Spline-Based Semiparametric Projected Generalized Estimating Equation Method for Panel Count Data

Lei Hua

FXB 514, Center for Biostatistics in AIDS Research/HSPH

651 Huntington Avenue, Boston, MA 02466, U.S.A.

lhua@sdac.harvard.edu

Ying Zhang

Department of Biostatistics, The University of Iowa

C22 GH, 200 Hawkins Drive, Iowa City, IA 52242, U.S.A.

ying-j-zhang@uiowa.edu

Summary

We propose to analyze panel count data using a spline-based semiparametric projected generalized estimating equation method with the semiparametric proportional mean model $E(\mathbb{N}(t)|Z) = \Lambda_0\,(t)\,e^{\beta_0^T Z}$. The natural logarithm of the baseline mean function, $\log \Lambda_0\,(t)$, is approximated by monotone cubic B-spline functions. The estimates of regression parameters and spline coefficients are obtained by projecting the generalized estimating equation estimates into the feasible domain using a weighted isotonic regression. The proposed method avoids assuming any parametric structure of the baseline mean function or the underlying counting process. Selection of the working-covariance matrix that represents the true correlation between the cumulative counts improves the estimating efficiency. Simulation studies are conducted to investigate finite sample performance of the proposed method and to compare the estimating efficiency using different working-covariance matrices in the generalized estimating equation. Finally, the proposed method is applied to a real dataset from a bladder tumor clinical trial.

Some key words: Semiparametric model; Generalized estimating equation; Monotone polynomial splines; Counting process; Over-dispersion;

1. Introduction

Panel count data are often seen in clinical trials, industrial reliability and epidemiologic studies. A well-known example is the bladder tumor randomized clinical trial 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 tumor were randomized into one of three treatment groups: placebo, pyridocine pills or thiotepa instillation. At subsequent follow-up visits, the number of newly recurrent tumors was counted, the new tumors were removed and the treatment was continued. The number of follow-up visits and the visit times may vary from subject to subject. The goal of this study was to determine the effects of different treatments on suppressing recurrence of the bladder tumor.

There are increasing interests in methodological research for panel count data in recent statistical literature. Various approaches were explored by, for example, Lee & Kim (1998), Thall (1988), Sun & Kalbfleisch (1995) and Wellner & Zhang (2000). Particularly, semiparametric regression analysis for panel count data with the proportional mean model, namely,

$$E(\mathbb{N}(t)|Z) = \Lambda_0\,(t)\,e^{\beta_0^T Z} \tag{1}$$

where $\Lambda_0(t)$ is the nondecreasing baseline mean function and $\beta_0 \in R^d$ is the d-dimension regression parameter, has drawn considerable attention among researchers in this field. Sun & Wei (2000) and Sun et al. (2005) studied estimating equation methods for making inference about the regression parameter β_0. But the validity of their methods relies on some assumptions of observation times which may be hard to justify in applications. Wellner & Zhang (2007) studied the semiparametric maximum pseudo-likelihood estimator and the semiparametric maximum likelihood estimator assuming the underlying counting process as a nonhomogeneous Poisson process with mean given by (1). Wellner & Zhang (2007) proved consistency and derived convergence rate of both estimators. They showed that the two estimators are robust to the underlying nonhomogeneous Poisson assumption. The maximum pseudo-likelihood estimator can be easily calculated but it can be very inefficient especially when the observation times are heavily tailed as discussed in Wellner et al. (2004). The maximum likelihood estimator is more efficient but it requires a doubly iterative algorithm which needs a large number of iterations to converge. Lu et al. (2009) studied the spline-based sieve version of the semiparametric maximum pseudo-likelihood estimator and the semiparametric maximum likelihood estimator of Wellner & Zhang (2007) by approximating the baseline mean function using monotone B-spline functions (Schumaker, 1981). Not only did they demonstrate a great numerical advantage in the sieve likelihood methods, they also showed good asymptotic behavior of their estimators. Moreover, the sieve estimators of the baseline mean function can have a better convergence rate than their counterparts studied by Wellner & Zhang (2007).

