^R - \Delta\Lambda_{i,j}^L\right)^2 + C_2\varepsilon^2 \\
P|f_{2,i,s}|^2 &\leq C_1 P\left(\sum_{j=1}^{K} \left(\Delta\Lambda_{i,j}^R - \Delta\Lambda_{i,j}^L\right)^2\right) + C_2\varepsilon^2 \leq C\varepsilon^2
\end{aligned}
$$

Again, if we allow σ^2 vary across $\mathcal{R}^+$, the bracket number with an enlarged bracket size is bounded by $Cexp\left(1/\varepsilon\right) \cdot \left(1/\varepsilon\right)^{d+1}$ and hence $\mathcal{M}_2$ is P-Donsker. For any $d_2\left(\beta, \Lambda, \sigma^2\right) \in \mathcal{M}_2$, we have

$$
\begin{aligned}
& d_2\left(\beta, \Lambda, \sigma^2\right) \\
=& m_2\left(\beta, \Lambda, \sigma^2\right)\left[h^*_{\sigma^2}\right] - m_2\left(\beta_0, \Lambda_0, \sigma^2\right)\left[h^*_{\sigma^2}\right] \\
=& S\left\{\left(\sum_{j=1}^{K} \frac{\Delta\mathbb{N}_j}{\Delta\Lambda_j}\Delta\Lambda_{0,j} - \frac{\mathbb{N}_K + 1/\sigma^2}{\Lambda_K e^{\beta^T Z} + 1/\sigma^2}\Lambda_{0,K} e^{\beta_0^T Z}\right) - \right. \\
& \left.\left(\sum_{j=1}^{K} \frac{\Delta\mathbb{N}_j}{\Delta\Lambda_{0,j}}\Delta\Lambda_{0,j} - \frac{\mathbb{N}_K + 1/\sigma^2}{\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^2}\Lambda_{0,K} e^{\beta_0^T Z}\right)\right\} \\
=& S\left\{\sum_{j=1}^{K} \Delta\mathbb{N}_j \Delta\Lambda_{0,j} \frac{\Delta\Lambda_{0,j} - \Delta\Lambda_j}{\Delta\Lambda_{0,j}\Delta\Lambda_j} + \right. \\
& \left. \left(\mathbb{N}_K + 1/\sigma^2\right)\Lambda_{0,K} e^{\beta_0^T Z} \frac{\Lambda_K e^{\beta^T Z} - \Lambda_{0,K} e^{\beta_0^T Z}}{\left(\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^2\right)\left(\Lambda_K e^{\beta^T Z} + 1/\sigma^2\right)}\right\} \\
\leq& C_2 \sum_{j=1}^{K} \left(\Delta\Lambda_{0,j} - \Delta\Lambda_j\right) + C_2\left(\Lambda_K e^{\beta^T Z} - \Lambda_{0,K} e^{\beta_0^T Z}\right) \\
=& C_2 \sum_{j=1}^{K} \left(\Delta\Lambda_{0,j} - \Delta\Lambda_j\right) + C_2\left[\left(\Lambda_K - \Lambda_{0,K}\right) e^{\beta^T Z} + \Lambda_{0,K}\left(e^{\beta^T Z} - e^{\beta_0^T Z}\right)\right] \\
\leq& C_2 \sum_{j=1}^{K} \left(\Delta\Lambda_{0,j} - \Delta\Lambda_j\right) + C_2\left(\beta - \beta_0\right)^T Z.
\end{aligned}
$$

Therefore,

$$P|d_2\left(\beta,\Lambda,\sigma^2\right)|^2 \leq C_2 P\left(\sum_{j=1}^{K}\left(\Delta\Lambda_j - \triangle\Lambda_{0,j}\right)^2\right) + C_2 P\left(\left(\beta-\beta_0\right)^T ZZ^T\left(\beta-\beta_0\right)\right)$$

$$\leq C\delta^2 \longrightarrow 0 \text{ as } \delta \to 0$$

By the Semi-equicontinuity Theorem (Lemma 3.5), this further implies,

$$\sup_{|\beta-\beta_0|\leq\delta_n, \|\Lambda-\Lambda_0\|\leq Cn^{\nu}, \sigma^2\in\mathcal{R}^+}\left|\mathbb{G}_n d_2\left(\beta,\Lambda,\sigma^2;X\right)\right| = o_p(1)$$

where ν can be chosen as the convergence rate shown in Section 3.3.

6. By Taylor expansion of $m_1\left(\beta,\Lambda,\sigma^2;X\right)$ at the point (β_0,Λ_0), we have

$$m_1\left(\beta,\Lambda,\sigma^2;X\right) =$$

$$\begin{aligned}
&m_1\left(\beta_0,\Lambda_0,\sigma^2;X\right) + m_{11}\left(\beta_0,\Lambda_0,\sigma^2;X\right)\left(\beta-\beta_0\right) + m_{12}\left(\beta_0,\Lambda_0,\sigma^2;X\right)\left[\Lambda-\Lambda_0\right]\\
&-\frac{1}{2}Z\left(\mathbb{N}_k+1/\sigma^2\right)1/\sigma^2\frac{1/\sigma^2-\Lambda_{0,K}e^{\beta_0^T Z}}{\left(\Lambda_{0,K}e^{\beta_0^T Z}+1/\sigma^2\right)^3}\Lambda_{0,K}e^{\beta_\xi^T Z}\left(\beta-\beta_0\right)^T ZZ^T\left(\beta-\beta_0\right)\\
&-\left(\mathbb{N}_k+1/\sigma^2\right)1/\sigma^2\frac{1/\sigma^2-\Lambda_{0,K}e^{\beta_0^T Z}}{\left(\Lambda_{0,K}e^{\beta_0^T Z}+1/\sigma^2\right)^3}e^{\beta_\xi^T Z}\left(\beta-\beta_0\right)^T ZZ^T\left(\Lambda_K-\Lambda_{0,K}\right)\\
&+\frac{\left(\mathbb{N}_k+1/\sigma^2\right)}{\left(\Lambda_{\xi,K}e^{\beta_0^T Z}+1/\sigma^2\right)^3}e^{2\beta_0^T Z}1/\sigma^2 Z\left(\Lambda_K-\Lambda_{0,K}\right)^2
\end{aligned}$$

So,

$$\begin{aligned}
&P\left|m_1\left(\beta,\Lambda,\sigma^2;X\right)-m_1\left(\beta_0,\Lambda_0,\sigma^2;X\right)-\right.\\
&\left.m_{11}\left(\beta_0,\Lambda_0,\sigma^2;X\right)\left(\beta-\beta_0\right)-m_{12}\left(\beta_0,\Lambda_0,\sigma^2;X\right)\left[\Lambda-\Lambda_0\right]\right|\\
=&P\left|-\frac{1}{2}Z\left(\mathbb{N}_k+1/\sigma^2\right)1/\sigma^2\frac{1/\sigma^2-\Lambda_{0,K}e^{\beta_0^T Z}}{\left(\Lambda_{0,K}e^{\beta_0^T Z}+1/\sigma^2\right)^3}\Lambda_{0,K}e^{\beta_\xi^T Z}\left(\beta-\beta_0\right)^T ZZ^T\left(\beta-\beta_0\right)\right.\\
&-\left(\mathbb{N}_k+1/\sigma^2\right)1/\sigma^2\frac{1/\sigma^2-\Lambda_{0,K}e^{\beta_0^T Z}}{\left(\Lambda_{0,K}e^{\beta_0^T Z}+1/\sigma^2\right)^3}\Lambda_{0,K}e^{\beta_\xi^T Z}\left(\beta-\beta_0\right)^T ZZ^T\left(\Lambda_K-\Lambda_{0,K}\right)
\end{aligned}$$

$$
\begin{aligned}
&+\left.\frac{\left(\mathbb{N}_K + 1/\sigma^2\right)}{\left(\Lambda_{\xi,K} e^{\beta_0^T Z} + 1/\sigma^2\right)^3} e^{2\beta_0^T Z} 1/\sigma^2 Z \left(\Lambda_K - \Lambda_{0,K}\right)^2 \right| \\
\leq & C \left\{ \left|\beta - \beta_0\right|^2 + \left\|\Lambda - \Lambda_0\right\|^2 \right\} \qquad \text{(by conditions 1, 2, and 5)}
\end{aligned}
$$

where $\beta_\xi = \beta_0 + \xi\left(\beta - \beta_0\right)$ and $\Lambda_{\xi,j} = \Lambda_{0,j} + \xi\left(\Lambda_j - \Lambda_{0,j}\right)$ for some $0 \leq \xi \leq 1$.

Similarly,

$$
\begin{aligned}
&m_2\left(\beta, \Lambda, \sigma^2; X\right) = m_2\left(\beta_0, \Lambda_0, \sigma^2; X\right)\left[h^*_{\sigma^2}\right] + m_{21}\left(\beta_0, \Lambda_0, \sigma^2; X\right)\left[h^*_{\sigma^2}\right]\left(\beta - \beta_0\right) + \\
&\quad m_{22}\left(\beta_0, \Lambda_0, \sigma^2; X\right)\left[h^*_{\sigma^2}, \Lambda - \Lambda_0\right] \\
&\quad - \frac{1}{2}\left(\mathbb{N}_K + 1/\sigma^2\right) 1/\sigma^2 \frac{1/\sigma^2 - \Lambda_{0,K} e^{\beta_0^T Z}}{\left(\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^2\right)^3} e^{\beta_\xi^T Z} \left(\beta - \beta_0\right)^T Z Z^T \left(\beta - \beta_0\right) h^*_{\sigma^2,K} \\
&\quad - 2\left(\mathbb{N}_K + 1/\sigma^2\right) 1/\sigma^2 \frac{e^{\beta_0^T Z}}{\left(\Lambda_{\xi,K} e^{\beta_0^T Z} + 1/\sigma^2\right)^3} Z \left(\beta - \beta_0\right)\left(\Lambda - \Lambda_0\right) h^*_{\sigma^2,K} \\
&\quad + \sum_{j=1}^{K} \frac{\Delta \mathbb{N}_j}{\Delta \Lambda_{\zeta,j}^3} \Delta h^*_{\sigma^2,j} \left(\Delta\Lambda_j - \Delta\Lambda_{0,j}\right)^2 - \frac{\mathbb{N}_K + 1/\sigma^2}{\left(\Lambda_{\zeta,K} e^{\beta_0^T Z} + 1/\sigma^2\right)^3} h^*_{\sigma^2,K} e^{3\beta_\xi^T Z} \left(\Lambda_K - \Lambda_{0,K}\right)^2
\end{aligned}
$$

So,

$$
\begin{aligned}
&P\left|m_2\left(\beta, \Lambda, \sigma^2; X\right)\left[h^*_{\sigma^2}\right] - m_2\left(\beta_0, \Lambda_0, \sigma^2; X\right)\left[h^*_{\sigma^2}\right] - \right. \\
&\left. m_{21}\left(\beta_0, \Lambda_0, \sigma^2; X\right)\left[h^*_{\sigma^2}\right]\left(\beta - \beta_0\right) + m_{22}\left(\beta_0, \Lambda_0, \sigma^2; X\right)\left[h^*_{\sigma^2}, \Lambda - \Lambda_0\right]\right| \\
=&P\left| - \frac{1}{2}\left(\mathbb{N}_K - 1/\sigma^2\right) 1/\sigma^2 \frac{1/\sigma^2 - \Lambda_{0,K} e^{\beta_0^T Z}}{\left(\Lambda_{0,K} e^{\beta_0^T Z} + 1/\sigma^2\right)^3} e^{\beta_\xi^T Z} \left(\beta - \beta_0\right)^T Z h^*_{\sigma^2} \left(\beta - \beta_0\right) \right. \\
&- 2\left(\mathbb{N}_K + 1/\sigma^2\right) 1/\sigma^2 \frac{2 e^{\beta_0^T Z}}{\left(\Lambda_{\zeta,K} e^{\beta_0^T Z} + 1/\sigma^2\right)^3} Z Z^T \left(\beta - \beta_0\right)\left(\Lambda - \Lambda_0\right) h^*_{\sigma^2,K} \\
&+ \sum_{j=1}^{K} \frac{\Delta \mathbb{N}_j}{\Delta \Lambda_{\zeta,j}^3} \Delta h^*_{\sigma^2,j} \left(\Delta\Lambda_j - \Delta\Lambda_{0,j}\right)^2 - \\
&\left. \frac{\mathbb{N}_K + 1/\sigma^2}{\left(\Lambda_{\zeta,K} e^{\beta_0^T Z} + 1/\sigma^2\right)^3} h^*_{\sigma^2,K} e^{3\beta_\xi^T Z} \left(\Lambda_K - \Lambda_{0,K}\right)^2 \right| \\
\leq & C \left\{ \left|\beta - \beta_0\right|^2 + \left\|\Lambda - \Lambda_0\right\|^2 \right\} \qquad \text{(by conditions 1, 2, and 5)}
\end{aligned}
$$

where $\beta_\xi = \beta_0 + \zeta\left(\beta - \beta_0\right)$ and $\Lambda_{\zeta,j} = \Lambda_{0,j} + \zeta\left(\Lambda_j - \Lambda_{0,j}\right)$ for some $0 \leq \zeta \leq 1$.

7. Since

$$m_1\left(\beta_0, \Lambda_0, \sigma^2; X\right) = \frac{\mathbb{N}_K - \Lambda_{0,K} e^{\beta_0^T Z}}{\sigma^2 \Lambda_{0,K} e^{\beta_0^T Z} + 1} \times Z$$
$$\frac{\partial}{\partial \sigma^2} m_1\left(\beta_0, \Lambda_0, \sigma^2; X\right) = -\frac{\mathbb{N}_K - \Lambda_{0,K} e^{\beta_0^T Z}}{\left(\sigma^2 \Lambda_{0,K} e^{\beta_0^T Z} + 1\right)^2} \Lambda_{0,K} e^{\beta_0^T Z} \times Z$$

Let $m_{1\sigma} = \left| \frac{\mathbb{N}_K - \Lambda_{0,K} e^{\beta_0^T Z}}{\left(\sigma^2 \Lambda_{0,K} e^{\beta_0^T Z} + 1\right)^2} \Lambda_{0,K} e^{\beta_0^T Z} \times Z \right|$, then

$$P m_{1\sigma}^4 = P\left\{ \frac{\left(\mathbb{N}_K - \Lambda_{0,K} e^{\beta_0^T Z}\right)^4}{\left(\sigma^2 \Lambda_{0,K} e^{\beta_0^T Z} + 1\right)^8} \Lambda_{0,K}^4 e^{4\beta_0^T Z} \times Z^4 \right\} < \infty$$

by the boundness of Z, Λ_0 and $E\left\{e^{C\mathbb{N}(t)}\right\}$.

Similarly

$$m_2\left(\beta_0, \Lambda_0, \sigma^2; X\right) = \frac{\mathbb{N}_K - \Lambda_{0,K} e^{\beta_0^T Z}}{\sigma^2 \Lambda_{0,K} e^{\beta_0^T Z} + 1} \times S$$
$$\frac{\partial}{\partial \sigma^2} m_2\left(\beta_0, \Lambda_0, \sigma^2; X\right) = -\frac{\mathbb{N}_K - \Lambda_{0,K} e^{\beta_0^T Z}}{\left(\sigma^2 \Lambda_{0,K} e^{\beta_0^T Z} + 1\right)^2} \Lambda_{0,K} e^{\beta_0^T Z} \times S + \frac{\mathbb{N}_K - \Lambda_{0,K} e^{\beta_0^T Z}}{\sigma^2 \Lambda_{0,K} e^{\beta_0^T Z} + 1} \frac{\partial S}{\partial \sigma^2}$$

Let $m_{2\sigma} = \left| -\frac{\mathbb{N}_K - \Lambda_{0,K} e^{\beta_0^T Z}}{\left(\sigma^2 \Lambda_{0,K} e^{\beta_0^T Z} + 1\right)^2} \Lambda_{0,K} e^{\beta_0^T Z} \times S + \frac{\mathbb{N}_K - \Lambda_{0,K} e^{\beta_0^T Z}}{\left(\sigma^2 \Lambda_{0,K} e^{\beta_0^T Z} + 1\right)^2} \frac{\partial S}{\partial \sigma^2} \right|$, then

$$\begin{aligned} P m_{2\sigma}^4 \leq & C P\left\{ \frac{\left(\mathbb{N}_K - \Lambda_{0,K} e^{\beta_0^T Z}\right)^4}{\left(\sigma^2 \Lambda_{0,K} e^{\beta_0^T Z} + 1\right)^8} \Lambda_{0,K}^4 e^{4\beta_0^T Z} \times S^4 \right\} + \\ & C P\left\{ \frac{\left(\mathbb{N}_K - \Lambda_{0,K} e^{\beta_0^T Z}\right)^4}{\left(\sigma^2 \Lambda_{0,K} e^{\beta_0^T Z} + 1\right)^8} \left(\frac{\partial S}{\partial \sigma^2}\right)^4 \right\} \\ = & C\left(I_1 + I_2\right) \\ I_1 = & P\left\{ \frac{\left(\mathbb{N}_K - \Lambda_{0,K} e^{\beta_0^T Z}\right)^4}{\left(\sigma^2 \Lambda_{0,K} e^{\beta_0^T Z} + 1\right)^8} \Lambda_{0,K}^4 e^{4\beta_0^T Z} \times S^4 \right\} \end{aligned}$$

Again, by the boundness of Z, Λ_0, $I_1 < \infty$.

$$I_2 = P\left\{ \frac{\left(\mathbb{N}_K - \Lambda_{0,K} e^{\beta_0^T Z}\right)^4}{\left(\sigma^2 \Lambda_{0,K} e^{\beta_0^T Z} + 1\right)^8} \left(\frac{\partial S}{\partial \sigma^2}\right)^4 \right\}$$

$$\frac{\partial S}{\partial \sigma^2} = -\frac{E\left(\frac{e^{\beta_0^T Z}\cdot Z}{\left(\sigma^2 \Lambda_{0,K} e^{\beta_0^T Z}+1\right)^2} \Lambda_{0,K} e^{\beta_0^T Z} | K, T\right)}{E\left(\frac{e^{\beta_0^T Z}\cdot Z}{\sigma^2 \Lambda_{0,K} e^{\beta_0^T Z}+1} | K, T\right)} + \frac{E\left(\frac{e^{\beta_0^T Z}\cdot Z}{\sigma^2 \Lambda_{0,K} e^{\beta_0^T Z}+1} | K, T\right)}{\left(E\left(\frac{e^{\beta_0^T Z}}{\sigma^2 \Lambda_{0,K} e^{\beta_0^T Z}+1} | K, T\right)\right)^2} \times E\left(\frac{e^{\beta_0^T Z}}{\left(\sigma^2 \Lambda_{0,K} e^{\beta_0^T Z}+1\right)^2} \Lambda_{0,K} e^{\beta_0^T Z} | K, T\right)$$

Let

$$J_1 = \frac{E\left(\frac{e^{\beta_0^T Z}\cdot Z}{\left(\sigma^2 \Lambda_{0,K} e^{\beta_0^T Z}+1\right)^2} \Lambda_{0,K} e^{\beta_0^T Z} | K, T\right)}{E\left(\frac{e^{\beta_0^T Z}\cdot Z}{\sigma^2 \Lambda_{0,K} e^{\beta_0^T Z}+1} | K, T\right)},$$

$$J_2 = \frac{E\left(\frac{e^{\beta_0^T Z}\cdot Z}{\sigma^2 \Lambda_{0,K} e^{\beta_0^T Z}+1} | K, T\right)}{\left(E\left(\frac{e^{\beta_0^T Z}}{\sigma^2 \Lambda_{0,K} e^{\beta_0^T Z}+1} | K, T\right)\right)^2}, \text{ and}$$

$$J_3 = E\left(\frac{e^{\beta_0^T Z}}{\left(\sigma^2 \Lambda_{0,K} e^{\beta_0^T Z}+1\right)^2} \Lambda_{0,K} e^{\beta_0^T Z} | K, T\right).$$

By the boundness of Z, Λ and σ^2, J_1, J_2 and J_3 are all bounded. And $\left(\frac{\partial S}{\partial \sigma^2}\right)^4 \leq C J_1^4 + C\left(J_2^4 \times J_3^4\right)$. Together with the boundness of K and T, this further implies I_2 is bounded.

$\square$

Note: In the simulation studies in Chapter 6, Zeger's method of moment is adopted in the estimation of the overdispersion paraemter. The definition of σ_0^2 and the consistency of $\hat{\sigma}_n^2$ is delegated in Section 5.2.

CHAPTER 4
VARIANCE ESTIMATION OF THE SPLINE-BASED SIEVE GEE ESTIMATOR

In Chapter 2, we show that the spline-based sieve GEE using either $V_1^{(i)}$, $V_2^{(i)}$ or $V_3^{(i)}$ coincide with the scores of different 'likelihood' functions. The asymptotic normality of the estimated regression parameter calculated from the sieve GEE is correspondingly established in Chapter 3. A consistent estimator of the asymptotic standard error of the sieve GEE estimator is needed to make inferences.

Three different methods are discussed to estimate the asymptotic standard error. Section 4.1 presents a projection method based on the general theorem for the maximum likelihood estimate of the finite dimensional parameter in the presence of a nuisance infinite-dimensional parameter. Different from the sieve GEE estimates using $V_1^{(i)}$ or $V_2^{(i)}$, the estimate from the sieve GEE using $V_3^{(i)}$ involves an extra over-dispersion parameter σ^2. Replacing σ^2 by its consistent estimate still provides a consistent estimate of the standard error.

Section 4.2 presents an ad hoc estimator of the standard error based on the ordinary sandwich formula in parametric GEE model. The spline coefficients are treated the same as the parametric regression parameters. Simulation results from Chapter 6 show the estimates based on GEE sandwich formula provide similar result as the estimates based on the projection algorithm from Section 4.1. Computationally, the sandwich estimator provides an easier standard error estimate for the spline-based sieve GEE estimator.

Spline-base sieve approximation largely reduces the dimension of the estimation, which makes it feasible to estimate the standard error of the estimated regression parameter using the bootstrap method. Section 4.3 briefly describes the bootstrap estimate of the standard error.

4.1 Projection Method

It is shown in Section 3.3 that the spline-based sieve GEE estimate of β_0, $\hat{\beta}_n$ satisfies

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \rightarrow_d N\left(0, A_0^{-1} B_0 \left(A_0^{-1}\right)^T\right)$$

Where

$$A_0 = A\left(\beta_0, \Lambda_0, \sigma_0^2\right) = -E\left(m_{11}\left(\beta_0, \Lambda_0, \sigma_0^2; X\right) - m_{21}\left(\beta_0, \Lambda_0, \sigma_0^2; X\right)[h^*_{\sigma_0^2}]\right)$$

$$B_0 = B\left(\beta_0, \Lambda_0, \sigma_0^2\right) = E\left(m_1\left(\beta_0, \Lambda_0, \sigma_0^2; X\right) - m_2\left(\beta_0, \Lambda_0, \sigma_0^2; X\right)[h^*_{\sigma_0^2}]\right)^{\otimes 2},$$

with $h^*_{\sigma_0^2} = \left(h^*_{\sigma_0^2,1}, \cdots, h^*_{\sigma_0^2,d}\right)^T$, $h^*_{\sigma_0^2,j} \in \mathcal{H}$ for $j = 1, \cdots, d$ satisfies the equation

$$P\left(m_{12}\left(\beta_0, \Lambda_0, \sigma_0^2; X\right)[h] - m_{22}\left(\beta_0, \Lambda_0, \sigma_0^2; X\right)[h^*_{\sigma_0^2}, h]\right) = 0 \quad \forall h \in \mathcal{H}$$

It is equivalent to the projection problem of solving $h^*_{\sigma_0^2,s}$ by

$$h^*_{\sigma_0^2,s} = \underset{h\in\mathcal{H}}{\operatorname{argmin}}\, P\left(m_{1,s}\left(\beta_0, \Lambda_0, \sigma_0^2; X\right) - m_2\left(\beta_0, \Lambda_0, \sigma_0^2; X\right)[h]\right)^2 \qquad \text{for } s = 1, 2, \cdots, d. \tag{4.1}$$

where $m_{1,s}$ is the s^{th} component of m_1.