The nonhomogeneous Poisson process model assumes the variance of the cumulative counts equals to the expected number of the counts, that is, no over-dispersion is accounted. Aforementioned likelihood-based methods, though leading to a consistent estimation, do not take into account the possible over-dispersion problem that often occurs in various applications of longitudinal count data. Although the maximum likelihood estimator of the regression parameter is robust and semiparametrically efficient as shown by both Wellner & Zhang (2007) and Lu et al. (2009) when the Poisson process model is true, it may not be the best estimator when the Poisson model assumption for the underlying counting process is violated.

In this manuscript, we consider a spline-based semiparametric regression method motivated by generalized estimating equation approach (GEE). Instead of assuming the underlying nonhomogeneous Poisson process, we only assume the proportional mean model and conjecture the covariance matrix that accounts for the over-dispersion. We will demonstrate that the proposed method improves the estimating efficiency when either over-dispersion or autocorrelation is present in the data.

The rest of the paper is organized as follows: Section 2 introduces the spline-based semiparametric projected GEE method. Three working-covariance matrices are discussed to accommodate different data structures. Section 3 proposes an easy-to-implement algorithm to compute the projected GEE estimate. Section 4 provides numerical results including simulation studies and an application to the bladder tumor example; Finally, we give some concluding remarks in section 5. Some technical results are given in Appendix.

2. Spline-based Semiparametric Projected GEE method

Suppose, $\mathbb{N} = \{\mathbb{N}(t) : t \geq 0\}$ is a univariate counting process. There are K random 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 counts 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 $\mathbb{N}$, conditional 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 $\mathbb{N}^{(i)} = \left(\mathbb{N}\left(T^{(i)}_{K_i,1}\right), \mathbb{N}\left(T^{(i)}_{K_i,2}\right), \cdots, \mathbb{N}\left(T^{(i)}_{K_i,K_i}\right)\right)$.

In this article, we consider to use monotone cubic B-spline functions to approximate the logarithm of the baseline mean function, $\log\Lambda_0(t)$. Suppose the observation times are restricted in a closed interval $[L, U]$. 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 $[L, U]$ into $m_n + 1$ subintervals, where $m_n \approx n^{\nu}$ is a positive integer such that $\max_{1\leq k\leq m_n} |t_{l+k} - t_{l+k-1}| = O(n^{-\nu})$. Denote

$\phi_{l,t}$ a class of polynomial spline functions of order l, $l \geq 1$. $\phi_{l,t}$ is spanned by a series of B-spline basis functions $\{B_i, 1 \leq i \leq q_n\}$ where $q_n = m_n + l$. 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 B-splines according to the variation diminishing property of B-splines (Schumaker, 1981) and hence is a proper feasible class from which the estimate of $\log \Lambda_0\left(t\right)$ can be found.

The Generalized Estimating Equation (GEE) method, developed by Liang & Zeger (1986) is widely used in parametric regression analysis of longitudinal data. It provides a robust inference with only weak assumptions of the underlying distributions. A large amount of literatures generalized the same idea to semiparametric regression analysis with the mean response model given by

$$E\left(Y|Z\right) = \mu\{\phi_0\left(T\right) + \beta_0^T Z\}. \tag{2}$$

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. The proportional mean model of (1) is a special case of (2) with the exponential link function μ and ϕ_0 being the logarithm of the baseline mean function. We approximate the proportional mean function by $\exp\{\sum_{l=1}^{q_n} \alpha_l B_l\left(t\right) + \beta_0^T Z\}$. The GEE for computing $\theta = (\beta, \alpha)$ is given by

$$U\left(\theta\right) = \sum_{i=1}^{n} \left(\frac{\partial \mu^{(i)}(\theta)}{\partial \theta}\right) V^{(i)^{-1}}(\theta) \left(\mathbb{N}\left(T_i\right) - \mu^{(i)}(\theta)\right) = 0 \tag{3}$$