To consistently estimate A_0 and B_0, we take advantage of the spline-based sieve method again and estimate each component of $h^*_{\sigma_0^2}$ by a set of linear spans of the cubic B-spline functions, e.g., $\hat{h}_{n,s} = \sum_{j=1}^{q_n} \gamma_{j,s} B_j$ for $s = 1, 2, \cdots, d$ where $\gamma_{j,s}, j = 1, \cdots, q_n$ are estimated by minimizing the empirical version of Equation (4.1), namely,

$$\mathbb{P}_n \left(m_{1,s}\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2; X\right) - m_2\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2; X\right)[\hat{h}_{n,s}] \right)^2,$$

where $\hat{\beta}_n, \hat{\Lambda}_n$ and $\hat{\sigma}_n^2$ are consistent estimates of β_0, Λ_0 and σ_0^2, respectively. Since m_2 is a bilinear operator, it is equivalent to solving a least square problem and the solution of $\underline{\gamma}_s = (\gamma_{1,s}, \gamma_{2,s}, \cdots, \gamma_{q_n,s})^T$ is given by

$$\begin{aligned} &\left(m_2^T\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2; X\right)[B] \times m_2\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2; X\right)[B] \right)^{-1} \times \\ &\qquad \left(m_2^T\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2; X\right)[B] \times m_{1,s}\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2; X\right) \right) \end{aligned}$$

where $m_2\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2; X\right)[B]$ is the $n \times q_n$ design matrix with $(i, m)^{th}$ entry being

$$\sum_{j=1}^{K_i} \frac{\Delta \mathbb{N}_{K_i,j}^{(i)}}{\Delta \Lambda_{K_i,j}^{(i)}} \Delta B_{m,K_i,j} - \frac{\mathbb{N}_{K_i,K_i}^{(i)}}{\Lambda_{K_i,K_i}^{(i)}} e^{\hat{\beta}_n^T Z_i} B_{m,K_i,j}$$

where $\Delta \mathbb{N}_{K_i,j}^{(i)}$ and $\Delta \Lambda_{K_i,j}^{(i)}$ are defined as same as in Equation (2.7), $B_{m,K_i,j} = B_m\left(T_{K_i,j}^{(i)}\right)$ and $\Delta B_{m,K_i,j} = B_{m,K_i,j} - B_{m,K_i,j-1}$ for $m = 1, 2, \cdots, q_n$. With this estimate of $h_{\sigma_0^2}^*$ we can empirically construct A and B respectively and show they are consistent.

Theorem 4.1. *Let* $\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2\right)$ *be a consistent estimate of* $(\beta_0, \Lambda_0, \sigma_0^2)$ *and* $\hat{h}_n = \left(\hat{h}_{n,1}, \hat{h}_{n,2}, \cdots, \hat{h}_{n,d}\right)^T$. *Under regularity conditions 1, 2 and 5,* $\hat{h}_n$ *is a consistent estimate of* $h_{\sigma_0^2}^*$. *Denote*

$$\begin{aligned} \hat{A}_n &= -\mathbb{P}_n \left(m_{12}\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2; X\right) - m_{22}\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2; X\right)[\hat{h}_n] \right) \\ \hat{B}_n &= \mathbb{P}_n \left(m_1\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2; X\right) - m_2\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2; X\right)[\hat{h}_n] \right)^{\otimes 2}. \end{aligned}$$

Then $\hat{A}_n \to_p A_0$ *and* $\hat{B}_n \to_p B_0$.

Proof. Denote $\rho_s\left(\beta,\Lambda,\sigma^2,h;X\right)=\left(m_{1,s}\left(\beta,\Lambda,\sigma^2;X\right)-m_2\left(\beta,\Lambda,\sigma^2,h;X\right)[h]\right)^2, s=1,2,\cdots,d$. First, we show a class of function $\mathfrak{G}=\{\rho_s\left(\beta,\Lambda,\sigma^2,h;X\right):\beta\in\mathcal{R}^d, log\Lambda\in\psi_{l,t},\sigma^2\in\mathcal{R}^+,h\in\phi_{l,t}\}$, is Glivenko-Cantelli by evaluating its bracket number with $L_1(\mathbb{P}_n)$ norm.

For the moment, we fix σ^2. By Lemma 3.8, $\psi_{l,t}$ is covered by

$$\left\{[\Lambda_i^L,\Lambda_i^R], i=1,\cdots,O\left(q_n^{1/2}/\varepsilon\right)^{cq_n}\right\}$$

and $\left\|\Lambda_i^R-\Lambda_i^L\right\|_{L_1(\mu)}=\int\left(\Lambda_i^R(t)-\Lambda_i^L(t)\right)d\mu(t)<\varepsilon$. Similarly we can construct a set of brackets $\left\{[h_l^L,h_l^R]:l=1,\cdots,O\left(q_n^{1/2}/\varepsilon\right)^{cq_n}\right\}$ and

$$\left\|h_l^R-h_l^L\right\|_{L_1(\mu)}=\int\left(h_l^R(t)-h_l^L(t)\right)d\mu(t)<\varepsilon$$

such that $\forall h\in\phi_{l,t}, h_l^L\le h\le h_l^R$ for some l. We can also construct an $\varepsilon-$net, $\{\beta_1,\beta_2,\cdots,\beta_p\}, p=O\left(1/\varepsilon^d\right)$ such that $\forall\beta\in\mathcal{R}^d,\exists s\in\{1,2,\cdots,p\}$ such that $\left|\beta^TZ-\beta_s^TZ\right|\le\varepsilon$ and $\left|exp\left(\beta^TZ\right)-exp\left(\beta_s^TZ\right)\right|\le C\varepsilon$. We further define

$$\Delta\Lambda_{i,j}^L=\Lambda_{i,j}^L-\Lambda_{i,j-1}^L;\qquad \Delta\Lambda_{i,j}^R=\Lambda_{i,j}^R-\Lambda_{i,j-1}^R;$$

$$\Delta\Lambda_{i,j}^{RL}=\Lambda_{i,j}^R-\Lambda_{i,j-1}^L;\qquad \Delta\Lambda_{i,j}^{LR}=\Lambda_{i,j}^L-\Lambda_{i,j-1}^R;$$

Following the same lines as those in Wellner & Zhang (1995), we can make these bracketing functions satisfy $\Lambda_i^R-\Lambda_i^L\le\gamma_1$ and $\Lambda_i^L\ge\gamma_2$ with $\gamma_1,\gamma_2>0$ for all $t\in[0,\tau]$ and $1\le i\le l$. And $\Delta\Lambda_{i,j}^{LR}\ge\gamma_3>0$. Similarly, we define

$$\Delta h_{l,j}^L=h_{l,j}^L-h_{l,j-1}^L;\qquad \Delta h_{l,j}^R=h_{l,j}^R-h_{l,j-1}^R;$$

$$\Delta h_{l,j}^{LR}=h_{l,j}^L-h_{l,j-1}^R;\qquad \Delta h_{l,j}^{RL}=h_{l,j}^R-h_{l,j-1}^L.$$

Let

$$
\begin{aligned}
M^L_{i,s,l} = {} & \frac{\mathbb{N}_K - \Lambda^R_{i,K}\left(e^{\beta_s^T Z} + C\varepsilon\right)}{\Lambda^R_{i,K}\left(e^{\beta_s^T Z} + C\varepsilon\right) + 1/\sigma^2} \cdot 1/\sigma^2 Z - \\
& \left[\sum_{j=1}^{K} \frac{\Delta\mathbb{N}_j}{\Delta\Lambda^{LR}_{i,j}} \Delta h^{RL}_{l,j} - \frac{\mathbb{N}_K + 1/\sigma^2}{\Lambda^R_{i,K}\left(e^{\beta_s^T Z} + C\varepsilon\right) + 1/\sigma^2} h^L_{l,K}\left(e^{\beta_s^T Z} - C\varepsilon\right)\right] \\
M^R_{i,s,l} = {} & \frac{\mathbb{N}_K - \Lambda^L_{i,K}\left(e^{\beta_s^T Z} - C\varepsilon\right)}{\Lambda^L_{i,K}\left(e^{\beta_s^T Z} - C\varepsilon\right) + 1/\sigma^2} \cdot 1/\sigma^2 Z - \\
& \left[\sum_{j=1}^{K} \frac{\Delta\mathbb{N}_j}{\Delta\Lambda^{RL}_{i,j}} \Delta h^{LR}_{l,j} - \frac{\mathbb{N}_K + 1/\sigma^2}{\Lambda^L_{i,K}\left(e^{\beta_s^T Z} + C\varepsilon\right) + 1/\sigma^2} h^R_{l,K}\left(e^{\beta_s^T Z} + C\varepsilon\right)\right]
\end{aligned}
$$

And we write $M^R_{i,s,l} - M^L_{i,s,l} = dM^1_{i,s,l} + dM^2_{i,s,l} + dM^3_{i,s,l}$, where

$$
\begin{aligned}
dM^1_{i,s,l} = {} & \left[\frac{\mathbb{N}_K - \Lambda^L_{i,K}\left(e^{\beta_s^T Z} - C\varepsilon\right)}{\Lambda^L_{i,K}\left(e^{\beta_s^T Z} - C\varepsilon\right) + 1/\sigma^2} - \frac{\mathbb{N}_K - \Lambda^R_{i,K}\left(e^{\beta_s^T Z} + C\varepsilon\right)}{\Lambda^R_{i,K}\left(e^{\beta_s^T Z} + C\varepsilon\right) + 1/\sigma^2}\right] \cdot 1/\sigma^2 Z \\
\leq {} & C\Big[\mathbb{N}_K e^{\beta_s^T Z}\left(\Lambda^R_{i,K} - \Lambda^L_{i,K}\right) + \mathbb{N}_K\left(\Lambda^R_{i,K} - \Lambda^L_{i,K}\right) C\varepsilon - \\
& \left(\Lambda^R_{i,K} - \Lambda^L_{i,K}\right) e^{\beta_s^T Z} 1/\sigma^2 + C\left(\Lambda^R_{i,K} - \Lambda^L_{i,K}\right) 1/\sigma^2 C\varepsilon\Big] \\
\leq {} & C_1\left(\Lambda^R_{i,K} - \Lambda^L_{i,K}\right) + C_2\varepsilon = C_1\sum_{j=1}^{K}\left(\Delta\Lambda^R_{i,j} - \Delta\Lambda^L_{i,j}\right) + C_2\varepsilon \\
dM^2_{i,s,l} = {} & \sum_{j=1}^{K} \frac{\Delta\mathbb{N}_j}{\Delta\Lambda^{LR}_{i,j}\Delta\Lambda^{RL}_{i,j}}\left(\Delta\Lambda^{RL}_{i,j}\Delta h^{RL}_{l,j} - \Delta\Lambda^{LR}_{i,j}\Delta h^{LR}_{l,j}\right) \\
= {} & \sum_{j=1}^{K} \frac{\Delta\mathbb{N}_j}{\Delta\Lambda^{LR}_{i,j}\Delta\Lambda^{RL}_{i,j}}\left(\Delta\Lambda^{RL}_{i,j}\left(\Delta h^{RL}_{l,j} - \Delta h^{LR}_{l,j}\right) + \Delta h^{LR}_{l,j}\left(\Delta\Lambda^{RL}_{i,j} - \Delta\Lambda^{LR}_{i,j}\right)\right) \\
\leq {} & C_1\sum_{j=1}^{K}\left(\Delta\Lambda^{RL}_{i,j} - \Delta\Lambda^{LR}_{i,j}\right) + C_2\sum_{l=1}^{K}\left(\Delta h^{RL}_{l,j} - \Delta h^{LR}_{l,j}\right) \\
\leq {} & C_1 K\sum_{j=1}^{K}\left(\Delta\Lambda^{R}_{i,j} - \Delta\Lambda^{L}_{i,j}\right) + C_2 K\sum_{l=1}^{K}\left(\Delta h^{R}_{l,j} - \Delta h^{L}_{l,j}\right) \qquad \text{(by inequality in (3.11))}
\end{aligned}
$$

$$\begin{aligned}
dM^3_{i,s,l} &= \left[\frac{h^R_{l,K}\left(e^{\beta_s^T Z}+C\varepsilon\right)}{\Lambda^L_{i,K}\left(e^{\beta_s^T Z}-C\varepsilon\right)+1/\sigma^2}-\frac{h^L_{l,K}\left(e^{\beta_s^T Z}-C\varepsilon\right)}{\Lambda^R_{i,K}\left(e^{\beta_s^T Z}+C\varepsilon\right)+1/\sigma^2}\right]\left(\mathbb{N}_K+1/\sigma^2\right) \\
&\le C\Big[\left(\Lambda^R_{i,K}h^R_{l,K}-\Lambda^L_{i,K}h^L_{l,K}\right)e^{2\beta_s^T Z}+2\left(\Lambda^R_{i,K}h^R_{l,K}+\Lambda^L_{i,K}h^L_{l,K}\right)e^{\beta_s^T Z}C\varepsilon+ \\
&\quad \left(\Lambda^R_{i,K}h^R_{l,K}-\Lambda^L_{i,K}h^L_{l,K}\right)e^{\beta_s^T Z}C\varepsilon^2+\left(h^R_{i,K}-h^L_{i,K}\right)e^{\beta_s^T Z}\cdot 1/\sigma^2+\left(h^R_{i,K}+h^L_{i,K}\right)\cdot 1/\sigma^2 C\varepsilon\Big] \\
&\le C_1\left(\Lambda^R_{i,K}-\Lambda^L_{i,K}\right)+C_2\left(h^R_{l,K}-h^L_{l,K}\right)+C_3\varepsilon \\
&\le C_1\sum_{j=1}^{K}\left(\Delta\Lambda^R_{i,j}-\Delta\Lambda^L_{i,j}\right)+C_2\sum_{j=1}^{K}\left(\Delta h^R_{l,j}-\Delta h^L_{l,j}\right)+C_3\varepsilon
\end{aligned}$$

Therefore

$$\begin{aligned}
M^R_{i,s,l}-M^L_{i,s,l} &\le C_1\sum_{j=1}^{K}\left(\Delta\Lambda^R_{i,j}-\Delta\Lambda^L_{i,j}\right)+C_2\sum_{j=1}^{K}\left(\Delta h^R_{l,j}-\Delta h^L_{l,j}\right)+C_3\varepsilon \\
\left(M^R_{i,s,l}-M^L_{i,s,l}\right)^2 &\le C_1\sum_{j=1}^{K}\left(\Delta\Lambda^R_{i,j}-\Delta\Lambda^L_{i,j}\right)^2+C_2\sum_{j=1}^{K}\left(\Delta h^R_{l,j}-\Delta h^L_{l,j}\right)^2+C_3\varepsilon^2
\end{aligned}$$

Therefore $m_{1,s}\left(\beta,\Lambda,\sigma^2,h;X\right)-m_2\left(\beta,\Lambda,\sigma^2,h;X\right)[h]$ is covered by $[M^L_{i,s,l},M^R_{i,s,l}]$. When both $M^L_{i,s,l}$ and $M^R_{i,s,l}$ are positive, $\left(m_{1,s}\left(\beta,\Lambda,\sigma^2,h;X\right)-m_2\left(\beta,\Lambda,\sigma^2,h;X\right)[h]\right)^2$ is covered by $[\left(M^L_{i,s,l}\right)^2,\left(M^R_{i,s,l}\right)^2]$.

$$\mathbb{P}_n\left(\left(M^L_{i,s,l}\right)^2-\left(M^R_{i,s,l}\right)^2\right)=\mathbb{P}_n\left(\left(M^L_{i,s,l}+M^R_{i,s,l}\right)\left(M^L_{i,s,l}-M^R_{i,s,l}\right)\right)\le C\varepsilon$$

Similarly, when both $M^L_{i,s,l}$ and $M^R_{i,s,l}$ are negative, the brackets are $[\left(M^R_{i,s,l}\right)^2,\left(M^L_{i,s,l}\right)^2]$. When $M^L_{i,s,l}<0<M^R_{i,s,l}$, the brackets are $[0,\left(-M^L_{i,s,l}\vee M^R_{i,s,l}\right)^2]$ and $\left(-M^L_{i,s,l}\vee M^R_{i,s,l}\right)^2\le\left(M^R_{i,s,l}-M^L_{i,s,l}\right)^2$. So

$$\mathbb{P}_n\left(-M^L_{i,s,l}\vee M^R_{i,s,l}\right)^2\le\mathbb{P}_n\left(M^R_{i,s,l}-M^L_{i,s,l}\right)^2\le C\varepsilon^2\le C\varepsilon$$

Therefore the bracket number of $\mathfrak{G}$ with fixed σ^2 is $O\left(1/\varepsilon\right)^{d+2Cq_n}$. Now we allow σ^2 to vary. As shown before $\frac{\partial}{\partial\sigma^2}\left[m_1\left(\beta,\Lambda,\sigma^2\right)-m_2\left(\beta,\Lambda,\sigma^2\right)[h]\right]$ is uniformly bounded.

We can find an enlarged brackets for $\tilde{\sigma}^2$ with $|\tilde{\sigma}^2-\sigma^2|\leq\varepsilon$. With the compactness of the parameter space of σ^2, we can select an ε-net, $\left\{\sigma_1^2,\sigma_2^2,\cdots,\sigma_q^2\right\},q=O(1/\varepsilon)$ over $\mathcal{R}^+$ and construct brackets for each σ_i^2 with this enlarged bracket size. So the total number of brackets of $\mathfrak{S}$ is $C\left(1/\varepsilon\right)^{d+2Cq_n+1}$. The entropy with bracketing $logN_{[]}\left(\varepsilon,\mathcal{F}_M,L_1\left(\mathbb{P}_n\right)\right)=o_p\left(n\right)$. Also $logN\left(\varepsilon,\mathcal{F}_M,L_1\left(\mathbb{P}_n\right)\right)=o_p\left(n\right)$. By Lemma 3.2 $\mathfrak{S}$ is a Glivenko-Cantelli. Similarly we can show $\widetilde{\mathfrak{S}}=\{\rho_s\left(\beta,\Lambda,\sigma^2,h^*_{\sigma^2};X\right):\beta\in\mathcal{R}^d,log\Lambda\in\mathcal{F},\sigma^2\in\mathcal{R}^+\}$, is a Glivenko-Cantelli as well.

Following the similar arguments used in the proof of consistency in Theorem 3.13 in Section 3.3, there exists a $h^*_{\sigma^2,n,s}\in\phi_{l,t}$ of order $m\geq p+2$ such that $\|h^*_{\sigma^2,n,s}-h^*_{\sigma^2,s}\|_\infty=O(n^{-p\nu})$. By definition, $\hat{h}_{n,s}=\text{argmin}_{h\in\phi_{l,t}}\mathbb{P}_n\rho_s\left(\hat{\beta}_n,\hat{\Lambda}_n,\hat{\sigma}_n^2,h;X\right)$, with the consistency of $\left(\hat{\beta}_n,\hat{\Lambda}_n,\hat{\sigma}_n^2\right)$, we have

$$\begin{aligned}
&\mathbb{P}_n\rho_s\left(\hat{\beta}_n,\hat{\Lambda}_n,\hat{\sigma}_n^2,\hat{h}_{n,s};X\right)-\mathbb{P}_n\rho_s\left(\hat{\beta}_n,\hat{\Lambda}_n,\hat{\sigma}_n^2,h^*_{\sigma^2,s};X\right)\\
=&\mathbb{P}_n\rho_s\left(\hat{\beta}_n,\hat{\Lambda}_n,\hat{\sigma}_n^2,\hat{h}_{n,s};X\right)-\mathbb{P}_n\rho_s\left(\hat{\beta}_n,\hat{\Lambda}_n,\hat{\sigma}_n^2,h^*_{\sigma^2,n,s};X\right)\\
&+\mathbb{P}_n\rho_s\left(\hat{\beta}_n,\hat{\Lambda}_n,\hat{\sigma}_n^2,h^*_{\sigma^2,n,s};X\right)-\mathbb{P}_n\rho_s\left(\hat{\beta}_n,\hat{\Lambda}_n,\hat{\sigma}_n^2,h^*_{\sigma^2,s};X\right)\\
\leq&\left(\mathbb{P}_n-P\right)\left(\rho_s\left(\hat{\beta}_n,\hat{\Lambda}_n,\hat{\sigma}_n^2,h^*_{\sigma^2,n,s};X\right)-\rho\left(\hat{\beta}_n,\hat{\Lambda}_n,\hat{\sigma}_n^2,h^*_{\sigma^2,s};X\right)\right)+\\
&P\left(\rho_s\left(\hat{\beta}_n,\hat{\Lambda}_n,\hat{\sigma}_n^2,h^*_{\sigma^2,n,s};X\right)-\rho_s\left(\hat{\beta}_n,\hat{\Lambda}_n,\hat{\sigma}_n^2,h^*_{\sigma^2,s};X\right)\right)=o_p(1).
\end{aligned}$$

This leads to

$$\begin{aligned}
&\mathbb{P}_n\rho_s\left(\hat{\beta}_n,\hat{\Lambda}_n,\hat{\sigma}_n^2,\hat{h}_{n,s};X\right)\leq\mathbb{P}_n\rho_s\left(\hat{\beta}_n,\hat{\Lambda}_n,\hat{\sigma}_n^2,h^*_{\sigma^2,s};X\right)+o_p(1)\\
=&\left(\mathbb{P}_n-P\right)\rho_s\left(\hat{\beta}_n,\hat{\Lambda}_n,\hat{\sigma}_n^2,h^*_{\sigma^2,s};X\right)+P\rho_s\left(\hat{\beta}_n,\hat{\Lambda}_n,\hat{\sigma}_n^2,h^*_{\sigma^2,s};X\right)+o_p(1)\\
=&P\rho_s\left(\hat{\beta}_n,\hat{\Lambda}_n,\hat{\sigma}_n^2,h^*_{\sigma^2,s};X\right)+o_p\left(1\right). \qquad (4.2)
\end{aligned}$$

Therefore, by the Glivenko-Cantelli Theorem, consistency of $\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2\right)$, continuous mapping and dominant convergence theorem (DCT),

$$\begin{aligned}
&P\left(\rho_s\left(\beta_0, \Lambda_0, \sigma_0^2, \hat{h}_{n,s}; X\right) - \rho_s\left(\beta_0, \Lambda_0, \sigma_0^2, h^*_{\sigma^2,s}; X\right)\right) \\
=&P\left(\rho_s\left(\beta_0, \Lambda_0, \sigma_0^2, \hat{h}_{n,s}; X\right) - \rho_s\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2, \hat{h}_{n,s}; X\right)\right) - P\left(\rho_s\left(\beta_0, \Lambda_0, \sigma_0^2, h^*_{\sigma^2,s}; X\right) - \right. \\
&\left.\rho_s\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2, h^*_{\sigma^2,s}; X\right)\right) + P\left(\rho_s\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2, \hat{h}_{n,s}; X\right) - \rho_s\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2, h^*_{\sigma^2,s}; X\right)\right) \\
=&o_p(1) + P\left(\rho_s\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2, \hat{h}_{n,s}; X\right) - \rho_s\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2, h^*_{\sigma^2,s}; X\right)\right) \\
&\text{(by continuous mapping and DCT)} \\
\leq&o_p(1) - \left(\mathbb{P}_n - P\right)\rho_s\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2, \hat{h}_{n,s}; X\right) \quad \text{(by inequality in (4.2))} \\
=&o_p(1) \quad \text{(by Glivenko-Cantelli Theorem)}
\end{aligned}$$

With the uniqueness of $h^*_{\sigma_0^2,s}$, the event $\left\|\hat{h}_{n,s} - h^*_{\sigma_0^2,s}\right\|_\infty > \varepsilon$ is a subset of the event $P\rho_s\left(\beta_0, \Lambda_0, \sigma_0^2, \hat{h}_{n,s}; X\right) > P\rho_s\left(\beta_0, \Lambda_0, \sigma_0^2, h^*_{\sigma_0^2,s}; X\right)$ and the latter goes to zero in probability as $n \to \infty$. Let $\varepsilon \to 0$ we conclude $\|\hat{h}_{n,s} - h^*_{\sigma_0^2,s}\|_\infty \to 0$.

Denote $\rho_1\left(\beta, \Lambda, \sigma^2, h; X\right) = \left(m_1\left(\beta, \Lambda, \sigma^2; X\right) - m_2\left(\beta, \Lambda, \sigma^2; X\right)[h]\right)^{\otimes 2}$ and $\mathfrak{S}_1 = \{\rho_1\left(\beta, \Lambda, \sigma^2, h; X\right) : \beta \in \mathcal{R}^d, log\Lambda \in \psi_{l,t}, \sigma^2 \in \mathcal{R}^+, h \in \phi_{l,t}\}$. By the consistency of $\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2$ and $\hat{h}_n$ and $\mathfrak{S}_1$ being a Glivenko-Cantelli, we can show

$$\begin{aligned}
\hat{B}_n &= \mathbb{P}_n\rho_1\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2, \hat{h}_n; X\right) \\
&= \left(\mathbb{P}_n - P\right)\rho_1\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2, \hat{h}_n; X\right) + P\rho_1\left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2, \hat{h}_n; X\right) \\
&\longrightarrow P\rho_1\left(\beta_0, \Lambda_0, \sigma_0^2, h^*_{\sigma_0^2}; X\right) = B_0
\end{aligned}$$

Let $\rho_2\left(\beta, \Lambda, \sigma^2, h; X\right) = m_{12}\left(\beta, \Lambda; \sigma^2, X\right) - m_{22}\left(\beta, \Lambda, \sigma^2; X\right)[h]$, we can similarly show that the class $\mathfrak{S}_2 = \{\rho_2\left(\beta, \Lambda, h; \sigma^2, X\right) : \beta \in \mathcal{R}^d, log\Lambda \in \psi_{l,t}, \sigma^2 \in \mathcal{R}^+, h \in$

$\phi_{l,t}\}$ is Glivenko-Cantelli. And

$$\begin{aligned}\hat{A}_n &= -\mathbb{P}_n \rho_2 \left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2, \hat{h}_n; X\right) \\ &= -\left(\mathbb{P}_n - P\right) \rho_2 \left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2, \hat{h}_n; X\right) - P\rho_2 \left(\hat{\beta}_n, \hat{\Lambda}_n, \hat{\sigma}_n^2, \hat{h}_n; X\right) \\ &\longrightarrow -P\rho_2 \left(\beta_0, \Lambda_0, \sigma_0^2, h^*_{\sigma_0^2}; X\right) = A_0\end{aligned}$$

□

4.2 GEE Sandwich Estimator

In Section 4.1 we prove the consistency of the spline-based sieve estimate of $\hat{h}_n$ using the general theorem of the maximum likelihood estimation and the projection algorithm. However, we may not be able to use this projection method if the generalized estimating equation does not coincide with the gradient of any objective function. The projection algorithm treats the baseline mean function as the infinite dimensional parameter. The estimation is complicated and another spline-based sieve approximation is needed to estimate the 'least favorable direction', i.e., $h^*_{\sigma^2}$ first.

In this section, we present an alternative ad hoc method for the estimation of the standard error of the estimated regression parameter. By treating the spline coefficients as same as the regression parameters, we propose to estimate the standard error of the estimated regression parameter s using the ordinary sandwich form in the generalized estimating equation for parametric model as follows.

In a parametric regression setting, we consider the observations (y_{ij}, x_{ij}) for times $t_{ij}, j = 1, \cdots, K_i, i = 1, \cdots, n$. y_{ij} is the outcome variable and x_{ij} is the covariate vector at t_{ij}. Let $Y_i = (y_{i1}, \cdots, y_{iK_i})^T$ be the outcome vector and $X_i =$

$(x_{i1}, \cdots, x_{iK_i})^T$ be the covariate matrix for subject i. Define μ_i to be the expectation of Y_i and suppose that $\mu_i = h(X_i\theta)$ with a know link function h. Denote the variance of Y_i as V_i. The GEE estimator of θ, $\hat{\theta}_n$ is the solution of the score-like equation system given by

$$U(\theta) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \theta}\right)^T V_i^{-1} (Y_i - \mu_i) \tag{4.3}$$

Liang & Zeger (1986) showed that $\hat{\theta}_n$ is a consistent estimator of θ and $\sqrt{n}\left(\hat{\theta}_n - \theta\right)$ is asymptotically multivariate normal with the covariance matrix give by a sandwich form

$$\lim_{n\to\infty} V_1^{-1} V_0 V_1^{-1},$$

where

$$\begin{aligned} V_1 &= \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \theta}\right)^T V_i^{-1} \left(\frac{\partial \mu_i}{\partial \theta}\right), \\ V_0 &= \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \theta}\right)^T cov(Y_i) \left(\frac{\partial \mu_i}{\partial \theta}\right). \end{aligned}$$

This asymptotic variance can be estimated consistently by replacing $cov(Y_i)$ by $(Y_i - \mu_i)(Y_i - \mu_i)^T$.

The proposed spline-based sieve GEE estimator, $\left(\hat{\beta}_n, \hat{\Lambda}_n\right)$ is estimated by solving Equation (2.6). They are similar to Equation (4.3). Heuristically, we could treat the spline coefficients as same as the regression parameters and estimate the standard error of $\hat{\beta}_n$ by the sandwich formula given above. That is, letting $\theta = (\beta, \alpha)^T_{(d+q_n)\times 1}$, the standard error of $\hat{\beta}_n$ could be estimated by the square root of the first d elements of the diagonal of $V_1^{-1} V_0 V_1^{-1}$.

Simulation results from Chapter 6 show that the standard error estimates based on the GEE sandwich form are similar to the estimates based on the projection method. Hence one might use the simplified ad hoc sandwich standard error estimation method instead of the projection method in practice. Theoretically, future work is needed to prove the estimation based on the GEE sandwich form and that based on the projection algorithm are asymptotically equivalent.

4.3 Bootstrap Method

The spline-based sieve approximation largely reduces the dimension of the estimation problem, which makes it feasible to estimate the standard error of β using the bootstrap method. In this manuscript, a case resampling bootstrap method is applied to estimate the standard error of the proposed spline-based sieve semiparametric GEE estimate of β. For a given dataset, the observations $\left(K_i, \underline{T}_{K_i}, \mathbb{N}^{(i)}, Z_i\right)$ are resampled with replacement 100 times. The bootstrap standard error is then calculated by the standard error of these 100 spline-based sieve GEE estimates of β based on the bootstrap datasets.

CHAPTER 5
NUMERICAL ALGORITHMS

5.1 Convex optimization algorithm with monotonicity constraint

Minimizing a smooth convex function ϕ over one of the cones $\mathcal{C}$ or $\mathcal{C}_+$ in $\mathbb{R}^n$, defined by

$$\mathcal{C} = \{x_i, i = 1, 2, \cdots n : x_1 \leq x_2 \leq \cdots \leq x_n\} \quad \text{or} \quad \mathcal{C}^+ = \{x \in \mathcal{C} : x_1 > 0\}$$

is often seen in statistical problems. Nonparametric and semiparametric maximum likelihood estimations, such as the estimations of hazard functions and distribution functions, can fit in this framework by taking ϕ to be the negative of the corresponding likelihood. See examples in Huang (1996), Wellner & Zhang (2000) and Wellner & Zhang (2007). Incorporating the monotone constraints into the computing algorithm is required to guarantee the validity of the estimation.

At any estimating iteration k, The convex function ϕ can be approximated locally at the current estimate $x^{(k)}$ by a quadratic form,

$$\tilde{\phi}\left(x, x^{(k)}\right) = \frac{1}{2}\left(x - f\left(x^{(k)}\right)\right)^T \tilde{W}\left(x^{(k)}\right)\left(x - f\left(x^{(k)}\right)\right)$$

Where

$$f\left(x^{(k)}\right) = x^{(k)} + g\left(x^{(k)}\right) \text{ and } g\left(x^{(k)}\right) = \lambda W\left(x^{(k)}\right)^{-1} \nabla\phi\left(x^{(k)}\right)$$

λ is a line search parameter with $0 < \lambda \leq 1$ such that $\phi\left(f\left(x^{(k)}\right)\right) \leq \phi\left(x^{(k)}\right)$. $\tilde{W}$ could be any positive definite matrix. The minimization of ϕ can be accomplished by iteratively minimizing $\tilde{\phi}$ subject to the monotone constraints, $x \in \mathcal{C}$ or $x \in \mathcal{C}^+$.

When $W\left(x^{(k)}\right)$ is the negative of the Hessian matrix, $f\left(x^{(k)}\right)$ is the Newton-Raphson update of the estimate. However the updates $f\left(x^{(k)}\right)$ does not automatically satisfy the monotone constraints. We present two different algorithms: the Generalized Rosen (GR) algorithm and the Convex Minorant (CM) algorithm to project this update onto the convex cone, $\mathcal{C}$ or $\mathcal{C}^+$.

GR-algorithm updates the gradient of the estimates, i.e., $g\left(x^{(k)}\right)$ onto the intersection of hyperplanes defined by some active constraints, which result in the updated estimates inside the convex cone. This method is utilized by Lu et al. (2007) and Lu et al. (2009) in the spline-based sieve maximum pseudolikelihood estimator and the spline-based sieve maximum likelihood estimator. Section 5.1.1 discusses GR-algorithm in detail and states its implementation in computing the spline-based sieve GEE estimates for panel count data. CM-algorithm projects the updated $f\left(x^{(k)}\right)$ directly to the convex cone determined by the monotone constraints. This algorithm can be viewed as a special form of the isotonic regression on a generalized gradient update. Section 5.1.2 explains the isotonic regression in detail and presents the implementation of the CM-algorithm developed by Jongbloed (1998) and a more generalized hybrid algorithm of Newton-Raphson iteration and isotonic regression.

5.1.1 Generalized Rosen (GR) Algorithm

Rosen (1960) first proposed a projection method for optimization problems with linear constraints. Jamshidian (2004) generalized Rosen's projection method to a general metric with the norm $\|x\| = x^T W x$. The GR-algorithm is based on the projections of $g\left(x^{(k)}\right)$ onto the intersection of hyperplanes determined by an active set $\mathcal{A}$, which is a set of indices of linear constraints. For example, $\mathcal{A} = \{j_1, \cdots, j_m\}$ for which $x_{j_i} = x_{j_i+1}$. $\mathcal{A}$ is allowed to be empty when $m = 0$. We start from defining the active set $\mathcal{A}$ and explain the projection algorithm afterwards. For the simplicity of the presentation, we suppress the dependence of f and g on x, denote $f^{(k)} = f\left(x^{(k)}\right)$, $g^{(k)} = g\left(x^{(k)}\right)$ and $\tilde{W}^{(k)} = \tilde{W}\left(x^{(k)}\right)$.

Given an estimate of x in the convex cone $\mathcal{C}$ or $\mathcal{C}^+$, $x^{(k)}$, if $g^{(k)}$ is nondecreasing, then $x^{(k+1)}$ satisfies the constraints automatically, no projection is needed.

If $g_j^{(k)} > g_{j+1}^{(k)}$ for some $j \in \{1, 2, \cdots, n\}$, to ensure $x_j^{(k+1)} \leq x_{j+1}^{(k+1)}$, we need to choose γ_j such that

$$x_j^{(k)} + \gamma_j g_j^{(k)} \leq x_{j+1}^{(k)} + \gamma_j g_{j+1}^{(k)}.$$

This implies

$$\gamma_j \leq \frac{x_{j+1}^{(k)} - x_j^{(k)}}{g_j^{(k)} - g_{j+1}^{(k)}}$$

If we choose

$$\gamma = \min_{\{j: g_j^{(k)} > g_{j+1}^{(k)}\}} \gamma_j,$$

it follows that $x^{(k+1)} \in \mathcal{C}$ or $\mathcal{C}^+$ and the active set $\mathcal{A}$ will expand to contain an extra index J where

$$J = \operatorname*{argmin}_{\{j: g_j^{(k)} > g_{j+1}^{(k)}\}} \gamma_j.$$

For the projection algorithm, a matrix A, whose rows correspond to the m active linear constraints and columns correspond to the n parameters, is defined as follows,

$$A = \begin{bmatrix} 0 & \cdots & -1 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & \cdots & \cdots & -1 & 1 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \end{bmatrix}_{m \times n}$$

The i^{th} row of A corresponds to the i^{th} active linear constraint $x_{j_i} = x_{j_i+1}$ in the active set $\mathcal{A} = \{j_1, \cdots, j_m\}$. It has -1 and 1 at the j_i^{th} and $j_i + 1^{th}$ element, zero elsewhere. This matrix may be updated as the algorithm proceed and hence is denoted as $A^{(k)}$ corresponding to the update $x^{(k)}$.

We need to project $g^{(k)}$ to the null space of the matrix $A^{(k)}$, defined as

$$\mathcal{N} = \{x : A^{(k)}x = 0\}$$

Let

$$\tilde{P}^{(k)} = \left(I - \left(\tilde{W}^{(k)}\right)^{-1} A^{(k)T} \left(A^{(k)} \left(\tilde{W}^{(k)}\right)^{-1} A^{(k)T} \right)^{-1} A^{(k)} \right)$$

It is easy to show that $\tilde{P}^{(k)}$ is idempotent, that is $\tilde{P}^{(k)}\tilde{P}^{(k)} = \tilde{P}^{(k)}$ and $A\tilde{P}^{(k)} = 0$ (Jamshidian 2004). Let $d = x - x^{(k)}$. For any $d \in \mathcal{N}$

$$\begin{aligned} \tilde{\phi}\left(x, x^{(k)}\right) &= \tilde{\phi}\left(d, g^{(k)}\right) = \left(d - g^{(k)}\right)^T \tilde{W}^{(k)} \left(d - g^{(k)}\right) \\ &= \left(d - \tilde{P}^{(k)} g^{(k)}\right)^T \tilde{W}^{(k)} \left(d - \tilde{P}^{(k)} g^{(k)}\right) \\ &\quad + \left(\tilde{P}^{(k)} g^{(k)} - g^{(k)}\right)^T \tilde{W}^{(k)} \left(\tilde{P}^{(k)} g^{(k)} - g^{(k)}\right) \\ &\geq \left(\tilde{P}^{(k)} g^{(k)} - g^{(k)}\right)^T \left(\tilde{W}^{(k)}\right) \left(\tilde{P}^{(k)} g^{(k)} - g^{(k)}\right), \end{aligned}$$

the equality holds at $d = \tilde{P}^{(k)} g^{(k)}$. Therefore $\tilde{P}^{(k)}$ is the projection matrix, the projected $g^{(k)}$ is $\tilde{P}^{(k)} g^{(k)}$ and $x^{(k+1)} = x^{(k)} + \tilde{P}^{(k)} g^{(k)}$. If the projected direction is not nondecreasing for those that are not already in the active set, γ needs to be estimated and an additional constraint needs to be added to the active set.

GR-algorithm updates the estimates by projections onto the vertices of active constraints. A Lagrange multiplier needs to be estimated at the end of the convergence to ensure the final estimate is the optimal solution. If the Lagrange multiplier is positive, the corresponding constraint is unnecessary and should be removed from the active set. The iteration continues with the updated active set. The steps used in GR-algorithm to solve for the spline-based sieve GEE subject to the monotone constraints are summarized in Table 5.1. In the simulation studies in Chapter 6, both $\tilde{W}$ and W are specified as the negative of the Hessian matrix.

5.1.2 Newton-Raphson/Isotonic Regression (NR/IR)

Groeneboom & Wellner (1992) first introduced the iterative convex minorant (ICM) algorithm to compute nonparametric maximum likelihood estimators (NPMLE). Jongbloed (1998) modified the ICM algorithm by inserting a line search parameter and showed the global convergence of the modified ICM algorithm. Other examples of applying ICM to estimation problems of censored or truncated data can be found in Pan (1999), Wellner & Zhang (2000) and Zhang & Jamshidian (2004) and the references therein.

ICM is based on the isotonic regression theory. Let $\mathcal{K}$ denote a convex cone of $\mathcal{C}$ or $\mathcal{C}^+$. $\hat{x} = \operatorname{argmin}_{x \in \mathcal{K}} \phi(x)$ if and only if $\hat{x}$ satisfies the Fenchel's optimality

Table 5.1: GR Algorithm for Spline-based Sieve GEE

Step 0: Start with an initial point $\theta^{(0)} = \left(\beta^{(0)}, \alpha^{(0)}\right)$ that satisfies the monotone constraint of the spline parameter, $\alpha^{(0)} = \left(\alpha_1^{(0)}, \alpha_2^{(0)}, \cdots, \alpha_{q_n}^{(0)}\right), \alpha_1^{(0)} \leq \alpha_2^{(0)} \leq \cdots \leq \alpha_{q_n}^{(0)}$. Iterate the algorithm through the following steps until convergence.

Step 1: Compute the feasible direction

$$d = \left\{ I - \tilde{W}^{-1} A^T \left(A \tilde{W}^{-1} A^T \right)^{-1} A \right\} W^{-1} U(\theta)$$

When there is no active constraint, take $d = W^{-1} U(\theta)$.

Step 2: If the resulted direction d is not nondecreasing, compute the biggest step

$$\gamma = \min_{i \notin \mathcal{A}, d_i > d_{i+1}} \left(- \frac{\alpha_{i+1} - \alpha_i}{d_{i+1} - d_i} \right)$$

This guarantees $\alpha_{i+1} + \gamma d_{i+1} \geq \alpha_i + \gamma d_i$, for $i = 1, 2, \cdots, q_n$

Step 3: Looking for the smallest integer $k \geq 0$ such that $\left\| U\left(\theta + (1/2)^k d\right) \right\| < \| U(\theta) \|$

Step 4: If $\gamma > (1/2)^k$, replace θ by $\tilde{\theta} = \theta + (1/2)^k d$ and go to Step 5.

If $\gamma \leq (1/2)^k$, replace θ by $\tilde{\theta} = \theta + \gamma d$, modify active set $\mathcal{A}$ and corresponding working matrix A by adding the new activated linear constraints.

Step 5: If $\|d\| \geq \varepsilon$ for a small $\varepsilon > 0$, go to Step 1. Otherwise, compute the Lagrange multiplier $\lambda = \left(A \tilde{W}^{-1} A^T \right)^{-1} A W^{-1} U(\theta)$.

i. If $\lambda_i \leq 0$ for all $i \in \mathcal{A}$, set $\hat{\theta} = \theta$ and stop.

ii. If at least one $\lambda_i > 0$ for $i \in \mathcal{A}$, remove the index corresponding to the largest λ_i from $\mathcal{A}$, and update A and go to Step 1.

condition, that is,

$$(\hat{x}, \nabla\phi(\hat{x})) = 0 \text{ and } (x, \nabla\phi(\hat{x})) \geq 0 \forall x \in \mathcal{K}$$

(Robertson et al. 1988). When $\phi(x)$ has a quadratic form, e.g.

$$\phi(x) = \frac{1}{2}(x-y)^T \tilde{W}(x-y)$$

and W is a diagonal matrix, the optimization reduces to estimating

$$\hat{x} = \operatorname*{argmin}_{x \in \mathcal{K}} \sum_{i=1}^{n} \tilde{w}_i (x_i - y_i)^2$$

where $\tilde{w}_i$ is the diagonal component of $\tilde{W}$.

The solution of this optimization has a nice graphic interpretation: it is the left derivative of the greatest convex minorant of the cumulative sum diagram, $\{P_i, i = 0, 1, \cdots, n\}$ where

$$P_0 = (0,0) \text{ and } P_i = \left(\sum_{l=1}^{i} \tilde{w}_l, \sum_{l=1}^{i} \tilde{w}_l y_l\right);$$

the left derivative of this diagram can be calculated by the pool adjacent violator algorithm (PAVA) described in Robertson et al. (1988) and the minimum-lower-set algorithm described in Brunk et al. (1957). As a matter of fact the solution can be expressed as

$$\hat{x}_i = \max_{j<i} \min_{l>i} \frac{\sum_{k=j}^{l} \tilde{w}_k y_k}{\sum_{k=j}^{l} \tilde{w}_k}.$$

In the nonparametric and semiparametric estimating problems as studied in Wellner & Zhang (2000) and Wellner & Zhang (2007), the number of parameters increases as the sample size increases. Storing and inverting the full high dimensional

Hessian matrix is daunting. The ICM-algorithm is implemented in which the matrix W in the generalized gradient update and $\tilde{W}$ are both diagonal with the negative diagonal elements of Hessian matrix, i.e., $W = \tilde{W} = D_H$.

In the spline-based sieve estimating problems, the dimension of the estimation increases much slower than the sample size. Instead of a diagonal matrix, the full Hessian matrix is used in the generalized gradient update which is essentially the Newton-Raphson update step. And a diagonal matrix $\tilde{W} = D_H$ is used to project the Newton-Raphson estimate onto the convex cone using the max-min formula. Obviously, the such a Newton-Raphson and Isotonic Regression hybrid algorithm would converge faster than the ICM-algorithm. The hybrid algorithm of Newton Raphson iteration and isotonic regression (NR/IR) algorithm tailored to the spline-based sieve GEE estimates is summarized in Table 5.2.

GR, ICM and the more general hybrid algorithm NR/IR are all based on the quadratic approximation. GR converts the inequality constraints of the estimates to an active set and update the active set during each iteration. ICM and NR/IR make a good use of the geometric interpretation of the isotonic regression and estimate the parameters subject to the monotone constraints directly. Best & Chakravarti (1990) shows that some isotonic regression methods, e.g. PAVA (Robertson et al. 1988) and the minimum-lower-set algorithm (Brunk et al. 1957) can also be fitted into the unifying framework of active set approach.

Table 5.2: NR/IR Algorithm for Spline-based Sieve GEE

Step 0: Start with an initial point $\theta^{(0)} = \left(\alpha^{(0)}, \beta^{(0)}\right)$ that satisfies the monotone constraint of the spline parameter,$\alpha^{(0)} = \left(\alpha_1^{(0)}, \alpha_2^{(0)}, \cdots, \alpha_{q_n}^{(0)}\right), \alpha_1^{(0)} \leq \alpha_2^{(0)} \leq \cdots \leq \alpha_{q_n}^{(0)}$. Iterate the algorithm through the following steps until convergence.