where $\mu^{(i)}(\theta) = \left(\mu^{(i)}_{K_i,1}(\theta), \mu^{(i)}_{K_i,2}(\theta), \cdots \mu^{(i)}_{K_i,K_i}(\theta)\right)^T$ with $\mu^{(i)}_{K_i,j}(\theta) = \exp\left(\sum_{l=1}^{q_n} \alpha_l B_l\left(T^{(i)}_{K_i,j}\right) + \beta^T Z_i\right)$ for $j = 1, 2, \cdots K_i$. However the solution of (3) does not necessarily provide an $\alpha = (\alpha_1, \cdots, \alpha_{q_n})$ that satisfies the monotone constraints. In order to produce a monotone non-decreasing estimate of α, we propose to project the GEE solution $\tilde{\alpha}_n = (\tilde{\alpha}_{n,1}, \cdots, \tilde{\alpha}_{n,q_n})$ from (3) into the feasible space $\Pi = \{\alpha : \alpha_1 \le \alpha_2 \le \cdots \le \alpha_{q_n}\}$ by a quadratic programming:

$$\hat{\alpha}_n = \mathrm{Proj}_W\left[\tilde{\alpha}_n, \Pi\right] = arg \min_{\alpha \in \Pi} (\alpha - \tilde{\alpha}_n)' W (\alpha - \tilde{\alpha}_n), \tag{4}$$

where W is a positive definite matrix. The spline-based semiparametric projected GEE estimator of Λ_0 is taken to be $\hat{\Lambda}(t) = \exp\left(\sum_{l=1}^{q_n} \hat{\alpha}_{n,l} B_l(t)\right)$ after the estimate of the spline coefficient $\hat{\alpha}_n = \{\hat{\alpha}_{n,l}, l = 1, 2, \cdots, q_n\}$ is obtained.

For the GEE method, $V^{(i)}$ is the working-covariance matrix for the panel counts from the i^{th} process and plays a pivotal role in determining the estimating efficiency. Different choices of this covariance matrix could accommodate the characteristics of different counting processes. The easiest choice of the covariance matrix is to use a diagonal matrix, in which the diagonal elements are determined by the variance function of a Poisson distribution, i.e.

$Var\left(\mathbb{N}\left(T_{K_i,j}\right)\right) = E\left(\mathbb{N}\left(T_{K_i,j}\right)\right)$ and

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

Using the diagonal matrix implies independence between cumulative counts, despite the cumulative counts are obviously positively correlated. The spline-based semiparametric GEE with this covariance matrix is exactly the score equation of the pseudo-likelihood studied by Lu et al. (2009) and the proof is given in Appendix 6.1. Instead of using the diagonal matrix that ignores the correlation among the cumulative counts, a working-covariance matrix that accommodates such correlation will intuitively produce more efficient estimate. The covariance function based on the Poisson counting process $Cov\left(\mathbb{N}\left(t_1\right), \mathbb{N}\left(t_2\right)\right) = E\left(\mathbb{N}\left(t_1\right)\right), \quad$ for $t_1 \leq t_2$ leads to the selection of the working-covariance matrix $V_2^{(i)}$ in the form of

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

Through simple but tedious algebra, the spline-based semiparametric GEE with this covariance matrix is exactly the score equation of the likelihood based on the nonhomogeneous

Poisson process model by Lu et al. (2009) and the proof is given in Appendix 6.2. Despite the improved estimation efficiency using $V_2^{(i)}$ compared to the one using $V_1^{(i)}$, it still imposes unrealistic assumptions to the covariance structure of the data: First, it assumes the variance of the counts equal to the mean, that is, no over-dispersion is accounted for the data; Second, it assumes independence of the counts between non-overlapping intervals. When either of these assumptions is violated, the estimate based on $V_2^{(i)}$ may not be very efficient.