Step 1: Look for a smallest integer k starting from 0 such that

$$\|U\left(\theta + (1/2)^k W^{-1} U(\theta)\right)\| < \|U(\theta)\|$$

Update the current estimates $\hat{\theta}^{(k)} = \left(\hat{\alpha}^{(k)}, \hat{\beta}^{(k)}\right)$ by

$$\tilde{\theta}^{(k+1)} = \left(\tilde{\alpha}^{(k+1)}, \tilde{\beta}^{(k+1)}\right) = \hat{\theta}^{(k)} + (1/2)^k W^{-1} U\left(\theta^{(k)}\right)$$

Step 2: Project the updated updated $\tilde{\alpha}^{(k+1)}$ using the isotonic regression by

$$\hat{\alpha}_i^{(k+1)} = \underset{x \in \mathcal{K}}{\operatorname{argmin}} \frac{1}{2}\left(x - \tilde{\alpha}^{(k+1)}\right) \tilde{W} \left(x - \tilde{\alpha}^{(k+1)}\right):$$

Construct the cumulative sum diagram $\{P_i, i = 0, 1, \cdots, n\}$ where

$$P_0 = (0,0) \text{ and } P_i = \left(\sum_{l=1}^{i} \tilde{w}_l, \sum_{l=1}^{i} \tilde{w}_l \tilde{\alpha}_l^{(k+1)}\right);$$

Calculate the left derivative of the greatest convex minorant of this cumulative sum diagram by

$$\hat{\alpha}_i^{(k+1)} = \max_{j<i} \min_{l>i} \frac{\sum_{m=j}^{l} \tilde{w}_m \tilde{\alpha}_m^{(k+1)}}{\sum_{m=j}^{l} \tilde{w}_m}$$

Since there is no constraints on β, let $\hat{\beta}^{(k+1)} = \tilde{\beta}^{(k+1)}$.

Step 3: Check the convergence criteria: Let $d = \|\hat{\theta}^{(k+1)} - \hat{\theta}^{(k)}\|$, if $d \geq \varepsilon$ for a small $\varepsilon > 0$ go to Step 1. Otherwise stop the algorithm.

5.2 Estimating the Over-Dispersion Parameter

The spline-based sieve semiparametric GEE with $V_3^{(i)}$ requires an estimator of the over-dispersion parameter in addition to the parameters in the proportional mean function. In Chapter 3, we show that as long as the estimated over-dispersion parameter is consistent, the spline-based sieve GEE estimates of (β_0, Λ_0) still have good asymptotic properties. In this section we discuss three different estimating methods.

Given a consistent estimate of (β_0, Λ_0), the over-dispersion parameter σ^2 could be estimated by maximizing the Gamma-Frailty Poisson likelihood as shown in Equation (2.7). It will be the most efficient estimator when the data are indeed generated from a Gamma-Frailty Poisson process. However, in order for the likelihood to be valid, the parameter space of σ^2 need to be restricted to $\mathcal{R}^+$. With only one additional parameter in the likelihood, we can simplify the estimation by a grid search, in which the MLE of σ^2 is the one that produces the largest likelihood.

In addition to the maximum likelihood estimator, method-of-moment is often used in parametric regression for estimating the over-dispersion in the literature of count data. Breslow (1984) used a method of moment to estimate this parameter by

$$\sum_{i=1}^{n}\sum_{j=1}^{K_i}\frac{\left(\mathbb{N}_{ij}-\hat{\mu}_{ij}\right)^2}{\hat{\mu}_{ij}+\sigma^2\hat{\mu}_{ij}^2}=\sum_{i=1}^{n}K_i-p$$

where $\hat{\mu}_{ij}$ is any consistent estimate of $E\left(\mathbb{N}\left(T_{ij}\right)\right)$, and p is the number of estimated parameters. In Breslow's method, the over-dispersion parameter can be computed iteratively using a self-consistent algorithm given by

$$\hat{\sigma}_n^2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{K_i} \frac{(\mathbb{N}_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}\left(\hat{\mu}_{ij} + \hat{\sigma}_n^{-2}\right)}}{\sum_{i=1}^{n} K_i - p}$$

Alternatively, σ^2 could also be estimated explicitly by

$$\hat{\sigma}_n^2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{K_i} \left\{(\mathbb{N}_{ij} - \hat{\mu}_{ij})^2 - \hat{\mu}_{ij}\right\}}{\sum_{i=1}^{n} \sum_{j=1}^{K_i} \hat{\mu}_{ij}^2}$$

as proposed by Zeger (1988). Both Zeger's method and Breslow's formula could end up with negative $\hat{\sigma}^2$. If that happens, $\hat{\sigma}_n^2$ is forced to be zero. Davis et al. (2000) pointed out that Zeger's method underestimates the over-dispersion parameter and provided an adjustment for the bias and showed that the modified estimator is consistent. As a matter of fact, the Breslow's method also underestimates the over-dispersion parameter. In our spline-based sieve semiparametric GEE method, this over-dispersion parameter is treated as a nuisance parameter and for the sake of numerical simplicity, Zeger's method is adopted in our calculations. We will show in Lemma 5.1 that this estimate converges to a positive value in $\mathcal{R}^+$.

Define the following two functions,

$$g_1(\beta, \Lambda; X) = \sum_{j=1}^{K} \left(\mathbb{N}(t_j) - \Lambda(t_j) e^{\beta^T Z}\right)^2 - \Lambda(t_j) e^{\beta^T Z}$$

$$g_2(\beta, \Lambda; X) = \sum_{j=1}^{K} \Lambda(t_j) e^{\beta^T Z}$$

and the two corresponding classes as

$$\mathcal{G}_1 = \left\{g_1(\beta, \Lambda; X) : \beta \in \mathcal{R}^d, log\Lambda \in \mathcal{F}\right\}$$

$$\mathcal{G}_2 = \left\{g_2(\beta, \Lambda; X) : \beta \in \mathcal{R}^d, log\Lambda \in \mathcal{F}\right\}$$

Let $\hat{\theta}_n^{(0)} = \left(\hat{\beta}_n^{(0)}, \hat{\Lambda}_n^{(0)}\right)$ be the estimates using the Poisson pseudolikelihood (or likelihood or estimates based on the GEE with frailty variance matrix $V_3^{(i)}$ using any arbitrary fixed σ^2 value), they are consistent estimates of $\theta_0 = (\beta_0, \Lambda_0)$. Zeger's estimator of the overdispersion parameter can be written as $\hat{\sigma}_n^2 = \frac{\mathbb{P}_n g_1\left(\hat{\beta}_n^{(0)}, \hat{\Lambda}_n^{(0)}; X\right)}{\mathbb{P}_n g_2\left(\hat{\beta}_n^{(0)}, \hat{\Lambda}_n^{(0)}; X\right)}$.

Lemma 5.1.

$$\hat{\sigma}^2 \to_p \sigma_0^2 = \frac{P g_1\left(\beta_0, \Lambda_0; X\right)}{P g_2\left(\beta_0, \Lambda_0; X\right)}.$$

Proof. By Lemma 3.2 the bracketing number of $\mathcal{F}$ with $L_1(P)$ norm is bounded by $C\left(exp\left(1/\varepsilon\right)\right)$. So $\mathcal{F}$ is covered by

$$\left\{[\Lambda_i^L, \Lambda_i^R] : i = 1, 2, \cdots, l\right\}, l = O\left(exp\left(1/\varepsilon\right)\right)$$

and $\|\Lambda_i^R - \Lambda_i^L\|_{L_1(\mu)} = \int \left(\Lambda_i^R(t) - \Lambda_i^L(t)\right) d\mu(t) < \varepsilon$. Since $\mathcal{R}^d$ is compact, there exists a ε-net, $\{\beta_1, \beta_2, \cdots, \beta_p\}$, $p = O\left(1/\varepsilon^d\right)$ such that $\forall \beta \in \mathcal{R}, \exists s \in \{1, 2, \cdots, p\}$ such that $|\beta^T Z - \beta_s^T Z| \leq \varepsilon$ and $|exp\left(\beta^T Z\right) - exp\left(\beta_s^T Z\right)| \leq C\varepsilon$. Let

$$g_{1,i,s}^L = \sum_{j=1}^{K} \left\{\mathbb{N}^2(t_j) - \left(2\mathbb{N}(t_j) + 1\right)\Lambda_i^R(t_j)\left(e^{\beta_s^T Z} + C\varepsilon\right) + \left[\Lambda_i^L(t_j)\left(e^{\beta_s^T Z} - C\varepsilon\right)\right]^2\right\}$$
$$g_{1,i,s}^R = \sum_{j=1}^{K} \left\{\mathbb{N}^2(t_j) - \left(2\mathbb{N}(t_j) + 1\right)\Lambda_i^L(t_j)\left(e^{\beta_s^T Z} - C\varepsilon\right) + \left[\Lambda_i^R(t_j)\left(e^{\beta_s^T Z} + C\varepsilon\right)\right]^2\right\}$$

$\mathcal{G}_1$ is covered by $\left\{[g^L_{1,i,s}, g^R_{1,i,s}] : i = 1, \cdots, l; s = 1, \cdots p\right\}$

$$\begin{aligned}\Delta g_{1,i,s} &= g^R_{1,i,s} - g^L_{1,i,s}\\ &= \sum_{j=1}^{K} \Big\{ \left(2\mathbb{N}\left(t_j\right) + 1\right) \Big[\left(\Lambda^R_i\left(t_j\right) - \Lambda^L_i\left(t_j\right)\right) e^{\beta^T_s Z} + C\left(\Lambda^R_i\left(t_j\right) + \Lambda^L_i\left(t_j\right)\right)\varepsilon \Big] + \\ &\qquad\qquad \Big[\left(\Lambda^R_i\left(t_j\right) + \Lambda^L_i\left(t_j\right)\right) e^{\beta^T_s Z} + C\left(\Lambda^R_i\left(t_j\right) - \Lambda^L_i\left(t_j\right)\right)\varepsilon \Big] \times \\ &\qquad\qquad \Big[\left(\Lambda^R_i\left(t_j\right) - \Lambda^L_i\left(t_j\right)\right) e^{\beta^T_s Z} + C\left(\Lambda^R_i\left(t_j\right) + \Lambda^L_i\left(t_j\right)\right)\varepsilon \Big] \Big\}\end{aligned}$$

$$P|\Delta g_{1,i,s}| \le C_1 P \sum_{j=1}^{K} \left[\Lambda^R_i\left(t_j\right) - \Lambda^L_i\left(t_j\right)\right] + C_2\varepsilon \le C\varepsilon$$

Similarly,

$$g^L_{2,i,s} = \sum_{j=1}^{K} \Big[\Lambda^L_i\left(t_j\right) \Big(e^{\beta^T_s Z} - C\varepsilon \Big) \Big]; \quad g^R_{2,i,s} = \sum_{j=1}^{K} \Big[\Lambda^L_i\left(t_j\right) \Big(e^{\beta^T_s Z} - C\varepsilon \Big) \Big]$$

$\mathcal{G}_2$ is covered by $\left\{[g^L_{2,i,s}, g^R_{2,i,s}] : i = 1, \cdots, l; s = 1, \cdots p\right\}$.

$$\Delta g_{2,i,s} = g^R_{2,i,s} - g^L_{2,i,s} = \sum_{j=1}^{K} \Big[\left(\Lambda^R_i\left(t_j\right) - \Lambda^L_i\left(t_j\right)\right) e^{\beta^T_s Z} + C\left(\Lambda^R_i\left(t_j\right) + \Lambda^L_i\left(t_j\right)\right)\varepsilon \Big]$$

$$P|\Delta g_{2,i,s}| \le C_1 P \sum_{j=1}^{K} \left[\Lambda^R_i\left(t_j\right) - \Lambda^L_i\left(t_j\right)\right] + C_2\varepsilon \le C\varepsilon$$

Both $\mathcal{G}_1$ and $\mathcal{G}_2$ have a finite $L_1(P)$ bracket number. They are Glivenko-Cantelli classes (Lemma 3.1). Then together with the consistency of $\left(\hat{\beta}^{(0)}_n, \hat{\Lambda}^{(0)}_n\right)$ and the continuous mapping theorem,

$$\begin{aligned}\mathbb{P}_n g_1\left(\hat{\beta}_n^{(0)}, \hat{\Lambda}_n^{(0)}; X\right) =& \left(\mathbb{P}_n - P\right) g_1\left(\hat{\beta}_n^{(0)}, \hat{\Lambda}_n^{(0)}; X\right) + \\ & P\left(g_1\left(\hat{\beta}_n^{(0)}, \hat{\Lambda}_n^{(0)}; X\right) - g_1\left(\beta_0, \Lambda_0; X\right)\right) + Pg_1\left(\beta_0, \Lambda_0; X\right) \\ =& Pg_1\left(\beta_0, \Lambda_0; X\right) + o_p(1) \\ \mathbb{P}_n g_2\left(\hat{\beta}_n^{(0)}, \hat{\Lambda}_n^{(0)}; X\right) =& \left(\mathbb{P}_n - P\right) g_2\left(\hat{\beta}_n^{(0)}, \hat{\Lambda}_n^{(0)}; X\right) + \\ & P\left(g_2\left(\hat{\beta}_n^{(0)}, \hat{\Lambda}_n^{(0)}; X\right) - g_2\left(\beta_0, \Lambda_0; X\right)\right) + Pg_2\left(\beta_0, \Lambda_0; X\right) \\ =& Pg_2\left(\beta_0, \Lambda_0; X\right) + o_p(1)\end{aligned}$$

we have

$$\hat{\sigma}_n^2 \longrightarrow_p \sigma_0^2$$

□

In the simulations conducted in Chapter 6, a two-stage estimating procedure is implemented with $V_3^{(i)}$ as the covariance matrix. At the first stage, due to its computational convenience, the spline-based sieve semiparametric GEE with $V_1^{(i)}$ is implemented to get consistent estimates of μ_{ij}. σ^2 is then estimated using Zeger's method. At the second stage, replacing σ^2 by its consistent estimates, $\hat{\sigma}_n^2$, the estimate of (β, Λ) is updated by solving a pseudo GEE, i.e.,

$$U\left(\theta; \hat{\sigma}_n^2\right) = \sum_{i=1}^{n} \left(\frac{\partial \mu^{(i)}}{\partial \theta}\right) V_i^{-1}\left(\theta; \hat{\sigma}_n^2\right)\left(\mathbb{N}\left(T_i\right) - \mu^{(i)}\right) = 0$$

with $\theta = (\beta, \alpha)$. The hybrid algorithm NR/IR is used at both stages to solve the sieve estimating equation subject to the monotone constraints, $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_{q_n}$.

CHAPTER 6
NUMERICAL RESULTS

6.1 Simulation Studies

6.1.1 Simulation Setup

Simulation studies are conducted to examine the performance of the spline-based sieve semiparametric GEE estimate in finite samples. For each subject, we generate $X_i = \left(K_i, \underline{T}_{K_i}, \mathbb{N}^{(i)}, Z_i\right)$ in the following manner: Six follow-up times are pre-scheduled at $T^\circ = \{T_j^\circ : T_j^\circ = 2j, j = 1, \cdots, 6\}$. The actual observation times T_{ij}° are generated from a normal distribution, $N(T_j^\circ, 1/3)$. Let $\xi_{ij} = 1_{[T_{ij-1}^\circ < T_{ij}^\circ]}$, for $i = 1, \cdots, 6$ and $T_{i0}^\circ = 0$. Let $\delta_{ij} = 1$ if the j^{th} visit actually happens and zero otherwise. $P(\delta_{ij} = 1) = \frac{1}{1+e^{T_{ij}^\circ - 10}}$. Each subject has $K_i = \sum_{j=1}^{6} \xi_{ij}\delta_{ij}$ observations at $\underline{T}_{K_i} = \left(T_{K_i,1}^{(i)}, T_{K_i,2}^{(i)}, \cdots, T_{K_i,K_i}^{(i)}\right)$, where $T_{K_i,j}^{(i)}$ are the j^{th} order observation time of $\{T_{ij}^\circ : \xi_{ij}\delta_{ij} = 1, j = 1, \cdots, 6\}$. The covariate vector $Z_i = (Z_{i1}, Z_{i2}, Z_{i3})$ is simulated by $Z_{i1} \sim \text{Uniform}\,(0,1)\,, Z_{i2} \sim N\,(0,1)\,,$ and $Z_{i3} \sim \text{Bernoulli}\,(0.5)$. The regression parameter $\beta_0 = (\beta_{0,1}, \beta_{0,2}, \beta_{0,3})^T = (-1.0, 0.5, 1.5)^T$. Given $\left(Z_i, K_i, \underline{T}_{K_i}\right)$, different scenarios are used to generate the panel counts $\mathbb{N}^{(i)} = \left(\mathbb{N}\left(T_{K_i,1}^{(i)}\right), \mathbb{N}\left(T_{K_i,2}^{(i)}\right), \cdots, \mathbb{N}\left(T_{K_i,K_i}^{(i)}\right)\right)$.

Scenario 1. The panel counts are generated from a Poisson process with the conditional mean function given by $\Lambda\left(t_{ij}|Z_i\right) = 2t_{ij}^{1/2}e^{\beta_0^T Z_i}$, that is,

$$\mathbb{N}\left(T_{K_i,j}^{(i)}\right) - \mathbb{N}\left(T_{K_i,j-1}^{(i)}\right) \sim \text{Poisson}\left\{2\left[\left(T_{K_i,j}^{(i)}\right)^{1/2} - \left(T_{K_i,j-1}^{(i)}\right)^{1/2}\right]e^{\beta_0^T Z_i}\right\}$$

for $j = 1, 2, \cdots, K_i$.

Scenario 2. Data are generated from a Gamma Frailty Poisson model. The frailty parameters $\gamma_1, \gamma_2, \cdots, \gamma_n$ are a random sample from a Gamma distribution, $\Gamma(12.5, 12.5)$, which means the over-dispersion parameter, σ^2, is 0.08. Conditioning on the frailty parameter γ_i as well as the covariates Z_i, the panel counts for each subject are drawn from a Poisson process, i.e.

$$\mathbb{N}\left(T^{(i)}_{K_i,j}\right) - \mathbb{N}\left(T^{(i)}_{K_i,j-1}\right) \sim \text{Poisson}\left\{2\gamma_i\left[\left(T^{(i)}_{K_i,j}\right)^{1/2} - \left(T^{(i)}_{K_i,j-1}\right)^{1/2}\right] e^{\beta_0^T Z_i}\right\}$$

for $j = 1, 2, \cdots, K_i$. In this scenario, the counting process given only the covariate is not a Poisson process. However, the conditional mean given the covariate vector still satisfies the proportional mean model specified in Equation (1.2). The marginal distribution of the counts follows a negative binomial distribution.

Scenario 3. Data are generated similar to *Scenario 2.* Instead of generating the frailty term γ from a Gamma distribution, it is generated from a discrete distribution $\{0.6, 1, 1.4\}$ with probabilities 0.25, 0.5 and 0.25, respectively. This scenario generates a so called mixed Poisson process as studied in Wellner & Zhang (2007) and Lu et al. (2009). The variance of the frailty variable is also 0.08. In this scenario, the counting process given the covariate is not a Poisson process, nor its marginal distribution follows a negative binomial distribution. However, the proportional mean structure still holds.

Scenario 4. Data are generated from a 'Negative-binomialized' counting process. Conditioning on Z, a random variable N is generated from a Negative binomial distribution, $NegBin(20e^{\beta_0^T Z}, 0.1)$. Given N, a random sample, $X_i, i = 1, 2, \cdots, N$, is generated from distribution function $F_x = t^{1/2}/90$. The count data is defined by the number of X_i's that is smaller than or equal to t, i.e.,

$$\mathbb{N}\left(t\right)=\sum_{i=1}^{N}I_{[x_i\leq t]}$$

It is easy to see the proportional mean model in Equation (1.2) still hold. The baseline mean function $\Lambda_0\left(t\right)=2t^{1/2}$, is the same as those in scenarios 1, 2 and 3. Under this setting, both over-dispersion and autocorrelation between non-overlapping increments are present. The covariance matrix has a similar form as matrix $V_3^{(i)}$, but the true over-dispersion parameter depends on the covariates.

In all these scenarios, the monotone cubic B-splines are used in computing the sieve semiparametric GEE estimators. The number of interior knots is chosen to be $m_n=\lceil N^{1/3}\rceil$, the smallest integer above $N^{1/3}$, where N is the number of distinct observation times. These knots are placed at the corresponding quantiles of the distinct observation times. In our simulation studies, we generate 1000 Monte Carlo samples with sample size of 50 and 100 for each scenario.

6.1.2 Simulation Results

Simulation results are summarized in Table 6.1 - Table 6.8 corresponding to the four scenarios with two different sample sizes. They include bias, Monte-Carlo standard error and the mean of the standard error estimate based on the proposed projection method in Section 4.1, the mean of the GEE sandwich estimator of the standard error discussed in Section 4.2 and the mean of the bootstrap estimator of the standard error described in Section 4.3 and the 95% empirical coverage probabilities calculated using these estimated standard errors. The square of the biases and the

Monte-Carlo variance of the spline-based sieve GEE estimates of the baseline mean function calculated at points $t = 2, 2.25, 2.50, \cdots, 9$ are plotted in Figure 6.1 - Figure 6.4 corresponding to the four different scenarios.

Table 6.1 and Table 6.2 summarize the results with regard to the regression parameters when the data are from a nonhomogeneous Poisson process. The bias is negligible compared to the standard error for all three different covariance matrices. The estimates using $V_1^{(i)}$ have a larger standard error than those using $V_2^{(i)}$. When using $V_3^{(i)}$ as the covariance matrix, the estimates have similar standard errors as those using $V_2^{(i)}$, since 80.5% and 75.5% of the times in the simulations, the estimated over-dispersion parameter, $\hat{\sigma}_n^2$, is zero for sample size 50 and 100 respectively. And it result in the estimates using $V_3^{(i)}$ being the same as those using $V_2^{(i)}$. The standard errors based on the projection method tend to underestimate the true values compared to the Monte-Carlo standard error. However, the underestimation lessens as the sample size increases. The standard error estimates using the standard GEE sandwich formula are similar to those based on the projection method. The bootstrap method produces a better standard error estimate than the two aforementioned method particularly when sample size is small. The coverage probability based on the bootstrap standard error estimate is the best among the three regarding its closeness to the nominal level. Figure 6.1 plots the squared bias and the variance of the estimated baseline mean function at the corresponding time points based on the three covariance matrices. Similar to the results of the regression parameters, the estimates based on $V_1^{(i)}$ has the largest standard deviation. And the estimates based on $V_2^{(i)}$ and $V_3^{(i)}$ are similar to each other.

When data come from a Poisson process, the estimator based on $V_3^{(i)}$ performs similar to the estimator based on $V_2^{(i)}$. When the over-dispersion is present as exemplified in Scenarios 2 and 3, then the estimator using $V_3^{(i)}$ clearly outperforms the estimators using $V_1^{(i)}$ or $V_2^{(i)}$. Table 6.3 and Table 6.4 show the simulation results of the estimated regression parameters for Scenario 2. Similar to the results for Poisson data, all three estimators are asymptotically unbiased. The estimator using covariance matrix $V_3^{(i)}$ has a smaller standard error compared to the estimators using $V_1^{(i)}$ or $V_2^{(i)}$. This is expected as the variance matrix $V_3^{(i)}$ correctly specifies the underlying true variance-covariance matrix among the cumulative panel counts. Both the projection and the parametric sandwich standard error estimators appear to underestimate the true standard error a little bit when sample size is small, which attributes to the coverage probability lower than the nominal level. The underestimation lessens as sample size increases. Among the three estimators, it seems that the standard error estimates of the spline-based sieve semiparametric GEE estimator with $V_3^{(i)}$ have the least bias. When using $V_1^{(i)}$ and $V_2^{(i)}$ as the working covariance matrix, the bootstrap method also underestimates the true standard error. While when using $V_3^{(i)}$, the bootstrap method produces a smaller bias, and the 95% coverage based on the bootstrap method is near to its nominal level. Figure 6.2 shows the squared bias and the variance of the estimated baseline mean function at corresponding time points. Similar to the regression parameters, their bias are negligible relative to their variances. The estimates based on $V_3^{(i)}$ are most efficient followed by the estimates based on $V_2^{(i)}$.

Simulation results for Scenario 3 displayed in Tables 6.5, 6.6 and Figure 6.3 are similar to the results for Scenario 2. The estimator using $V_3^{(i)}$ is again most efficient compared to the other selections for the working covariance matrix. In this case, the working covariance matrix $V_3^{(i)}$ is still the true covariance matrix between the cumulative panel counts, even though the underlying frailty variable is not Gamma distributed.