In the literature of count data, 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 choice in parametric regression analysis to account for possible over-dispersion. Chan & Ledolter (1995) and Hay & Pettitt (2001) discussed a log normal frailty model by assuming a lognormal distribution of the frailty term γ. But there is no close form for the marginal distribution of count and the estimation with this frailty variable is computationally intensive. 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 cumulative count. Zhang & Jamshidian (2003) introduced a gamma frailty term to nonparametric estimation of the mean function of counting process. Zeger (1988) considered a latent frailty process while assuming only the first and second moments of the frailty term. We adopt a similar idea in our semiparametric GEE setting. We specify γ with $E\left(\gamma\right) = 1$, which guarantees the identifiability of the model and does not violate the proportional mean model specified in (1). Denote $Var\left(\gamma\right) = \sigma^2$, the marginal variance function based on the 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 accounted for by the frailty parameter γ as well, 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 of the form

$$V_3^{(i)} = \begin{pmatrix} \mu_{K_i,1}^{(i)} + \sigma^2 \mu_{K_i,1}^{(i)} \mu_{K_i,1}^{(i)} & \mu_{K_i,1}^{(i)} + \sigma^2 \mu_{K_i,1}^{(i)} \mu_{K_i,2}^{(i)} & \cdots & \mu_{K_i,1}^{(i)} + \sigma^2 \mu_{K_i,1}^{(i)} \mu_{K_i,K_i}^{(i)} \\ \mu_{K_i,1}^{(i)} + \sigma^2 \mu_{K_i,1}^{(i)} \mu_{K_i,2}^{(i)} & \mu_{K_i,2}^{(i)} + \sigma^2 \mu_{K_i,2}^{(i)} \mu_{K_i,2}^{(i)} & \cdots & \mu_{K_i,2}^{(i)} + \sigma^2 \mu_{K_i,2}^{(i)} \mu_{K_i,K_i}^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{K_i,1}^{(i)} + \sigma^2 \mu_{K_i,1}^{(i)} \mu_{K_i,K_i}^{(i)} & \mu_{K_i,2}^{(i)} + \sigma^2 \mu_{K_i,2}^{(i)} \mu_{K_i,K_i}^{(i)} & \cdots & \mu_{K_i,K_i}^{(i)} + \sigma^2 \mu_{K_i,K_i}^{(i)} \mu_{K_i,K_i}^{(i)} \end{pmatrix}_{K_i \times K_i}$$

and it can be rewritten as

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

$V_2^{(i)}$ is, therefore, a special case of $V_3^{(i)}$ with $\sigma^2 = 0$.

The estimating equation with $V_3^{(i)}$ turns out to be the score equation of the marginal likelihood of panel count data under the Gamma-Frailty nonhomogeneous Poisson model, that is, given the gamma distribution of the frailty term, $\gamma \sim \Gamma\left(1/\sigma^2, 1/\sigma^2\right)$, the cumulative count follows a nonhomogeneous Poisson process with mean $\gamma \Lambda\left(t\right) e^{\beta^T Z}$. The proof is given in Appendix 6.3. Because of this property and $V_2^{(i)}$ is a special case of $V_3^{(i)}$, it is likely that the spline-based semiparametric projected GEE method with $V_3^{(i)}$ will lead to a more efficient estimate than the spline-based maximum likelihood estimate studied by Lu et al. (2009), when the over-dispersion exists.

3. Numerical Algorithm

For computing the proposed projected GEE estimate of $\theta = (\beta, \Lambda)$, estimation of the over-dispersion parameter σ^2 is needed. It is possible to create an extra estimating equation using the second moment to jointly solve for $(\beta, \alpha, \sigma^2)$. It is, however, numerically cumbersome. We propose to estimate σ^2 externally to the GEE.

Breslow (1984) used a method of moment to estimate the over-dispersion parameter σ^2 by solving.

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

where $\mathbb{N}_{ij} = \mathbb{N}\left(T_{ij}\right)$, $\hat{\mu}_{ij}$ is any consistent estimate of $E\left(\mathbb{N}_{ij}\right)$, 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

$$\hat{\sigma}_n^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{K_1} \frac{\left(\mathbb{N}_{ij} - \hat{\mu}_{ij}\right)^2}{\hat{\mu}_{ij}\left(\hat{\mu}_{ij} + \hat{\sigma}_n^{-2}\right)}}{\sum_{i=1}^{n} K_i - p}$$