Table 6.7 and Table 6.8 summarize the simulation results for Scenario 4. Again, the bias is negligible. The estimates based on $V_2^{(i)}$ and $V_3^{(i)}$ are comparable to each other. Both the projection method and the parametric GEE sandwich method underestimate the standard error of the spline-based sieve GEE estimates. The bootstrap method provides a better estimate of the standard error. Figure 6.4 plots the squared bias and the variance of the estimated baseline mean function.

Table 6.1: Monte-Carlo simulations results of the B-splines based sieve GEE estimator using different covariance matrix for Poisson data with sample size n=50

	Bias			M-C sd			Projection ASE			95% coverage		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
$V_1^{(i)}$	-0.0001	0.0023	-0.0002	0.0314	0.1016	0.0694	0.0276	0.0879	0.0651	0.9180	0.9130	0.9180
$V_2^{(i)}$	0.0003	0.0016	-0.0008	0.0291	0.0969	0.0657	0.0261	0.0908	0.0647	0.9090	0.8950	0.9160
$V_3^{(i)}$	0.0002	0.0015	-0.0008	0.0292	0.0972	0.0657	0.0261	0.0849	0.0622	0.9080	0.8950	0.9160
	GEE Sandwich ASE			95% coverage			Bootstrap ASE			95% coverage		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
$V_1^{(i)}$	0.0276	0.0879	0.0651	0.9180	0.9130	0.9180	0.0346	0.1045	0.0728	0.9680	0.9500	0.9480
$V_2^{(i)}$	0.0259	0.0826	0.0612	0.9100	0.8960	0.9160	0.0324	0.0976	0.0680	0.9660	0.9410	0.9420
$V_3^{(i)}$	0.0260	0.0828	0.0613	0.9090	0.8960	0.9170	0.0329	0.0986	0.0684	0.9680	0.9420	0.9430

Table 6.2: Monte-Carlo simulations results of the B-splines based sieve GEE estimator using different covariance matrix for Poisson data with sample size n=100

	Bias			M-C sd			Projection ASE			95% coverage		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
$V_1^{(i)}$	0.0012	0.0000	0.0014	0.0209	0.0687	0.0495	0.0189	0.0625	0.0462	0.9080	0.9190	0.9240
$V_2^{(i)}$	0.0010	0.0000	0.0014	0.0195	0.0636	0.0449	0.0176	0.0584	0.0431	0.9220	0.9320	0.9290
$V_3^{(i)}$	0.0009	−0.0002	0.0015	0.0196	0.0641	0.0450	0.0177	0.0585	0.0432	0.9230	0.9300	0.9300

	GEE Sandwich ASE			95% coverage			Bootstrap ASE			95% coverage		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
$V_1^{(i)}$	0.0189	0.0625	0.0462	0.9080	0.9190	0.9240	0.0218	0.0683	0.0488	0.9550	0.9400	0.9370
$V_2^{(i)}$	0.0176	0.0584	0.0431	0.9220	0.9320	0.9290	0.0203	0.0637	0.0454	0.9630	0.9510	0.9430
$V_3^{(i)}$	0.0177	0.0584	0.0432	0.9230	0.9300	0.9290	0.0206	0.0643	0.0456	0.9650	0.9540	0.9430

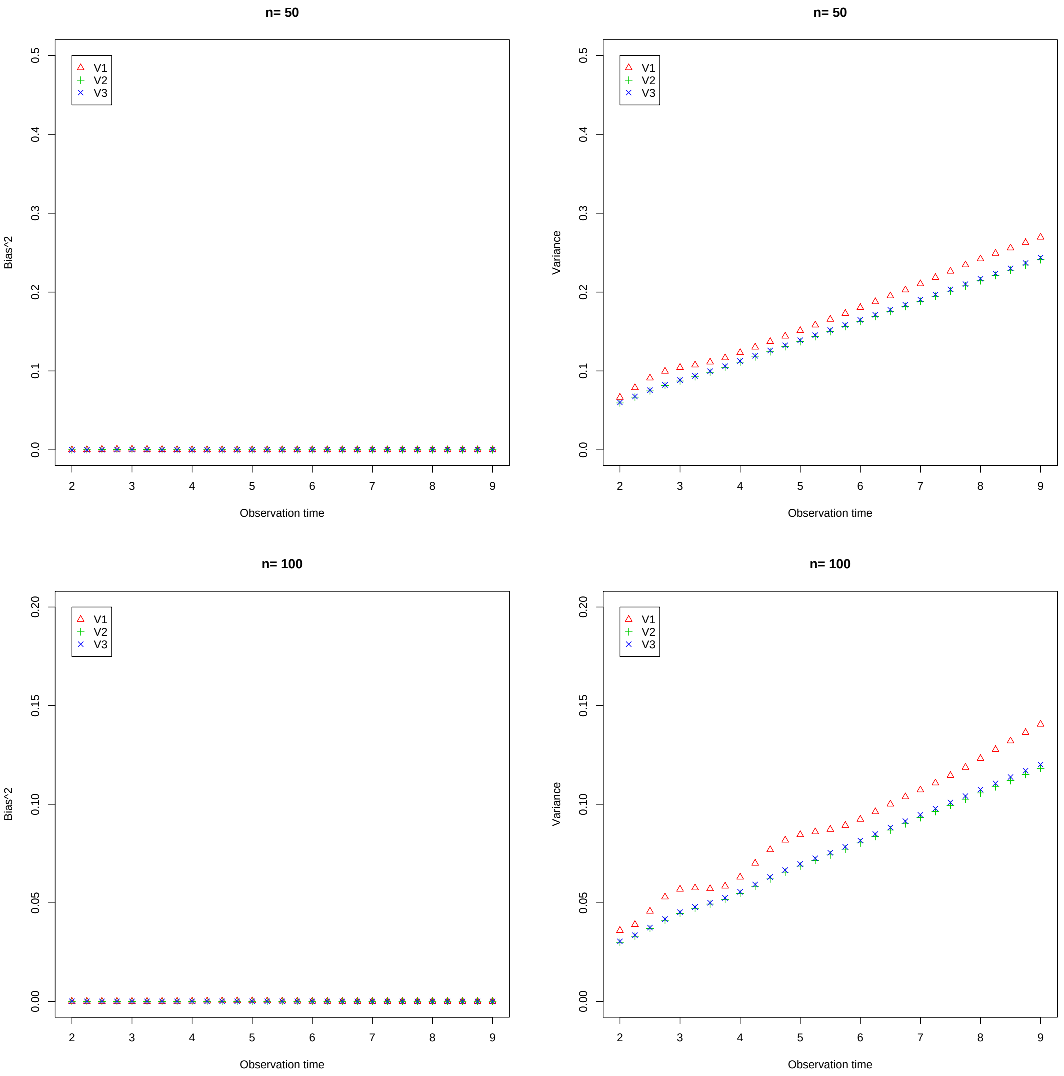

Figure 6.1: Scenario 1, with Data from the Poisson Model: $\Lambda_0\left(t\right)=2t^{1/2}$

Table 6.3: Monte-Carlo simulations results of the B-splines based sieve GEE estimator using different covariance matrix for Gamma Frailty Poisson data with sample size n=50

	Bias			M-C sd			Projection ASE			95% coverage		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
$V_1^{(i)}$	0.0023	-0.0066	0.0026	0.0829	0.2446	0.1454	0.0561	0.1957	0.1183	0.7830	0.8510	0.8800
$V_2^{(i)}$	0.0027	-0.0047	0.0019	0.0824	0.2411	0.1424	0.0555	0.1943	0.1167	0.7760	0.8640	0.8840
$V_3^{(i)}$	-0.0017	-0.0004	0.0017	0.0617	0.1958	0.1193	0.0549	0.1812	0.1104	0.8870	0.9180	0.9240

	GEE Sandwich ASE			95% coverage			Bootstrap ASE			95% coverage		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
$V_1^{(i)}$	0.0561	0.1957	0.1183	0.7820	0.8510	0.8800	0.0729	0.2273	0.1316	0.9150	0.9100	0.9180
$V_2^{(i)}$	0.0555	0.1942	0.1167	0.7760	0.8630	0.8840	0.0726	0.2267	0.1306	0.9060	0.9100	0.9200
$V_3^{(i)}$	0.0544	0.1776	0.1098	0.8900	0.9160	0.9240	0.0643	0.2007	0.1203	0.9380	0.9420	0.9520

Table 6.4: Monte-Carlo simulations results of the B-splines based sieve GEE estimator using different covariance matrix for Gamma Frailty Poisson data with sample size n=100

	Bias			M-C sd			Projection ASE			95% coverage		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
$V_1^{(i)}$	0.0031	-0.0031	-0.0007	0.0612	0.1858	0.0974	0.0433	0.1511	0.0883	0.8000	0.8840	0.9280
$V_2^{(i)}$	0.0030	-0.0029	0.0004	0.0603	0.1846	0.0944	0.0431	0.1505	0.0867	0.8180	0.8910	0.9260
$V_3^{(i)}$	0.0005	-0.0013	0.0016	0.0429	0.1385	0.0805	0.0396	0.1293	0.0786	0.9290	0.9370	0.9390
	GEE Sandwich ASE			95% coverage			Bootstrap ASE			95% coverage		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
$V_1^{(i)}$	0.0433	0.1511	0.0883	0.8000	0.8840	0.9270	0.0511	0.1607	0.0924	0.8920	0.9080	0.9430
$V_2^{(i)}$	0.0431	0.1505	0.0867	0.8190	0.8900	0.9260	0.0509	0.1601	0.0909	0.9100	0.9100	0.9410
$V_3^{(i)}$	0.0395	0.1291	0.0785	0.9240	0.9350	0.9400	0.0427	0.1348	0.0812	0.9420	0.9410	0.9480

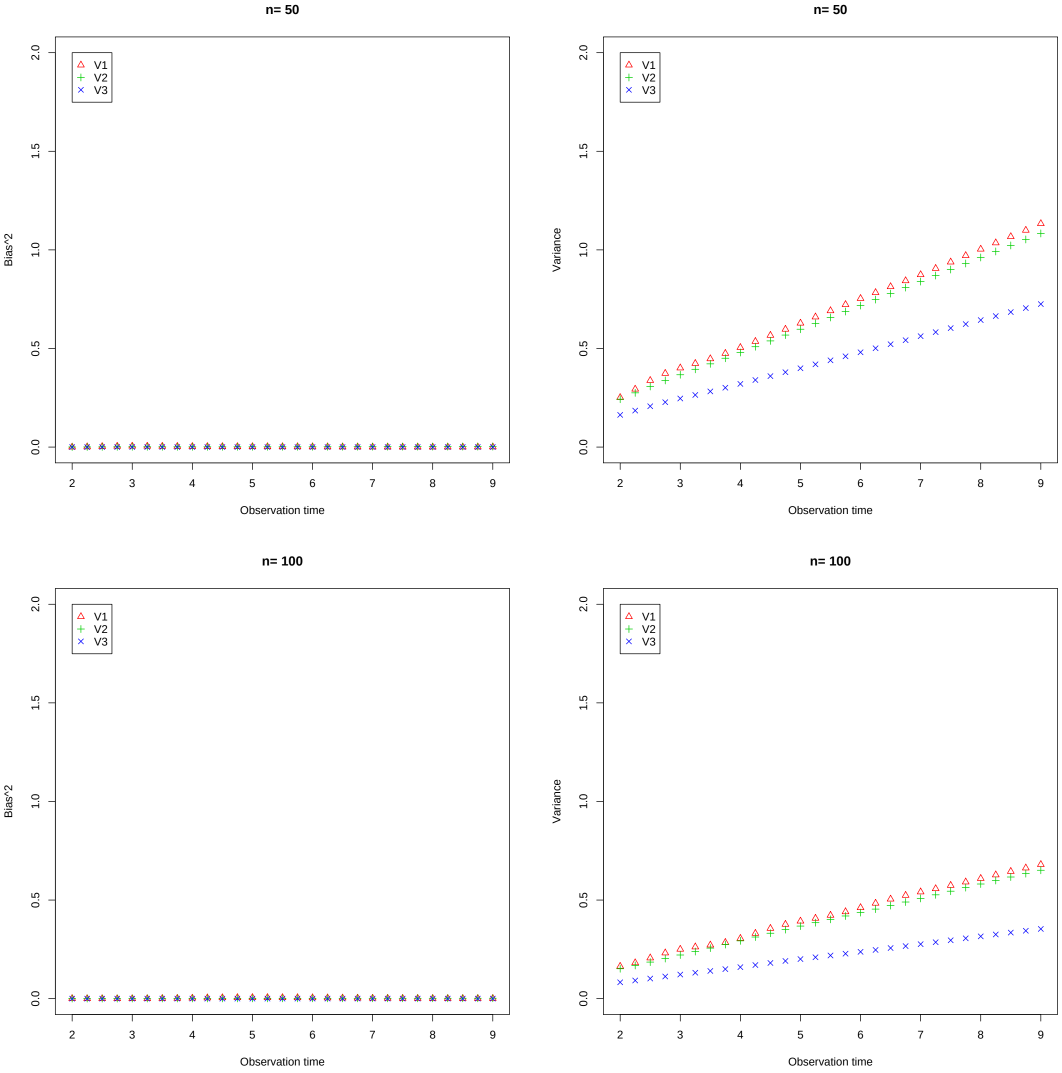

Figure 6.2: Scenario 2, with Data from the Gamma Frailty Poisson Model: $\Lambda_0\left(t\right) = 2t^{1/2}$

Table 6.5: Monte-Carlo simulations results of the B-splines based sieve GEE estimator using different covariance matrix for Mixture Poisson data with sample size n=50

	Bias			M-C sd			Projection ASE			95% coverage		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
$V_1^{(i)}$	0.0043	0.0017	0.0012	0.0810	0.2540	0.1443	0.0564	0.1999	0.1178	0.8010	0.8640	0.8750
$V_2^{(i)}$	0.0049	0.0010	0.0015	0.0809	0.2504	0.1407	0.0559	0.1991	0.1157	0.7980	0.8630	0.8760
$V_3^{(i)}$	0.0020	−0.0020	0.0029	0.0640	0.2015	0.1232	0.0549	0.1827	0.1107	0.8980	0.9140	0.9070

	GEE Sandwich ASE			95% coverage			Bootstrap ASE			95% coverage		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
$V_1^{(i)}$	0.0564	0.1999	0.1179	0.7990	0.8640	0.8750	0.0732	0.2312	0.1321	0.9050	0.9190	0.9150
$V_2^{(i)}$	0.0559	0.1990	0.1156	0.7980	0.8630	0.8740	0.0728	0.2307	0.1297	0.9080	0.9260	0.9170
$V_3^{(i)}$	0.0546	0.1800	0.1098	0.8980	0.9150	0.9070	0.0649	0.2045	0.1208	0.9370	0.9390	0.9340

Table 6.6: Monte-Carlo simulations results of the B-splines based sieve GEE estimator using different covariance matrix for Mixture Poisson data with sample size n=100

	Bias			M-C sd			Projection ASE			95% coverage		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
$V_1^{(i)}$	-0.0023	0.0016	0.0022	0.0601	0.1840	0.1015	0.0454	0.1565	0.0908	0.8299	0.8973	0.9170
$V_2^{(i)}$	-0.0026	-0.0001	0.0027	0.0606	0.1812	0.0981	0.0452	0.1556	0.0892	0.8299	0.8973	0.9232
$V_3^{(i)}$	-0.0019	0.0007	0.0035	0.0425	0.1384	0.0805	0.0405	0.1317	0.0798	0.9336	0.9378	0.9492
	GEE Sandwich ASE			95% coverage			Bootstrap ASE			95% coverage		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
$V_1^{(i)}$	0.0454	0.1565	0.0908	0.8320	0.8973	0.9170	0.0529	0.1675	0.0954	0.9035	0.9212	0.9274
$V_2^{(i)}$	0.0452	0.1556	0.0892	0.8299	0.8973	0.9232	0.0528	0.1668	0.0939	0.8994	0.9315	0.9336
$V_3^{(i)}$	0.0405	0.1314	0.0796	0.9305	0.9367	0.9471	0.0434	0.1384	0.0830	0.9429	0.9419	0.9585

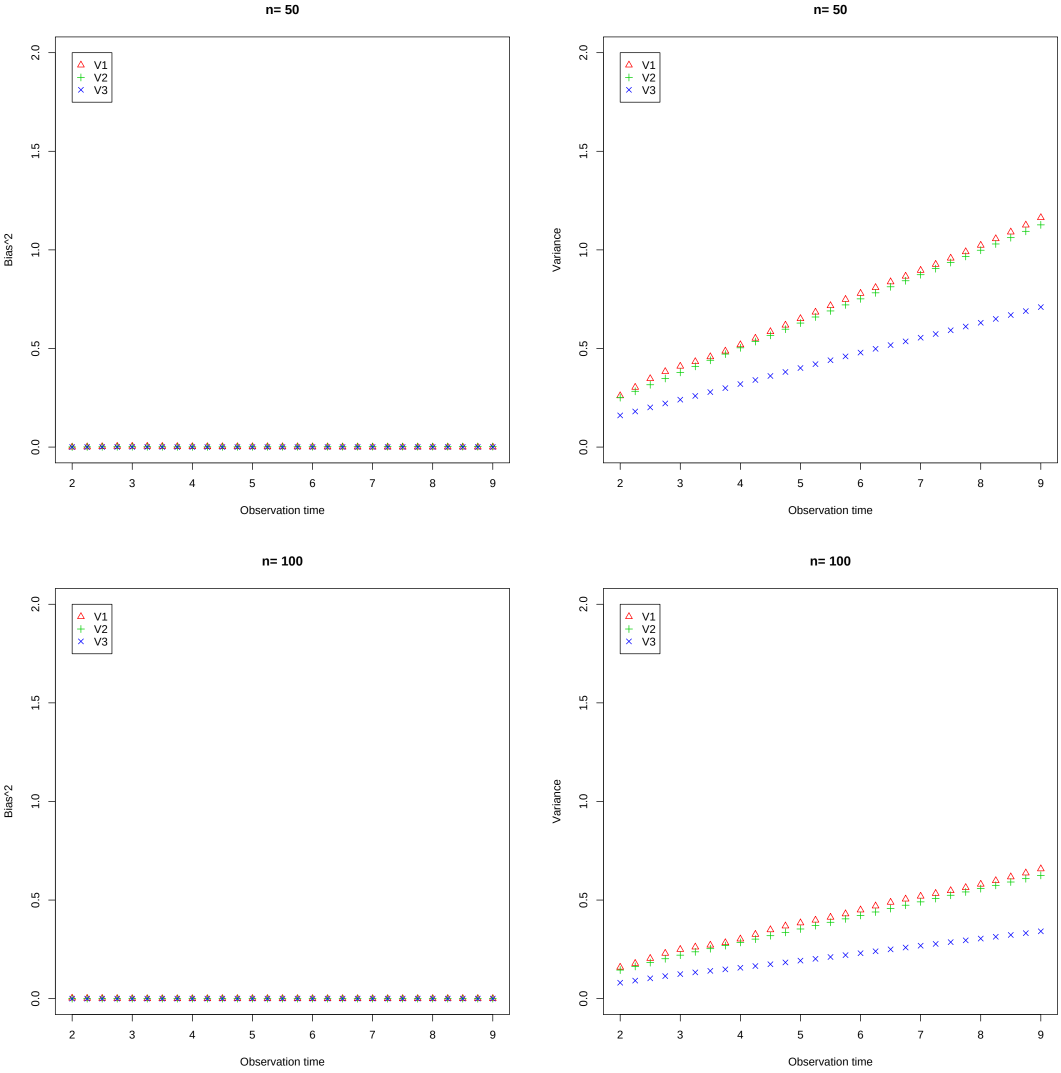

Figure 6.3: Scenario 3, with Data from the Mixture Poisson Model: $\Lambda_0(t) = 2t^{1/2}$

Table 6.7: Monte-Carlo simulations results of the B-splines based sieve GEE estimator using different covariance matrix for Negative Binomial data with sample size n=50

	Bias			M-C sd			Projection ASE			95% coverage		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
$V_1^{(i)}$	-0.0004	0.0014	-0.0038	0.0357	0.1180	0.0799	0.0312	0.0991	0.0738	0.9070	0.8950	0.9190
$V_2^{(i)}$	-0.0004	0.0020	-0.0038	0.0344	0.1100	0.0743	0.0296	0.0947	0.0702	0.9010	0.9090	0.9240
$V_3^{(i)}$	-0.0008	0.0017	-0.0037	0.0350	0.1120	0.0749	0.0302	0.0958	0.0707	0.9040	0.9110	0.9220
	GEE Sandwich ASE			95% coverage			Bootstrap ASE			95% coverage		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
$V_1^{(i)}$	0.0312	0.0991	0.0738	0.9070	0.8950	0.9190	0.0390	0.1178	0.0822	0.9630	0.9380	0.9480
$V_2^{(i)}$	0.0296	0.0947	0.0701	0.9000	0.9080	0.9230	0.0369	0.1122	0.0780	0.9590	0.9480	0.9570
$V_3^{(i)}$	0.0302	0.0957	0.0706	0.9040	0.9100	0.9220	0.0383	0.1150	0.0791	0.9630	0.9490	0.9570

Table 6.8: Monte-Carlo simulations results of the B-splines based sieve GEE estimator using different covariance matrix for Negative Binomial data with sample size n=100

	Bias			M-C sd			Projection ASE			95% coverage		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
$V_1^{(i)}$	0.0008	−0.0019	0.0014	0.0248	0.0791	0.0568	0.0212	0.0707	0.0525	0.9120	0.9200	0.9200
$V_2^{(i)}$	0.0009	−0.0024	0.0016	0.0231	0.0744	0.0538	0.0200	0.0671	0.0496	0.9070	0.9280	0.9090
$V_3^{(i)}$	0.0009	−0.0026	0.0017	0.0236	0.0752	0.0540	0.0206	0.0679	0.0499	0.9110	0.9280	0.9160
	GEE Sandwich ASE			95% coverage			Bootstrap ASE			95% coverage		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
$V_1^{(i)}$	0.0212	0.0707	0.0525	0.9120	0.9200	0.9200	0.0246	0.0771	0.0553	0.9480	0.9360	0.9330
$V_2^{(i)}$	0.0200	0.0671	0.0496	0.9080	0.9280	0.9090	0.0232	0.0730	0.0522	0.9490	0.9420	0.9290
$V_3^{(i)}$	0.0206	0.0679	0.0499	0.9100	0.9270	0.9160	0.0242	0.0747	0.0527	0.9600	0.9490	0.9360

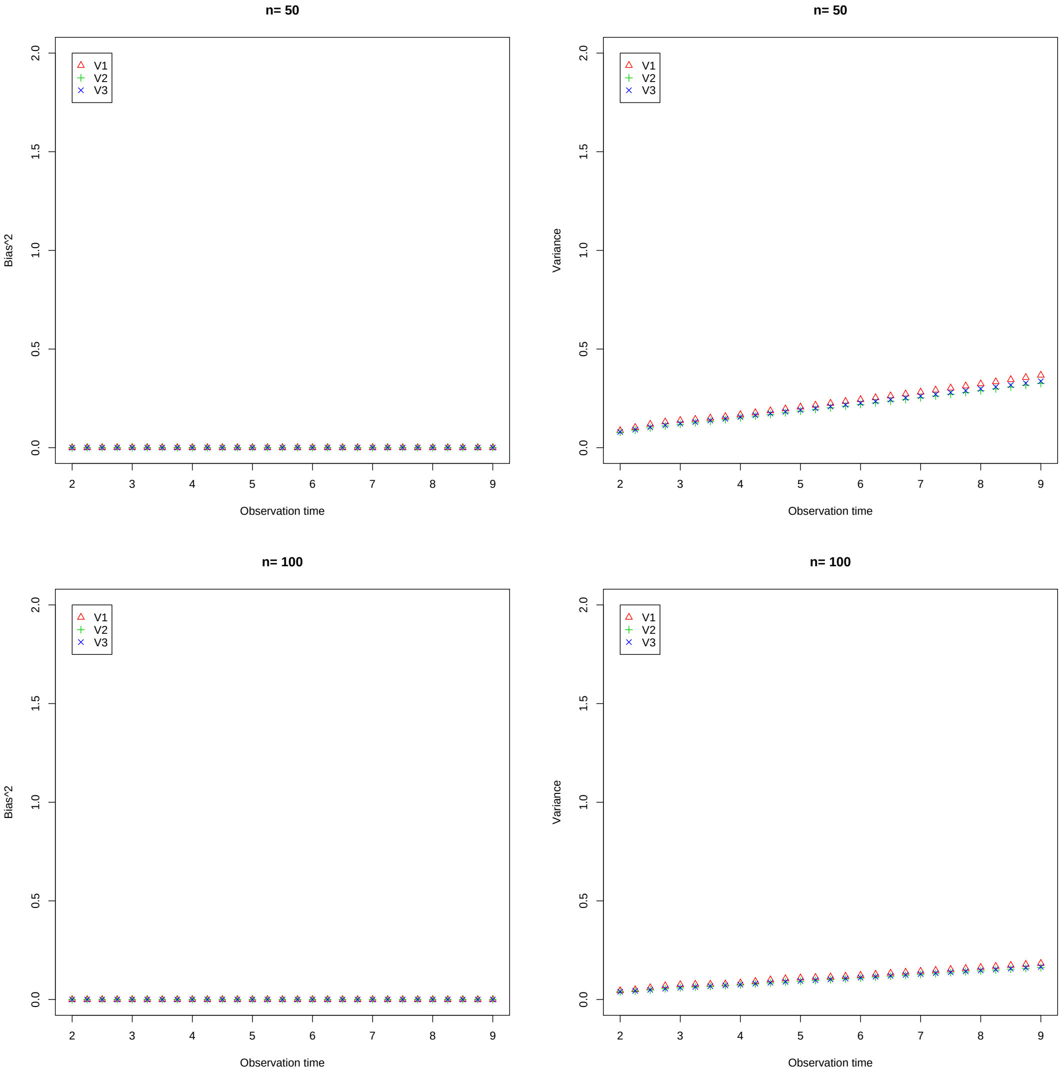

Figure 6.4: Scenario 4, with Data from the Negative Binomial Model: $\Lambda_0(t) = 2t^{1/2}$

6.2 Comparison of different algorithms

6.2.1 Comparison among ICM, NR/IR and GR algorithms

We conducted a simulation study comparing the computing time among the three algorithms, ICM, NR/IR and GR, in solving the spline-based sieve GEE using all three working covariance matrices. The comparison is based on 10 datasets of each of the four scenarios described in the previous section. Simulation results are shown in Table 6.9. In all these scenarios, the hybrid algorithm NR/IR is the most efficient in terms of computing time, followed by GR. ICM algorithm is a lot slower than the other two methods in solving the spline-based sieve GEE in all 4 scnarios.