Alternatively, σ^2 could also be estimated explicitly by

$$\hat{\sigma}_n^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{K_i} \left\{\left(\mathbb{N}_{ij} - \hat{\mu}_{ij}\right)^2 - \hat{\mu}_{ij}\right\}}{\sum_{i=1}^{n}\sum_{j=1}^{K_1} \hat{\mu}_{ij}^2} \tag{5}$$

as proposed by Zeger (1988). Both Zeger's and Breslow's methods could underestimate the over-dispersion parameter and even end up with a negative $\hat{\sigma}_n^2$. If that happens, $\hat{\sigma}_n^2$ is forced to be zero. In our spline-based semiparametric projected GEE method, this over-dispersion parameter is a nuisance parameter and has little impact on the consistency of the estimate of (β, Λ). Hence, for the sake of numerical simplicity, Zeger's method is adopted in our calculation.

A two-stage estimating procedure is implemented when $V_3^{(i)}$ is chosen as the working-covariance matrix. At the first stage, due to its computational convenience, the spline-based semiparametric projected GEE method with $V_1^{(i)}$, or equivalently the spline-based semiparametric maximum pseudo-likelihood estimate studied by Lu et al. (2009) is implemented to get an initial consistent estimate of $\theta = (\beta, \alpha)$, $\theta^{(0)} = (\beta^{(0)}, \alpha^{(0)})$. Then an estimate of σ^2, $\hat{\sigma}_n^2$ is obtained using Zeger's method (5) in which $\hat{\mu}_{ij} = \mu_{K_i,j}^{(i)}(\theta^{(0)})$. At the second stage, replacing σ^2 by the estimate, $\hat{\sigma}_n^2$, the estimate of $\theta = (\beta, \alpha)$ is obtained by projecting the GEE update of

$$U\left(\theta; \hat{\sigma}_n^2\right) = \sum_{i=1}^{n} \left(\frac{\partial \mu^{(i)}(\theta)}{\partial \theta}\right) V_3^{(i)^{-1}}(\theta; \hat{\sigma}_n^2) \left(\theta; \hat{\sigma}_n^2\right) \left(\mathbb{N}\left(T_i\right) - \mu^{(i)}(\theta)\right) = 0$$

into the feasible space $\Theta = R^d \times \Pi$.

A hybrid algorithm of Newton-Raphson type method and Isotonic Regression (NR/IR) is used to compute the spline-based projected GEE estimate. Newton-Raphson (NR) algorithm is a widely used iterative algorithm for finding the optimizer of nonlinear equations as it is known to have a quadratical convergence rate. However it cannot guarantee the monotonicity of the iterates. So after each NR iteration, the projection step is made by an easy-to-implement isotonic regression. At the current estimate $\theta^{(k)} = (\beta^{(k)}, \alpha^{(k)})$, denote

$$\begin{aligned} H\left(\theta^{(k)}; \hat{\sigma}_n^2\right) &= -E\left\{\nabla_\theta U(\theta^{(k)}; \hat{\sigma}_n^2)\right\} \\ &= \sum_{i=1}^{n} \left(\frac{\partial \mu^{(i)}(\theta^{(k)})}{\partial \theta}\right) V_3^{(i)^{-1}}(\theta; \hat{\sigma}_n^2) \left(\frac{\partial \mu^{(i)}(\theta^{(k)})}{\partial \theta}\right)^T \\ &= \begin{pmatrix} H_{\beta\beta}(\theta^{(k)}; \hat{\sigma}_n^2) & H_{\beta\alpha}(\theta^{(k)}; \hat{\sigma}_n^2) \\ H_{\alpha\beta}(\theta^{(k)}; \hat{\sigma}_n^2) & H_{\alpha\alpha}(\theta^{(k)}; \hat{\sigma}_n^2) \end{pmatrix}, \end{aligned}$$

the negative expectation of the derivative of estimating function which is the Fisher information if the underlying stochastic model is indeed Gamma-frailty Poisson model. We choose

$$W = \mathrm{diag}(w_1, w_2, \cdots, w_{q_n}) = \mathrm{diag}\left(H_{\alpha\alpha}(\theta^{(k)}; \hat{\sigma}_n^2)\right)$$