6.2.2 Comparison of different over-dispersion estimation methods

To compare the effect of different over-dispersion estimation methods on the estimates of the regression parameter, we conducted a simulation study using MLE and the two methods of moment described in Section 5.2. Table 6.10 shows the simulation results. When data are generated from the Gamma-frailty nonhomogeneous Poisson process as in Scenarios 2, the maximum likelihood estimator of the regression parameter has a slightly smaller standard error. When data are generated from the nonhomogeneous Poisson process, the Mixture Poisson process or the Negative binomial process as in Scenarios 1, 3 and 4 respectively, all three methods give similar results.

Table 6.9: Comparison of the average computing time in seconds among ICM, NR/IR and GR

		Poisson			Mixture Poisson			Gamma Frailty			Negative Binomial		
		ICM	NR/IR	GR	ICM	NR/IR	GR	ICM	NR/IR	GR	ICM	NR/IR	GR
n=50	$V_1^{(i)}$	9.78	1.50	2.34	11.82	1.55	1.95	11.82	1.53	2.19	12.16	1.81	2.62
	$V_2^{(i)}$	20.86	2.50	3.83	21.15	2.60	4.30	20.58	2.55	3.92	26.32	3.07	5.26
	$V_3^{(i)}$	10.21	4.14	5.01	20.82	5.72	5.81	67.82	4.80	4.85	11.68	3.78	5.12
n=100	$V_1^{(i)}$	19.25	3.17	3.80	25.75	3.27	4.52	20.02	3.13	5.44	19.79	2.98	4.89
	$V_2^{(i)}$	131.34	5.11	7.47	93.42	5.09	7.81	91.62	5.19	8.90	108.72	4.90	8.31
	$V_3^{(i)}$	14.90	4.97	7.15	76.53	10.24	10.62	29.33	9.66	9.91	20.72	6.09	7.52

Table 6.10: Standard deviation of the regression parameters using different methods estimating the overdipsersion parameter

	Mean standard error				Mean standard error			
Stand error	β_1	β_2	β_3	$\overline{\hat{\sigma}_n^2}(s.e.)$	β_1	β_2	β_3	$\overline{\hat{\sigma}_n^2}(s.e.)$
	Poisson				Gamma-Frailty Poisson			
Zeger's method	0.0310	0.0966	0.0634	0.0006(0.0021)	0.0652	0.1950	0.1191	0.0625(0.0468)
Breslow's method	0.0318	0.0994	0.0641	0.0040(0.0085)	0.0648	0.1942	0.1188	0.0757(0.0324)
MLE	0.0309	0.0969	0.0636	0.0015(0.0018)	0.0648	0.1942	0.1188	0.0685(0.0252)
	Mixture Poisson				Negative Binomial			
Zeger's method	0.0633	0.1973	0.1229	0.0651(0.0433)	0.0340	0.1089	0.0781	0.0023(0.0045)
Breslow's method	0.0625	0.1959	0.1216	0.0773(0.0317)	0.0354	0.1127	0.0800	0.0158(0.0209)
MLE	0.0626	0.1957	0.1217	0.0733(0.0252)	0.0338	0.1085	0.0784	0.0040(0.0058)

6.3 Application To A Real Data

The proposed estimating method is applied to the bladder tumor data introduced in the Section 1. A total of 116 patients were randomized into three treatment groups, with 31 using pyridoxin pills, 38 instilled with thiotepa and 47 in placebo group. Their follow-up times vary from one week to sixty-four weeks. Four variables, including the number (Z_1) and size (Z_2) of the tumor at baseline, and two indicator variables, one for pyridoxin (Z_3), one for thiotepa (Z_4), are included in the proportional mean model, i.e.,

$$E(\mathbb{N}(t)|Z_1, Z_2, Z_3, Z_4) = \Lambda_0\,(t)\,exp\,(\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4)$$

Regression analysis results using the three covariance matrices are shown in Table 6.11. The number of the tumors at the entrance of the study is positively related to the recurrence of the bladder tumor. With one more tumor at their diagnosis, the number of tumors at follow-up visits increases by 15.5%, 23.1% and 39.1% on average using covariance matrix $V_1^{(i)}, V_2^{(i)}$ and $V_3^{(i)}$ respectively. Thiotepa instillation effectively decreases the number of recurrent tumors. The number of recurrent tumor in patients with thiotepa instillation is 49.5%, 45.1% and 32.5% of those in control group using $V_1^{(i)}, V_2^{(i)}$ and $V_3^{(i)}$, respectively. The size of tumors and pyridoxin pills are not significantly related to the number of recurrent tumors at follow-up visits. The results using the diagonal covariance matrix ($V_1^{(i)}$) and the covariance matrix based on Poisson process ($V_2^{(i)}$) are consistent with those based on the sieve pseudolikelihood and the sieve likelihood methods proposed by Lu et al. (2009). The spline-based sieve semiparametric GEE estimates using the frailty Poisson covariance matrix ($V_3^{(i)}$)

provides an estimate of the over-dispersion parameter as 1.32. It implies the over-dispersion of the panel count and the potential positive correlation between non-overlapping increments in the counting process. The effect of the number of the tumors at the study entrance and the treatment of thiotepa are more significant when accounting for the correlation between cumulative counts using the frailty variable.

Table 6.11: The spline-based sieve semiparametric inference for bladder tumor data

	V_1			V_2			V_3 ($\hat{\sigma}_n^2 = 1.29$)		
	Est.	Std.	p-value	Est.	Std.	p-value	Est.	Std.	p-value
Z1	0.1444	0.0518	0.0053	0.2075	0.0677	0.0022	0.3289	0.0702	0.0000
Z2	−0.0447	0.0488	0.3595	−0.0353	0.0732	0.6299	0.0054	0.0767	0.9437
Z3	0.1776	0.2246	0.4292	0.0637	0.3502	0.8556	0.0213	0.4069	0.9583
Z4	−0.6966	0.2397	0.0037	−0.7960	0.2952	0.0070	−1.0692	0.3389	0.0016

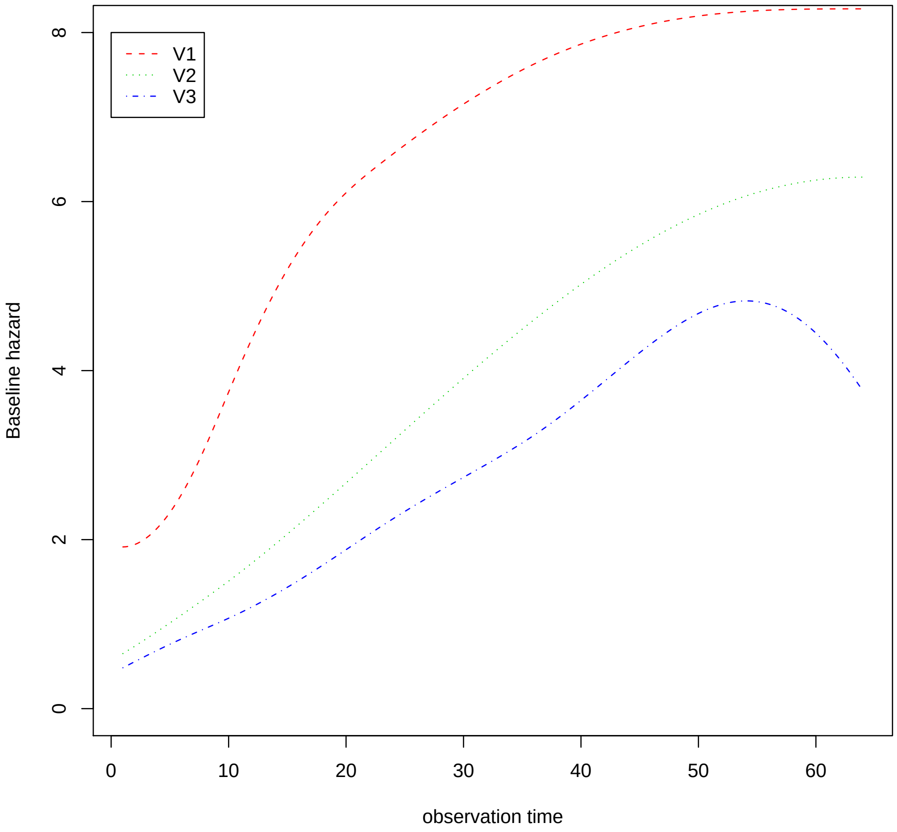

Figure 6.5: Bladder tumor: Estimates of baseline mean function based on different working covariance matrices

CHAPTER 7
DISCUSSIONS

In this dissertation we propose to analyze panel count data using the generalized estimating equation method with the semiparametric proportional mean model. With limited choices of the available counting processes, the proposed method provides a more flexible approach to model the data. As emphasized in the previous chapters, different covariance matrices could be adopted to allow for different data structures. The covariance matrix that captures the true correlation between the repeated measures increases the efficiency of the estimation. Similar idea could also be generalized to a multivariate panel count data setting. More complicated covariance matrices could be used to account for the correlations between multiple levels.

In the proof of the asymptotic properties of the proposed method, we make use of the fact that the generalized estimating equations coincide with scores of different likelihood functions. A maximum pseudo likelihood estimating procedure is applied to solve the estimating equations using $V_3^{(i)}$ as the covariance matrix. The estimators are still consistent and converge at the same convergence rate as the maximum likelihood estimators. In general, the estimator that maximizes pseudo likelihood has a bigger variance than the true maximum likelihood estimator. However in the proposed Gamma-Frailty Poisson model, the estimated parameters in the semiparametric proportional mean model and the estimated over-dispersion parameter are asymptotically independent, which attributes to the fact that the estimated regression parameter based on the pseudo likelihood has the same asymptotic variance as the maximum likelihood estimator. This property serves as the underpin of the two-stage algorithm.

The cubic B-spline estimator of the baseline mean function improves the convergence rate compared to the estimation using the traditional nonparametric step functions. At the same time, it decreases the dimension of the estimation which contributes to the computing efficiency.

Despite the fact that different working covariance matrices could be used in the generalized estimating equation, how to construct them so that they represent the true covariance matrix require more efforts. We show that the estimation using $V_3^{(i)}$ as the working covariance matrix generally outperforms the estimates using $V_1^{(i)}$ or $V_2^{(i)}$. $V_3^{(i)}$ is constructed under the assumption that conditioning on the frailty term, the increments are independent. It certainly covers a broader model than the methods using $V_1^{(i)}$ and $V_2^{(i)}$. However this assumption may still be unrealistic in view of medical applications. As future research, it would be useful to investigate how to relax this assumption and possibly to incorporate the autoregressive structures into the covariance matrix.

APPENDIX A
AGREEMENT OF GEE AND SCORE FUNCTIONS

In the derivation of the equivalence between GEE using $V_1^{(i)}, V_2^{(i)}$ and $V_3^{(i)}$ and different 'likelihood' functions, we adopt the following notations

$$\begin{aligned}
B^{(i)}_{K_i,j} &= \left(B_1\left(T^{(i)}_{K_i,j}\right), \cdots, B_{q_n}\left(T^{(i)}_{K_i,j}\right)\right)^T; & B^{(i)} &= \left(B^{(i)}_{K_i,1}, \cdots, B^{(i)}_{K_i,K_i}\right)^T \\
\mu^{(i)}_{K_i,j} &= exp\left(\beta^T Z_i + \alpha^T B^{(i)}_{K_i,j}\right); & \mu^{(i)} &= \left(\mu^{(i)}_{K_i,1}, \cdots, \mu^{(i)}_{K_i,K_i}\right)^T \\
\Delta\mu^{(i)}_{K_i,j} &= \mu^{(i)}_{K_i,j} - \mu^{(i)}_{K_i,j-1}; & \Delta\mu^{(i)} &= \left(\Delta\mu^{(i)}_{K_i,1}, \cdots, \Delta\mu^{(i)}_{K_i,K_i}\right)^T \\
\Delta\mathbb{N}^{(i)}_{K_i,j} &= \mathbb{N}\left(T^{(i)}_{K_i,j}\right) - \mathbb{N}\left(T^{(i)}_{K_i,j-1}\right); & \Delta\mathbb{N}^{(i)} &= \left(\mathbb{N}^{(i)}_{K_i,1}, \cdots, \mathbb{N}^{(i)}_{K_i,K_i}\right)^T
\end{aligned}$$

Also let $1_{K_i} = (1, 1, \cdots, 1)^T_{K_i \times 1}$, we have

$$\begin{aligned}
\frac{\partial \mu^{(i)}_{K_i,j}}{\partial \theta} &= exp\left(\beta^T Z_i + \alpha^T B^{(i)}_{K_i,j}\right)\left(Z_i^T, B^{(i)T}_{K_i,j}\right)^T; \\
\frac{\partial \mu^{(i)}}{\partial \theta} &= \left(\frac{\partial \mu^{(i)}_{K_i,1}}{\partial \theta}, \cdots, \frac{\partial \mu^{(i)}_{K_i,K_i}}{\partial \theta}\right)^T = diag\left(\mu^{(i)}_{K_i,1}, \cdots, \mu^{(i)}_{K_i,K_i}\right)\left(1_{K_i} Z_i^T, B^{(i)}\right)
\end{aligned}$$

A.1 Agreement between sieve GEE using $V_1^{(i)}$ and the score of the sieve pseudolikelihood

Using $V_1^{(i)}$ as the working covariance matrix, Equation (2.6) can be rewritten as

$$\begin{aligned}
U(\theta; D) &= \sum_{i=1}^{n} \left(1_{K_i} Z_i^T, B^{(i)}\right)^T diag\left(\mu^{(i)}_{K_i,1}, \cdots, \mu^{(i)}_{K_i,K_i}\right) \times \\
&\qquad \left(diag\left(\mu^{(i)}_{K_i,1}, \cdots, \mu^{(i)}_{K_i,K_i}\right)\right)^{-1}\left(\mathbb{N}(T_i) - \mu^{(i)}\right) \\
&= \sum_{i=1}^{n} \left(1_{K_i} Z_i^T, B^{(i)}\right)^T \left(\mathbb{N}(T_i) - \mu^{(i)}\right)
\end{aligned}$$

This is the score function based on the pseudolikelihood shown in Equation (2.3).

A.2 Agreement between sieve GEE using $V_2^{(i)}$ and the score of the sieve likelihood

When using $V_2^{(i)}$ as the working covariance matrix, the estimating equation from Equation (2.6) can be rewritten as

$$U(\theta) = \sum_{i=1}^{n} \left(1_{K_i} Z_i^T, B^{(i)}\right)^T diag\left(\mu_{K_i,1}^{(i)}, \cdots, \mu_{K_i,K_i}^{(i)}\right) V_i^{(2)^{-1}} \left(\mathbb{N}^{(i)} - \mu^{(i)}\right)$$

A careful examination of the likelihood function in Equation (2.5) shows its score function can be rewritten in a matrix form,

$$\frac{\partial}{\partial\theta}\tilde{l}_n(\theta; D) = \sum_{i=1}^{n} \left(\frac{\partial\Delta\mu^{(i)}}{\partial\theta}\right)^T \left(diag\left(\Delta\mu_{K_i,1}^{(i)}, \cdots, \Delta\mu_{K_i,K_i}^{(i)}\right)\right)^{-1} \left(\Delta\mathbb{N}^{(i)} - \Delta\mu^{(i)}\right)$$

Since

$$\begin{aligned}
\frac{\partial\Delta\mu_{K_i,j}^{(i)}}{\partial\theta} &= \mu_{K_i,j}^{(i)}\left(Z_i^T, B_{K_i,j}^{(i)^T}\right)^T - \mu_{K_i,j-1}^{(i)}\left(Z_i^T, B_{K_i,j-1}^{(i)^T}\right)^T \\
&= \left\{\left(-\mu_{K_i,j-1}^{(i)}, \mu_{K_i,j}^{(i)}\right)\begin{pmatrix} Z_i^T & B_{K_i,j-1}^{(i)^T} \\ Z_i^T & B_{K_i,j}^{(i)^T}\end{pmatrix}\right\}^T \\
\frac{\partial\Delta\mu^{(i)}}{\partial\theta} &= \left(\frac{\partial\Delta\mu_{K_i,1}^{(i)}}{\partial\theta}, \cdots, \frac{\partial\Delta\mu_{K_i,K_i}^{(i)}}{\partial\theta}\right)^T \\
&= \begin{pmatrix} \mu_{K_i,1}^{(i)} & 0 & \cdots & 0 \\ -\mu_{K_i,1}^{(i)} & \mu_{K_i,2}^{(i)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & -\mu_{K_i,K_i-1}^{(i)} & \mu_{K_i,K_i}^{(i)} \end{pmatrix} \left(1_{k_i} Z_i^T, B^{(i)}\right) \\
&= \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & -1 & 1 \end{pmatrix} diag\left(\mu_{K_i,1}^{(i)}, \cdots, \mu_{K_i,K_i}^{(i)}\right)\left(1_{k_i} Z_i^T, B^{(i)}\right)
\end{aligned}$$

The score function can be further written as

$$\frac{\partial}{\partial\theta}\tilde{l}_n\left(\theta;D\right)=\sum_{i=1}^{n}\left(1_{K_i}Z_i^T,B^{(i)}\right)^T diag\left(\mu_{K_i,1}^{(i)},\cdots,\mu_{K_i,K_i}^{(i)}\right)\Sigma\left(\mathbb{N}^{(i)}-\mu^{(i)}\right)$$

Where

$$\begin{aligned}
\Sigma &= \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & -1 & 1 \end{pmatrix}^T diag\left(\Delta\mu_{K_i,1}^{(i)},\cdots,\Delta\mu_{K_i,K_i}^{(i)}\right)^{-1}\begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & -1 & 1 \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{\mu_{K_i,1}^{(i)}} & -\frac{1}{\mu_{K_i,2}^{(i)}-\mu_{K_i,1}^{(i)}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\mu_{K_i,2}^{(i)}-\mu_{K_i,1}^{(i)}} & -\frac{1}{\mu_{K_i,3}^{(i)}-\mu_{K_i,2}^{(i)}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & -\frac{1}{\mu_{K_i,K_i}^{(i)}-\mu_{K_i,K_i}^{(i)}} \\ 0 & 0 & 0 & \cdots & \frac{1}{\mu_{K_i,K_i}^{(i)}-\mu_{K_i,K_i}^{(i)}} \end{pmatrix}\begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & -1 & 1 \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{\mu_{K_i,1}^{(i)}}+\frac{1}{\mu_{K_i,2}^{(i)}-\mu_{K_i,1}^{(i)}} & -\frac{1}{\mu_{K_i,2}^{(i)}-\mu_{K_i,1}^{(i)}} & \cdots & \cdots & 0 \\ -\frac{1}{\mu_{K_i,2}^{(i)}-\mu_{K_i,1}^{(i)}} & \frac{1}{\mu_{K_i,2}^{(i)}-\mu_{K_i,1}^{(i)}}+\frac{1}{\mu_{K_i,3}^{(i)}-\mu_{K_i,2}^{(i)}} & -\frac{1}{\mu_{K_i,3}^{(i)}-\mu_{K_i,2}^{(i)}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & \frac{1}{\mu_{K_i,K_i}^{(i)}-\mu_{K_i,K_i-1}^{(i)}} \end{pmatrix}
\end{aligned}$$

It is easy to verify Σ is the inverse of $V_2^{(i)}$, so sieve GEE using the covariance matrix $V_2^{(i)}$ is the same as the score function of the likelihood function in Equation (2.5).

A.3 Agreement between sieve GEE using $V_3^{(i)}$ and the score of the likelihood of Gamma-Frailty Poisson model

By the derivation of the equivalence between sieve GEE using $V_2^{(i)}$ and the score function of likelihood in Equation (2.5), we have

$$\left(\frac{\partial\mu^{(i)}}{\partial\theta}\right)^T V_2^{(i)^{-1}}\left(\mathbb{N}^{(i)}-\mu^{(i)}\right)=\sum_{j=1}^{K_i}\left(\frac{\partial\Delta\mu_{K_i,j}^{(i)}}{\partial\theta}\right)\left(\frac{\Delta\mathbb{N}_{K_i,j}^{(i)}}{\Delta\mu_{K_i,j}^{(i)}}-1\right)$$

This equality holds for any nonnegative nondecreasing process $\mathbb{N}^{(i)}$. Let $\mathbb{N}^{(i)}=2\mu^{(i)}$ then

$$\left(\frac{\partial\mu^{(i)}}{\partial\theta}\right)^T V_2^{(i)^{-1}}\mu^{(i)}=\sum_{j=1}^{K_i}\frac{\partial\Delta\mu_{K_i,j}^{(i)}}{\partial\theta}=\frac{\partial\mu_{K_i,K_i}^{(i)}}{\partial\theta}=\left(Z_i^T,B_{K_i,K_i}^{(i)T}\right)^T\mu_{K_i,K_i}^{(i)} \tag{A.1}$$

Also

$$\left(\frac{\partial\mu^{(i)}}{\partial\beta}\right)^T V_2^{(i)^{-1}}\left(\mathbb{N}^{(i)}-\mu^{(i)}\right)=\sum_{j=1}^{K_i}\left(\frac{\partial\Delta\mu_{K_i,j}^{(i)}}{\partial\beta}\right)\left(\frac{\Delta\mathbb{N}_{K_i,j}^{(i)}}{\Delta\mu_{K_i,j}^{(i)}}-1\right)$$

The left hand side (LHS) can be rewritten as

$$\begin{aligned} LHS &= Z_i 1_{K_i}^T diag\left(\mu_{K_i,1}^{(i)}, \cdots, \mu_{K_i,K_i}^{(i)}\right) V_2^{(i)^{-1}} \left(\mathbb{N}^{(i)} - \mu^{(i)}\right) \\ &= Z_i \mu^{(i)^T} V_2^{(i)^{-1}} \left(\mathbb{N}^{(i)} - \mu^{(i)}\right) \end{aligned}$$

The right hand side (RHS) can also be rewritten as

$$\begin{aligned} RHS &= \sum_{j=1}^{K_i} \left(\mu_{K_i,j}^{(i)} Z_i - \mu_{K_i,j-1}^{(i)} Z_i\right) \left(\frac{\Delta\mathbb{N}_{K_i,j}^{(i)}}{\Delta\mu_{K_i,j}^{(i)}} - 1\right) \\ &= Z_i \left(\mathbb{N}_{K_i,K_i}^{(i)} - \mu_{K_i,K_i}^{(i)}\right) \end{aligned}$$

Therefore

$$\begin{aligned} Z_i \mu^{(i)^T} V_2^{(i)^{-1}} \left(\mathbb{N}^{(i)} - \mu^{(i)}\right) &= Z_i \left(\mathbb{N}_{K_i,K_i}^{(i)} - \mu_{K_i,K_i}^{(i)}\right) \\ \Rightarrow \quad \mu^{(i)^T} V_2^{(i)^{-1}} \left(\mathbb{N}^{(i)} - \mu^{(i)}\right) &= \mathbb{N}_{K_i,K_i}^{(i)} - \mu_{K_i,K_i}^{(i)} \end{aligned} \tag{A.2}$$

Again if we let $\mathbb{N}^{(i)} = 2\mu^{(i)}$ then

$$\mu^{(i)} V_2^{(i)^{-1}} \mu^{(i)} = \mu_{K_i,K_i}^{(i)} \tag{A.3}$$

The estimating equation with $V_3^{(i)}$ as the covariance matrix can then be written as,

$$\begin{aligned} U(\theta; X_i) &= \left(\frac{\partial\mu^{(i)}}{\partial\theta}\right)^T \left(V_2^{(i)} + \sigma^2 \mu^{(i)} \mu^{(i)^T}\right)^{-1} \left(\mathbb{N}^{(i)} - \mu^{(i)}\right) \\ &= \left(\frac{\partial\mu^{(i)}}{\partial\theta}\right)^T \left(V_2^{(i)} - \frac{\sigma^2}{1 + \sigma^2 \mu^{(i)^T} V_2^{-1} \mu^{(i)}} V_2^{-1} \mu^{(i)} \mu^{(i)^T} V_2^{-1}\right) \left(\mathbb{N}(T_i) - \mu^{(i)}\right) \\ &= \left(\frac{\partial\mu^{(i)}}{\partial\theta}\right)^T V_2^{(i)^{-1}} \left(\mathbb{N}^{(i)} - \mu^{(i)}\right) - \frac{\sigma^2}{1 + \sigma^2 \mu^{(i)^T} V_2^{-1} \mu^{(i)}} \left(\frac{\partial\mu^{(i)}}{\partial\theta}\right)^T V_2^{(i)^{-1}} \mu^{(i)} \times \\ &\quad \mu^{(i)^T} V_2^{(i)^{-1}} \left(\mathbb{N}^{(i)} - \mu^{(i)}\right) \\ &= \sum_{j=1}^{K_i} \left(\mu_{K_i,j}^{(i)} \left(Z_i^T, B_{K_i,j}^{(i)^T}\right)^T - \mu_{K_i,j-1}^{(i)} \left(Z_i^T, B_{K_i,j-1}^{(i)^T}\right)^T\right) \left(\frac{\Delta\mathbb{N}_{K_i,j}^{(i)}}{\Delta\mu_{K_i,j}^{(i)}} - 1\right) - \\ &\quad \frac{\sigma^2}{1 + \sigma^2 \mu_{K_i,K_i}^{(i)}} \left(Z_i^T, B_{K_i,K_i}^{(i)^T}\right)^T \mu_{K_i,K_i}^{(i)} \left(\mathbb{N}_{K_i,K_i}^{(i)} - \mu_{K_i,K_i}^{(i)}\right) \\ &\quad \text{(by Equations (A.1)-(A.3))} \\ &= \sum_{j=1}^{K_i} \left(\mu_{K_i,j}^{(i)} \left(Z_i^T, B_{K_i,j}^{(i)^T}\right)^T - \mu_{K_i,j-1}^{(i)} \left(Z_i^T, B_{K_i,j-1}^{(i)^T}\right)^T\right) \frac{\Delta\mathbb{N}_{K_i,j}^{(i)}}{\Delta\mu_{K_i,j}^{(i)}} - \\ &\quad \frac{1 + \sigma^2 \mathbb{N}_{K_i,K_i}^{(i)}}{1 + \sigma^2 \mu_{K_i,K_i}^{(i)}} \left(Z_i^T, B_{K_i,K_i}^{(i)^T}\right)^T \mu_{K_i,K_i}^{(i)} \end{aligned}$$

This is exactly the score function of the Gamma-frailty Poisson likelihood in Equation (2.7).

APPENDIX B
R FUNCTIONS FOR NUMERICAL RESULTS

In this section, we list some important functions used for the simulation studies presented in Chapter 6.

Table B.1: Important Functions Used in Simulation Studies

Name	Function
genData	Generate data used in the simulation studies in section 6
covariate	Organize data into a matrix form
GEE†	Calculate the value of estimation equation as well as the sandwich form based on current parameter estimates
GR†	Generalized Rosen algorithm
ICM†	ICM algorithm
ICM-NR†	ICM/NR algorithm
sandwich†	Estimate the variance based on projection method using the pseudolikelihood approach
sigma.est1	Zeger's method of estimating overdispersion parameter
sigma.est2	Breslow's method of estimating overdispersion parameter
sigma.est3	MLE of overdispersion parameter assuming a gamma-distributed frailty term
V1	GEE covariance matrix, GEE using V_1 is the same as scores based on pseudolikelihood
V2	GEE covariance matrix, GEE using V_2 is the same as scores based on likelihood
V3	GEE covariance matrix, GEE using V_3 is the same as scores based on Gamma-frailty Poisson likelihood
semiResult†	Run simulations in batch
bsvar	Generate boostrap samples to get bootstrap variances

Note: † There are corresponding functions using frailty covariance matrix. They are similar to the listed functions. For the simplicity of the dissertation, they are omitted here.

```
#=============================================================================#
# Function name: genData                                                      #
#-----------------------------------------------------------------------------#
# It works with different definition of myfunc and myfunc.inv                 #
#=============================================================================#

genData<-function(method, scenario, n, beta=c(0,0,0), seed=0, sigma2.inv=2,
            distribution='gamma', r0=10, p=0.2, diff.K=0){
  z1 <- rnorm(n)
  z2 <- runif(n)
  z3 <- rbinom(n, 1, 1/2)
  z <- cbind(z1,z2,z3)

  proportion<-exp(z%*%beta)
# Create observation times
  T.schedule <- 2*(1:6)
  T.real <- matrix(round(rnorm(6*n, T.schedule, 1/3), digits=2), nrow=n, byrow=T)
  solveprob <- function(x){
    different <- c(x[1], diff(x))
    x[different<=0] <- 0
    x
    }
  T.real <- t(apply(T.real, 1, solveprob))
  miss <- matrix(rbinom(6*n, 1, exp(T.real-10)/(1+exp(T.real-10))),nrow=n, byrow=F)
  T.obs.pre <- T.real*(1-miss)

  count <- function(x) sum(x!=0)
  K <- apply(T.obs.pre, 1, count)
  T.obs <- matrix(NA, nrow=n, ncol=max(K))
# get the increment T
  dT.obs<-matrix(0,nrow=n, ncol=max(K))
  for (i in 1:n){
    T.obs[i,1:K[i]] <- T.obs.pre[i,T.obs.pre[i,]!=0]
    dT.obs[,1]<-T.obs[,1]
    for (j in 2:ncol(T.obs)) dT.obs[,j]<-T.obs[,j]-T.obs[,j-1]
    }
# scenario 1:  The panel counts are generated from Poisson Process
  if (scenario==1){
      dN <- matrix(nrow=n,ncol=max(K))
      for (i in 1:n){
        dN[i,1] <- rpois(1, myfunc(T.obs[i,1])*proportion[i])
        if (K[i]!=1){
          for (j in 2:K[i]){
            dN[i,j] <- rpois(1, (myfunc(T.obs[i,j])-myfunc(T.obs[i,j-1]))*
                         proportion[i])
            }
          }
        }
  }else if (scenario==2){
# scenario 2:  The panel counts are generated from a mixed Poisson process
  dN<-matrix(nrow=n, ncol=max(K))
  for (i in 1:n){
    gama<- c(-0.8,0,0,0.8)[floor(runif(1,min=1,max=5))]
    dN[i,1] <- rpois(1, (2+gama)/2*myfunc(T.obs[i,1])*proportion[i])
    if (K[i]!=1){
      for (j in 2:K[i]){
        dN[i,j] <- rpois(1, (2+gama)/2*(myfunc(T.obs[i,j])-myfunc(T.obs[i,j-1]))
                    *proportion[i])
        }
      }
    }
  }else if (scenario==3){
# scenario 3:  The panel counts are generated from a Gamma Frailty Poisson process
# we can also specify the frailty term from a lognormal distribution
  dN<-matrix(nrow=n, ncol=max(K))
  for (i in 1:n){
    if (distribution==’gamma’) gama<- rgamma(1, sigma2.inv, sigma2.inv) else
    if (distribution==’lognormal’) gama <- exp(rnorm(1, -1/2*log(1/sigma2.inv+1),
                                            sqrt(log(1/sigma2.inv+1))))
    dN[i,1] <- rpois(1, gama*myfunc(T.obs[i,1])*proportion[i])
    if (K[i]!=1){
      for (j in 2:K[i]){
        dN[i,j] <- rpois(1, gama*(myfunc(T.obs[i,j])-myfunc(T.obs[i,j-1]))
                    *proportion[i])
        }
      }
    }
   }else if (scenario==4){
# scenario 4: The panel count data are generated from a ’negative-binomialization’
#    of  the empirical counting process, from Zhang’s paper!
   dN <- matrix(nrow=n, ncol=max(K))
   r <- r0*proportion
   for (i in 1:n){
    total <- rnbinom(1,r[i], p)
    Fx <- runif(total)
    x <- myfunc.inv(r0*(1-p)/p*Fx)
    for (j in 1:K[i]){
      if (j==1) dN[i,j] <- sum(x<=T.obs[i,j]) else
      dN[i,j] <- sum(T.obs[i,j-1] <x & x<=T.obs[i,j])
      }
    }
  }
  list(K=K, T=T.obs, Z=z, dN=dN)
}

myfunc1 <- function(x) 2*x
myfunc1.inv <- function(x) x/2 #This is the inverse function of myfunc1
myfunc2 <- function(x) 2*x^(1/2)
myfunc2.inv <- function(x) (x/2)^2 #This is the inverse function of myfunc2

#myfunc2 and myfunc2.inv are the functions used in simulation studies in Section 6.
#We can easily change them to any baseline functions
#====================== End of Function (genData) ======================#

#=======================================================================#
# Function name: covariate                                              #
#-----------------------------------------------------------------------#
# It produces the covariate matrix                                      #
# the first qn column are B-spline values                               #
# the last 3 column are covariates                                      #
#=======================================================================#

covariate<-function(K, myT, Z, N, method, cumulative=1,n.knot=1/3,
    position.knot='quantile'){
  n<-length(K)
  myt<-myT[order(myT)]
  myt<-myt[!is.na(myt)]
  qn<-ceiling(sum(K)^n.knot)

  X<-matrix(nrow=sum(K), ncol=qn+3)
  if (position.knot=='quantile') myknots <-  c(rep(min(myt),3), quantile(myt,
    seq(0,1,len=qn-4+2)),rep(max(myt),3)) else
  if (position.knot=='uniform') myknots<- c(rep(min(myt),3), seq(min(myt), max(myt),
    length=qn-4+2), rep(max(myt),3))
  for (i in 1:n){
    b<-splineDesign(knots=myknots, x=myT[i,!is.na(myT[i,])])
    db<-matrix(0,nrow=nrow(b), ncol=ncol(b));
    db[1,]<-b[1,]
    if (nrow(b)>1){
      for (j in 2:nrow(b)) db[j,] <- b[j,]-b[j-1,]
      }

    z<-matrix(rep(Z[i,],K[i]),nrow=K[i], byrow=T)
    if (cumulative==0) x<-cbind(db, z) else
    x<-cbind(b, z)

    row.start<-ifelse(i==1, 1, cumsum(K)[i-1]+1)
    row.end<-cumsum(K)[i]
    X[row.start:row.end,]<-x
    }
  X2<-cbind(rep(1:n, K),t(myT)[!is.na(t(myT))], X, t(N)[!is.na(t(N))])
  colnames(X2)<-c('subj', 't',paste('B', 1:qn, sep=''), paste('Z', 1:3,sep=''),'N')
  if (method=='nonparametric') X2 <- X2[, -((1+1+qn+1):(1+1+qn+3))]
  list(X=X2, alpha.dim=qn)   #X2[,3:(qn+5)]
  }
#=== usage ===#
#cov<-covariate(K, T, Z, N)
#====================== End of Function (covariate) ====================#

#=======================================================================#
# Function name: GEE                                                    #
#-----------------------------------------------------------------------#
# It calculates the value of estimating equation and the sandwich form  #
# based on the current estimates of the parameter                       #
#=======================================================================#

# GEE calculate U and W conditioning on the current gama #
GEE<-function(gama, alpha.dim, dataset, method='semiparametric', varfunc){
      gama.dim <- length(gama)
      n <- length(unique(dataset[,'subj']))
      gee.table <- matrix(0, nrow=gama.dim,ncol=n)
      W<-matrix(0, nrow=gama.dim, ncol=gama.dim)
      for (i in 1:n){
        subj.dataset<-dataset[dataset[,'subj']==i,,drop=F]
        subj.n <- nrow(subj.dataset)
        subj.N <- subj.dataset[, 'N']
        subj.T <- subj.dataset[,'t']
        subj.covariate <- subj.dataset[, -c(1,2,2+gama.dim+1),drop=F]

        if (method=='nonparametric'){
          subj.mu <- subj.covariate%*%gama
          subj.dmu <- subj.covariate
          }else
        if (method=='semiparametric'){
          subj.mu <- exp(subj.covariate%*%gama)
          subj.dmu <- diag(as.vector(subj.mu), nrow=subj.n, ncol=subj.n)%*%
            subj.covariate
          }

        subj.result <- varfunc(subj.mu)
        if (subj.result$error!=1){
          subj.var <- subj.result$subj.var
          subj.var.inv <- subj.result$subj.var.inv

          gee.i <- t(subj.dmu)%*%subj.var.inv%*%(subj.N-subj.mu)
          gee.table[,i] <- gee.i
          W <- W+t(subj.dmu)%*%subj.var.inv%*%subj.dmu
          }
        }
      u<-apply(gee.table, 1, sum)
      list(U=u, W=W)
      }
#======================= End of Function (GEE) ===========================#

#=========================================================================#
# Function name: GR                                                       #
#-------------------------------------------------------------------------#
# It implements the Generalized Rosen algorithm to the GEE settings       #
#=========================================================================#

GR<-function(dataset, alpha.dim, gama.ini, method, varfunc, likelihood.func){
  converge.status <- 1
  error <- 0
  A1<-cbind(rep(0,alpha.dim-1), diag(1, nrow=alpha.dim-1))
  A2<-cbind(diag(-1,nrow=alpha.dim-1),rep(0,alpha.dim-1))
  A.ori<-cbind(A1+A2, matrix(rep(0,(alpha.dim-1)*beta.dim),ncol=beta.dim))

  gama<-gama.ini
  gama.dim<-length(gama)
  active.set<-numeric(length=0)
  active <- numeric(length=0)
  lamda<-rep(1, alpha.dim-1)
  A<-A.ori[active.set, ,drop=F]

  count2<-0
  while (max(lamda)>0){
  delta<-1
  count1<-0
  while (max(abs(delta))>=1e-5){
      active.set<-unique(append(active.set, active)[order(append(active.set,
        active))])
# Step 0 computing feasible search direction
      UW<-GEE(gama=gama, alpha.dim=alpha.dim, dataset=dataset, method=method,
        varfunc=varfunc)
      U<-UW$U
      W<-UW$W
      if (is.infinite(max(W))) {error <- 1; break; }
      if (missing(W)) {error<-1; break;}
      if (sum(is.na(W))>0) {error <-1 ; break;}
      if (min(abs(eigen(W)$values))<1e-5|max(abs(eigen(W)$values))>1e20 ) {
        error <- 1; break}
      W.inv<-solve(W)
```

```
   if (length(active.set)==0) d<- W.inv%*%U else{
    A<-A.ori[active.set, ,drop=F]
    d <- (diag(1, gama.dim, gama.dim)-
       W.inv%*%t(A)%*%solve(A%*%W.inv%*%t(A))%*%A
       )%*%W.inv%*%U}
# Step 1
   ratio<- -(A.ori%*%gama)/(A.ori%*%d)
   step <- ifelse(max(ratio,na.rm=T)<=0,1, min(ratio[ratio>0],na.rm=T))
# Step 2
   ksi<- min(step, 1)
   gama.update <- gama+ksi*d
   while (crossprod(GEE(gama=gama.update, alpha.dim=alpha.dim, dataset=dataset,
    method=method, varfunc=varfunc)$U) >
       crossprod(GEE(gama=gama, alpha.dim=alpha.dim, dataset=dataset,
       method=method, varfunc=varfunc)$U)){
    ksi <- ksi/2
    gama.update <- gama+ksi*d
    if (ksi<1e-5) break
    }
# Step 3  & Step 4
   if (step>ksi) delta<-ksi*d else{
    delta<-step*d
    active <- which(ratio==step)
    }
   gama<-gama+delta
   count1<-count1+1

   if (count1>20) {converge.status <- 0; break}
   if (missing(d)) break
   }
# Step 5: checking the stopping criterion
  if (length(active.set)==0) break else{
   lamda <-  solve(A%*%W.inv%*%t(A))%*%A%*%W.inv%*%U
   inactive <- which.max(lamda)
   active.set <- active.set[-inactive]
   }
  count2 <- count2+1
  if (count2>5) {
   converge.status <- 0
   break
   }
 }
 GEE.result <- GEE(gama=gama.update, alpha.dim=alpha.dim, dataset=dataset,
  method=method, varfunc=varfunc)
 W <- GEE.result$W
 U <- GEE.result$U
```

```
  sigma2 <- GEE.result$sigma2
  list(gama=gama.update, W=W, error=error, converge.status=converge.status,
    U=U, sigma2=sigma2)
}
#======================== End of Function (GR) ============================#

#==========================================================================#
# Function name: ICM                                                       #
#--------------------------------------------------------------------------#
# It implements the ICM algorithm to the GEE settings                      #
#==========================================================================#

ICM <- function(dataset, alpha.dim, gama.ini, method, varfunc, likelihood.func){
  beta.update <- gama.ini[-(1:alpha.dim)]
  alpha.update <- gama.ini[1:alpha.dim]
  d.beta <- 1
  beta.iter <- 0
  while(max(abs(d.beta))>1e-5){
    d.alpha<-1
    alpha.iter <- 0
    while(max(abs(d.alpha))>1e-5){
      gama <- c(alpha.update, beta.update)
      UW <- GEE(gama=gama, alpha.dim=alpha.dim, dataset=dataset, method=method,
        varfunc=varfunc)
      W <- UW$W
      U <- UW$U
      mydiag <- diag(W)[1:alpha.dim]
      x.axis <- c(0,cumsum(mydiag))
      y.axis <- c(0,cumsum(mydiag*alpha.update + U[1:alpha.dim]))
      ratio <- vector(length=length(x.axis)-1)
      i<-1
      while (i <length(x.axis)){
        derivative <- (y.axis[-(1:i)]-y.axis[i])/(x.axis[-(1:i)]-x.axis[i])
        position <- which.min(abs(derivative))
        ratio[i:(i+position-1)] <- min(abs(derivative))
        i<- i+position
        }
      ratio.update <- line.search(dataset, gama, ratio, beta.update, likelihood.func)
      d.alpha <- ratio.update-alpha.update
      alpha.update <- ratio.update
      alpha.iter <- alpha.iter+1
      if (alpha.iter>20) break
      }
    d.beta<- (solve(W)%*%U)[-(1:alpha.dim)]
    beta.update <- beta.update+ d.beta
    beta.iter <- beta.iter+1
    if (beta.iter>20) break
    }
  gama.update <- c(alpha.update, beta.update)
  list(gama=gama.update, W=W, U=U, alpha.iter=alpha.iter, beta.iter=beta.iter)
  }
#===================================
line.search <- function(dataset, gama, ratio, beta.update, likelihood.func){
  epsilon <- 0.4
  ksi <- 1
  alpha.dim <- length(ratio)
  while(likelihood.func(dataset, c(ratio, beta.update), alpha.dim)$l<
    likelihood.func(dataset, gama, alpha.dim)$l){
    ksi <- ksi/2
    ratio <- gama[1:alpha.dim] + ksi*(ratio-gama[1:alpha.dim])
    }
  ratio
  }
#======================== End of Function (ICM) ============================#

#===========================================================================#
# Function name: ICM-NR                                                     #
#---------------------------------------------------------------------------#
# It implements the ICM/NR algorithm to the GEE settings                    #
#===========================================================================#

# This is the modified ICM algorithm, combined with Newton-Raphson algrithm.
# we estimate alpha and beta through one loop!
ICM-NR <- function(dataset, alpha.dim, gama.ini, method, varfunc,
    likelihood.func=NA){
  gama <- gama.ini
  d.gama<-1
  iter <- 0
  while(max(abs(d.gama))>1e-5){
      UW <- GEE(gama=gama, alpha.dim=alpha.dim, dataset=dataset, method=method,
        varfunc=varfunc)
      W <- UW$W
      U <- UW$U
      mydiag <- diag(W)[1:alpha.dim]
      x.axis <- c(0,cumsum(mydiag))
      y.pre <- NR.gama(dataset, gama, alpha.dim, varfunc)
      y.axis <- c(0, cumsum(mydiag*y.pre[1:alpha.dim]))
      ratio <- vector(length=length(x.axis)-1)
      i<-1
      while (i <length(x.axis)){
        derivative <- (y.axis[-(1:i)]-y.axis[i])/(x.axis[-(1:i)]-x.axis[i])
        position <- which.min(abs(derivative))
        ratio[i:(i+position-1)] <- min(abs(derivative))
        i<- i+position
        }
      d.gama <- c(ratio, y.pre[-(1:alpha.dim)])-gama
      gama  <- c(ratio, y.pre[-(1:alpha.dim)])
      iter <- iter+1
      if (iter>5) break
    }
  list(gama=gama, W=W, U=U, iter=iter)
  }
#==================================
NR.gama <- function(dataset, gama, alpha.dim, varfunc){
  d<-1
  UW <-GEE(gama, alpha.dim, dataset, method='semiparametric', varfunc)
  if (min(eigen(UW$W)$values)>1e-10){
    direction <-solve(UW$W)%*%(UW$U)
    ksi <- 1
    gama.update <- gama + ksi*direction
    while(crossprod(GEE(gama=gama.update, alpha.dim=alpha.dim, dataset=dataset,
      varfunc=varfunc)$U) >
              crossprod(GEE(gama=gama, alpha.dim=alpha.dim, dataset=dataset,
                  varfunc=varfunc)$U)){
      ksi <- ksi/2
      gama.update<- gama+ksi*direction
      }
    d <- gama.update-gama
    gama <- gama.update
    }
  gama
}
#======================== End of Function (ICM/NR) ========================#

#==========================================================================#
# Function name: sandwich                                                  #
#--------------------------------------------------------------------------#
# It estimates the variance of the regression parameter based on the       #
# projection algorithm described in Wellner & Zhang (2007)                 #
#==========================================================================#

#=== using the least sqaure to calculate alpha         ===#
sandwich <- function(K, T, Z, dN, gama, varfunc){
  N<-matrix(nrow=nrow(dN), ncol=ncol(dN))
  n<-length(K)
  for (i in 1:n) N[i,]<-cumsum(dN[i,])
  cov<-covariate(K, T, Z, N, method)
  dataset<-cov$X
  alpha.dim<-cov$alpha.dim

  n <- length(unique(dataset[,'subj']))
  m1.matrix <- matrix(nrow=n, ncol=beta.dim)
  m2.star.matrix <- matrix(nrow=n, ncol=alpha.dim)
  m11.matrix <- matrix(0, nrow=beta.dim, ncol=beta.dim)
  m21.star.matrix <- matrix(0, nrow=alpha.dim, ncol=beta.dim)

  for (i in 1:n){
    Xi <- dataset[dataset[,’subj’]==i,,drop=F]
    subj.n <- nrow(Xi)
    subj.N<- Xi[,’N’, drop=F]
    subj.B <- Xi[,c(paste(’B’, 1:alpha.dim, sep=’’)), drop=F]
    subj.Z <- Xi[,c(’Z1’,’Z2’,’Z3’), drop=F]
    subj.mu <- exp(Xi[,c(paste(’B’, 1:alpha.dim, sep=’’),’Z1’,’Z2’,’Z3’)]%*%gama)

    subj.var.inv <- varfunc(subj.mu)$subj.var.inv
    m1 <- t(subj.Z)%*%diag(as.vector(subj.mu), nrow=subj.n)%*%subj.var.inv%*%
        (subj.N-subj.mu)
    m1.matrix[i,] <- m1
    m2.star <- t(subj.B)%*%subj.var.inv%*%(subj.N-subj.mu)*as.numeric(exp(
        subj.Z[1,]%*%gama[-(1:alpha.dim)]))
    m2.star.matrix[i,] <- m2.star

    m11 <- t(subj.Z)%*%diag(as.vector(subj.mu),nrow=subj.n)%*%subj.var.inv%*%
        diag(as.vector(subj.mu),nrow=subj.n)%*%subj.Z
    m11.matrix <- m11.matrix + m11
    m21.star.matrix <- m21.star.matrix + t(subj.B)%*%diag(as.vector(exp(subj.Z%*%
        gama[-(1:alpha.dim)])),nrow=subj.n)%*%subj.var.inv%*%diag(as.vector(subj.mu),
            nrow=subj.n)%*%subj.Z
    }
  alpha <- solve(crossprod(m2.star.matrix), crossprod(m2.star.matrix, m1.matrix))
  m2.matrix <- m2.star.matrix%*%alpha
  A.hat <- (m11.matrix-t(alpha) %*% m21.star.matrix)/n
  B.hat <- (t(m1.matrix-m2.matrix)%*%(m1.matrix-m2.matrix))/n
  A.hat.inv <- solve(A.hat)
  sandwich.ABA<- (A.hat.inv%*%B.hat%*%t(A.hat.inv))/n
  sandwich.B <- solve(B.hat)/n
  list(sandwich.ABA=diag(sandwich.ABA), sandwich.B=diag(sandwich.B),
    A=A.hat, B=B.hat)
  }
#======================== End of Function (sandwich) ======================#

#==========================================================================#
# Function name: sigma.est                                                 #
#--------------------------------------------------------------------------#
# Different methods of estimating the overdispersion parameter             #
#==========================================================================#

#--------------------------------------------------------------------------#
```

```
# sigma.est1: from Zeger (Biometrika 1988) & Davis et al. (Biometrika 2000)#
#-------------------------------------------------------------------------#
sigma.est1 <- function(N, mu, gama.dim=NA, K=NA){
      residual <- N - mu
      sigma2 <- sum(residual^2-mu)/sum(mu^2)
      sigma2 <- ifelse(sigma2>0, sigma2, 0)
      sigma2
      }

#-------------------------------------------------------------------------#
# sigma.est2: from Breslow(Appl. Statist. 1984); Breslow (JASA 1990)      #
#-------------------------------------------------------------------------#
sigma.est2 <- function(N, mu, gama.dim, K=NA){
  sigma2 <- 1
  d<-1
  while(d>1e-5){
    sigma2.update <- sum((N-mu)^2/(mu/sigma2+mu^2))/(length(N)- gama.dim)
    d <- abs(sigma2.update-sigma2)
    sigma2 <- sigma2.update
    }
  sigma2 <- ifelse(sigma2>0, sigma2, 0)
  sigma2
  }

#-------------------------------------------------------------------------#
# sigma.est3: MLE of overdispersion parameter based on --------------------#
# gamma poisson assumption                                                #
#-------------------------------------------------------------------------#
sigma.est3 <- function(N.vec, mu, gama.dim, K){
  n <- length(K)
  sigma.vec <- seq(0, 0.2, length=201)[-1]
  lmatrix <- matrix(nrow=n, ncol=200)
  for (i in 1:n){
    if (i==1) subj.start <- 1 else subj.start <- cumsum(K)[i-1]+1
    subj.end <- cumsum(K)[i]
    subj.mu <- mu[subj.start: subj.end]
    subj.dmu <- c(subj.mu[1], diff(subj.mu))
    subj.muk <- subj.mu[length(subj.mu)]
    subj.N <- N.vec[subj.start:subj.end]
    subj.dN <- c(subj.N[1], diff(subj.N))
    subj.Nk <- subj.N[length(subj.N)]
    l.est <- function(sigma2){
       sum(subj.dN*log(subj.dmu), na.rm=T)- (subj.Nk+1/sigma2)*log(subj.muk+
        1/sigma2)+1/sigma2*log(1/sigma2)+ lgamma(subj.Nk+1/sigma2)-lgamma(1/sigma2)
       }
    lmatrix[i,]<- sapply(sigma.vec, l.est)
    }
    lvec <- apply(lmatrix,2,sum)
  sigma2 <- sigma.vec[which.max(lvec)]
  sigma2
}
#======================== End of Function (sigma.est) ======================#

#===========================================================================#
# Function name: V1, V2, V3                                                 #
#---------------------------------------------------------------------------#
# Different Variance-covariance matrices                                    #
#===========================================================================#

#---------------------------------------------------------------------------#
# Covariance matrix V1, GEE with V1 is the score of pseudolikelihood        #
#---------------------------------------------------------------------------#
V1<-function(subj.mu){
  error<-0
  subj.n<-length(subj.mu)
  subj.var<-diag(as.vector(subj.mu), nrow=subj.n, ncol=subj.n)
  if (min(eigen(subj.var)$values)<1e-8 | max(eigen(subj.var)$values)>1e10){
    subj.var.inv <- NA
    error <- 1 }else
  subj.var.inv<-diag(as.vector(1/subj.mu), nrow=subj.n, ncol=subj.n)
  list(subj.var=subj.var, subj.var.inv=subj.var.inv, error=error)
  }

#---------------------------------------------------------------------------#
# Covariance matrix V2, GEE with V2 is the score of likelihood              #
#---------------------------------------------------------------------------#
V2<-function(subj.mu){
  error <- 0
  subj.n<-length(subj.mu)
  subj.var<-rep(subj.mu[1], subj.n)
  if (subj.n>1){
    for (i in 2:subj.n) subj.var <- cbind(subj.var, c(subj.mu[1:(i-1)],
        rep(subj.mu[i], subj.n-(i-1))))
      }
  if (min(eigen(subj.var)$values)<1e-8 | max(eigen(subj.var)$values)>1e8){
    subj.var.inv <- ginv(subj.var)
    error <- 1 }else
  subj.var.inv <- solve(subj.var)
  list(subj.var=subj.var, subj.var.inv=subj.var.inv, error=error)
  }

#---------------------------------------------------------------------------#
# Covariance matrix V3, GEE with V3 is the score of Gamma-frailty Poisson   #
# likelihood                                                                  #
#-----------------------------------------------------------------------------#
V3<-function(subj.mu, sigma2){
  error <- 0
  subj.n<-length(subj.mu)
  subj.var<-rep(subj.mu[1], subj.n)
  if (subj.n>1){
    for (i in 2:subj.n) subj.var <- cbind(subj.var, c(subj.mu[1:(i-1)],
        rep(subj.mu[i], subj.n-(i-1))))
      }
  subj.var <- sigma2*tcrossprod(subj.mu)+ subj.var
  if (min(eigen(subj.var)$values)<1e-8 | max(eigen(subj.var)$values)>1e8){
    subj.var.inv <- NA #ginv(subj.var)
    error <- 1 }else
  subj.var.inv <- solve(subj.var)
  list(subj.var=subj.var, subj.var.inv=subj.var.inv, error=error)
  }
#======================= End of Function (V1, V2, V3) ====================#

#=============================================================================#
# Function name: semiResult                                                   #
#-----------------------------------------------------------------------------#
# Run simulations in batch                                                    #
#=============================================================================#
semiResult<-function(K, T, Z, dN, method, varfunc, likelihood.func=NA, n.knot=1/3,
    position.knot='quantile'){
  N<-matrix(nrow=nrow(dN), ncol=ncol(dN))
  n<-length(K)
  for (i in 1:n) N[i,]<-cumsum(dN[i,])

  cov<-covariate(K, T, Z, N, method, n.knot=n.knot)
  dataset<-cov$X
  alpha.dim<-cov$alpha.dim

  beta.dim<-3
  gama.dim<-alpha.dim+beta.dim
  gama.ini<-c((1:alpha.dim)/alpha.dim, -1,0.5,1.5)
  result<-algorithm(dataset, alpha.dim, gama.ini=gama.ini, method='semiparametric',
    varfunc, likelihood.func=likelihood.func)
# estimates of the gama parameters
  gama <- result$gama
  alpha<-gama[1:(length(gama)-3)]
  beta.est <- gama[-(1:(length(gama)-3))]
# estimates of the variance of beta
  W1 <- result$W
  if (min(abs((eigen(W1)$values)))>1e-10){
    W1.inv <- solve(W1)
    W0 <- Varest(gama, alpha.dim, dataset, method=’semiparametric’, varfunc=varfunc)
    W <- W1.inv%*%W0%*% t(W1.inv)
    naive.v <- diag(W1.inv[-(1:(nrow(W1.inv)-3)),-(1:(ncol(W1.inv)-3))])
    v <- diag(W)[-(1:(nrow(W)-3))]
    } else {v <- NA; naive.v <- NA}
# estimates of the baseline hazard
  myt<-T[order(T)]
  myt<-myt[!is.na(myt)]
  qn<-ceiling(sum(K)^n.knot)
  if (position.knot==’quantile’) myknots <-  c(rep(min(myt),3),quantile(myt,
    seq(0,1,len=qn-4+2)),rep(max(myt),3)) else
  if (position.knot==’uniform’) myknots <- c(rep(min(myt),3), seq(min(myt),
    max(myt), length=qn-4+2), rep(max(myt),3))
  S<-splineDesign(knots=myknots, seq(2, 9, by=0.25))
  est<-S%*%alpha
  list(gama=gama, beta.est=beta.est, beta.variance=v, beta.variance.naive=
    naive.v, est=est, converge.status=result$converge.status, error=result$error,
    U=result$U)
  }
#======================== End of Function (semiResult) =====================#

#===========================================================================#
# Function name: bsvar                                                      #
#---------------------------------------------------------------------------#
# Generate boostrap samples to get bootstrap variances                      #
#===========================================================================#
bsvar<-function(K, T, Z, dN, n, varfunc, bs.n=200){
  bs.est<-matrix(nrow=bs.n, ncol=3)
  for (i in 1:bs.n){
    rep <- floor(runif(n, 1, n+1))
    replicate.data <- list(K=K[rep], T=T[rep,], Z=Z[rep,], dN=dN[rep,])
    replicate.result <- with(replicate.data, semiResult(K=K, T=T, Z=Z, dN=dN,
        method=method, varfunc=varfunc))
    bs.est[i,] <- replicate.result$beta.est
    }
  bs.var <- apply(bs.est, 2, var)
  bs.var
  }
#======================== End of Function (bsvar) ==========================#
```

REFERENCES

Andersen, P. K. & Gill, R. D. (1982), 'Cox's regression model for counting processes: A large sample study', *The Annals of Statistics* **10**(4), 1100–1120.

Best, M. & Chakravarti, N. (1990), 'Active set algorithms for isotonic regresion; a unifying framwork', *Mathematical Programming* **47**, 425–439.

Breslow, N. (1984), 'Extra-poisson variation in log-linear models', *Applied Statistics* **33**(1), 38–44.

Brunk, H., Ewing, G. & Utz, W. (1957), 'Minimizing intergrals in certain classes of monotone functions', *Pacific Journal of Mathematics* **7**, 833–847.

Byar, D., Blackard, C. & Vacurg (1980), 'Comparisons of placebo, pyridoxine, and topical thiotepa in preventing stage i bladder cancer', *Urology* **10**, 556–561.

Chan, K. & Ledolter, J. (1995), 'Monte carlo em estimation for time series models involving counts', *Journal of American Statistical Association* **90**, 242–52.

Cox, D. (1972), 'Regression models and life-table (with discussion)', *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

Davis, R., Dunsmuir, W. & Wang, Y. (2000), 'On autocorrelation in a poisson regression model', *Biometrika* **87**(3), 491–505. They discuss a time series with latent process, as discussed in Zeger (1988). He pointed Zeger's method of mement in underestimates the overdispersion parameter and provides a correction.

de Boor, C. (2001), *A Practical Guide to Splines*, Springer-Verlag.

Diggle, P., Liang, K.-Y. & Zegger, S. (1994), *Analysis of longitudinal data*, Oxford Press.

Fan, J. & Li, R. (2004), 'New estimation and model selection procedures for semiparameric modeling in longitudinal data analysis', *Journal of the American Statistical Association* **99**(467), 710–723.

Gentleman, R., Lawless, J., Lindsey, J. & Yan, P. (1994), 'Multi-state markov models for analysing incomplete disease history data with illustrations for hiv disease', *Statistics in Medicine* **13**(8), 805–821.

Gladman, D., VT, F. & C., N. (1995), 'Clinical indicators of progression in psoriatic arthritis: multivariate relative risk model', *Journal of Rheumatology* **22**(4), 675–679.

Groeneboom, P. & Wellner, J. (1992), *Information Bounds and Nonparametric Maximum Likelihood Estimation*, Basel: Birkhauser.

Hay, J. & Pettitt, A. (2001), ‘Bayesian analysis of a time series of counts with covariates: an application to the control of an infectious disease’, *Biostatistics* **2**, 433–444.

Hoover, D. R., Rice, J. A., Wu, C. O. & Yang, L.-P. (1998), ‘Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data’, *Biometrika* **85**(4), 809–822.

Huang, J. (1996), ‘Efficient estimation for the proportional hazards model with interval censoringq’, *The Annals of Statistics* **24**(2), 540–568.

Jamshidian, M. (2004), ‘On algorithms for restricted maximum likelihood estimation’, *Computational Statistics and Data Analysis* **45**, 137–157.

Jongbloed, G. (1998), ‘The iterative convex minorant algorithm for nonparametric estimation’, *Journal of Computational and Graphical Statistics* **7**(3), 310–321.

Kalbfleisch, J. D. & Lawless, J. F. (1985), ‘The analysis of panel data under a markov assumption’, *Journal of the American Statistical Association* **80**(392), 863–871.

Lawless, J. F. (1987), ‘Regression methods for poisson process data’, *Journal of the American Statistical Association* **82**(399), 808–815.

Lawless, J. F. & Nadeau, J. (1995), ‘Some simple robust methods for the analysis of recurrent events’, *Thechnometrics* **37**, 158–168.

Lee, E. W. & Kim, M. Y. (1998), ‘The analysis of correlated panel data using a continuous-time markov model’, *Biometrics* **54**(4), 1638–1644.

Liang, K. & Zeger, S. L. (1986), ‘Longitudinal data analysis using generalized linear models’, *Biometrika* **97**(1), 13–22.

Lin, D., Wei, L., Yang, I. & Ying, Z. (2000), ‘Semiparametric regression for the mean and rate functions of recurrent events’, *J. R. Statist. Soc. B* **62**, 711–730.

Lin, D. & Ying, Z. (2001), ‘Semiparametric and nonparametric regression analysis of longitudinal data’, *Journal of the American Statistical Association* **96**(453), 103–113.

Lin, X. & Carroll, R. (2000), ‘Nonparametric function estimation for clustered data when the predictor is measured without/with error’, *Journal of the American Statistical Association* **95**(450), 520–534.

Lin, X. & Carroll, R. (2001), ‘Semiparametric regression for clustered data using generalized estimating equations’, *Journal of the American Statistical Association* **96**(455), 1045–1056.

Lin, X. & Zhang, D. (1999), ‘Inference in generalized aditive mixed models by using smoothing splines’, *J. R. Statist. Soc. B* **61**(Part 2), 381–400.

Lu, M., Zhang, Y. & Huang, J. (2007), 'Estimation of the mean function with panel count data using monotone polynomial splines', *Biometrika* **94**, 705–718.

Lu, M., Zhang, Y. & Huang, J. (2009), 'Semiparametric estimation methods for panel count data using monotone polynomial splines', *Journal of the American Statistical Association* **27**, 1–11.

Pan, W. (1999), 'Extending teh iterative convex minorant algorithm to the cox model for interval-censored data', *Journal of Computational and Graphical Statistics* **8**(1), 109–120.

Pepe, M. & Cai, J. (1993), 'Some graphical displays and marginal regression and analyses for recurrent failure times and time dependent covariates', *Journal of the American Statistical Association* **88**, 811–820.

Prentice, R., Williams, B. & Peterson, A. (1981), 'On the regression analysis of multivariate failure time data', *Biometrika* **62**(2), 373–379.

Rice, J. A. & Wu, C. O. (2001), 'Nonparametric mixed effects models for unequally sampled noisy curves', *Biometrics* **57**, 253–259. Mixed effect, spline, eigenfunction.

Robertson, T., Write, F. & Dykstra, R. (1988), *Order Restricted Statistical Inference*, New York: Wiley.

Rosen, J. (1960), 'The gradient projection method for nonlinear progarmming. part i linear constraints', *Journal of the Society for Industrial and Applied Mathematics* **8**(1), 181–217.

Rudemo, M. (1982), 'Empirical choice of histograms and kernel density estimators', *Scandinavian Journal of Statistics* **9**(2), 65–78.

Schumaker, L. (1981), *Spline Functions: Basic Theory*, New York: Wiley.

Severini, T. & Staniswalis, J. (1994), 'Quasilikelihood estimation in semiparametric models', *Journal of the American Statistical Association* **89**(426), 501–511.

Shen, X. & Wong, W. (1994), 'Convergence rate of sieve estimates', *The Annals of Statistics* **22**(2), 580–615.

Silverman, B. (1986), *Density estimation for statistics and data analysis*, Chapman and Hall, London.

Sun, J. & Fang, H. (2003), 'A nonparametric test for panel count data', *Biometrika* **90**(1), 199–208.

Sun, J. & Kalbfleisch, J. (1995), 'Estimation of the mean function of point processes based on panel count data', *Statistica Sinica* **5**, 279–290.

Sun, J. & Wei, L, J. (2000), 'Regression analysis of panel count data with covariate-dependent observation and censoring times', *J. R. Statist. Soc. B* **62**(2), 293–302.

Terrell, G. R. (1990), 'The maximal smoothing principle in density estimation', *Journal of the American Statistical Association* **85**(410), 470–477.

Thall, P. F. (1988), 'Mixed poisson likelihood regression models for longitudinal interval count data', *Biometrics* **44**(1), 197–209. 0006341X.

Thall, P. F. & Lachin, J. (1988), 'Analysis of recurrent events: Nonparametric methods for random-interval count data', *Journal of the American Statistical Association* **83**(402), 339–347.

van der Vaart, A. (1998), *Asymptotic Statistics*, Cambridge University Press.

van der Vaart, A. & Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, Springer.

Wang, N. (2003), 'Marginal nonparametric kernel regression accounting for within-subject correlation', *Biometrika* **90**(1), 43–52.

Wang, N., Carroll, R. & Lin, X. (2005), 'Efficient semiparametric marginal estimation for longitudinal/clustered data', *Journal of the American Statistical Association* **100**, 147–157.

Wei, L., Lin, D. & Weissfeld, L. (1989), 'Regression analysis of multivariate incomplete failure time data by modeling marginal distribution', *Journal of the American Statistical Association* **1989**(84).

Wellner, J. A. & Zhang, Y. (1995), 'Two likelihood-based semiparametric estimation methods for panel count data with covariates'.

Wellner, J. A. & Zhang, Y. (2000), 'Two estimators of the mean of a counting process with anel count data', *The Annals of Statistics* **28**(3), 779–814.

Wellner, J. A. & Zhang, Y. (2007), 'Two likelihood-based semiparametric estimation methods for panel count data with covariates', *The Annals of Statistics* **35**(5), 2106–2142.

Wild, C. & Yee, T. (1996), 'Additive extensions to generalized estimating equation methods', *J. R. Statist. Soc. B* **58**(4), 711–725.

Wu, H. & Zhang, J. (2002), 'Local polynomial mixed-effects models for longitudinal data', *Journal of the American Statistical Association* **97**, 883–897.

Zeger, S. L. (1988), 'A regression model for time series of counts', *Biometrika* **75**(4), 621–629.

Zeger, S. L. & Diggle, P. (1994), 'Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters', *Biometrics* **50**.

Zeger, S. L. & Liang, K. (1986), 'Longitudinal data analysis for discrete and continuous outcomes', *Biometrics* **42**(1), 121–130.

Zhang, D., Lin, X., Raz, J. & Sours, M. (1998), 'Semiparametric stochastic mixed models for longitudinal data', *Journal of the American Statistical Association* **93**(442), 710–719.

Zhang, H. (1997), 'Multivariate adaptive splines for analysis of longitudinal data', *Journal of Computational and Graphical Statistics* **6**(1), 74–91.

Zhang, Y. (2002), 'A semiparametric pseudo likelihood estimation method for panel count data', *Biometrika* **89**, 39–48.

Zhang, Y. (2006), 'Nonparametric k-sample tests with panel count data', *Biometrika* **93**(4), 777–790.

Zhang, Y. & Jamshidian, M. (2004), 'On algorithms for the nonparameric maximum likelihood estimator of the failure function with censored data', *Journal of Computational and Graphical Statistics* **3**(1), 123–140.

Zhu, Z., Fung, W. K. & He, X. (2008), 'On the asymptotics of marginal regression splines with longitudinal data', *Biometrika* **95**(4), 907–917.