for the weight matrix in (4). The projection is actually the weighted isotonic regression problem and the solution has a nice interpretation: it is the left derivative of the greatest minorant of the cumulative sum diagram $\{P_i, i = 0, 1, \cdots, n\}$ (Groeneboom & Wellner, 1992) where

$$P_0 = (0, 0) \text{ and } P_i = \left(\sum_{l=1}^{i} w_l, \sum_{l=1}^{i} w_l \alpha_l^{(k)}\right);$$

and can be expressed as

$$\hat{\alpha}_i = \max_{j<i} \min_{l>i} \frac{\sum_{m=j}^{l} w_m \alpha_m^{(k)}}{\sum_{m=j}^{l} w_m}$$

The NR/IR algorithm tailored to the spline-based projected GEE estimation is summarized in the following steps.

Step 0 (Initial Values): Obtain an initial estimate $\theta^{(0)} = \left(\alpha^{(0)}, \beta^{(0)}\right)$ by the projected GEE with the working-covariance matrix $V_1^{(i)}$ and obtain an estimate of σ^2, $\hat{\sigma}_n^2$ with $\theta^{(0)} = (\beta^{(0)}, \alpha^{(0)})$ using the Zeger's method (5). Iterate the algorithm through the following steps until convergence.

Step 1 (Newton-Raphson Type Update): Update the current estimate $\theta^{(k)} = \left(\beta^{(k)}, \alpha^{(k)}\right)$ by Newton-Raphson type algorithm,

$$\tilde{\theta}^{(k+1)} = \left(\tilde{\beta}^{(k+1)}, \tilde{\alpha}^{(k+1)}\right) = \theta^{(k)} + H^{-1}\left(\theta^{(k)}; \hat{\sigma}_n^2\right) U\left(\theta^{(k)}; \hat{\sigma}_n^2\right).$$

Step 2 (Isotonic Regression): Construct the cumulative sum diagram $\{P_i, i = 0, 1, \cdots, n\}$ where

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

where $w_l, l = 1, 2, \cdots, q_n$ are the diagonal elements of $H_{\alpha\alpha}\left(\theta^{(k)}; \hat{\sigma}_n\right)$.The update of α is obtained by the left derivative of the convex minorant of this cumulative sum diagram, that is,

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

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

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

4. Numerical Results

4.1 *Simulation Studies*

Simulation studies are conducted to examine the performance of the spline-based semiparametric projected 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: (i) The simulation of observation times mimics a possible scenario in clinical follow-up study in which the chance of skipping the follow-up visit may increase as the study goes along. 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 with $P(\delta_{ij} = 1) = \frac{1}{1+e^{T_{ij}^\circ - 10}}$. 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 $\{T_{ij}^\circ : \xi_{ij}\delta_{ij} = 1, j = 1, \cdots, 6\}$; (ii) The covariate vector $Z_i = (Z_{i1}, Z_{i2}, Z_{i3})$ is simulated by $Z_{i1} \sim \text{Uniform}\,(0,1)\,, Z_{i2} \sim N\,(0,1)\,,\text{and } Z_{i3} \sim \text{Bernoulli}\,(0.5)$; (iii) Set the regression parameter $\beta_0 = (\beta_{0,1}, \beta_{0,2}, \beta_{0,3})^T = (-1.0, 0.5, 1.5)^T$ and given $\left(Z_i, K_i, \underline{T}_{K_i}\right)$, four 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. Data are generated from a Gamma-Frailty Poisson process. The frailty variables $\gamma_1, \gamma_2, \cdots, \gamma_n$ are a random sample from Gamma distribution, $\Gamma\,(0.5, 0.5)$ that results in the over-dispersion parameter equal to 2. Conditioning on the frailty variable γ_i as well as the covariates Z_i, the panel counts for each subject are drawn from a Poisson process, i.e.

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

for $j = 1, 2, \cdots, K_i$. 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 (1) and $E\left(\mathbb{N}\left(t\right)|Z\right) = 2t^{1/2}e^{\beta_0^T Z_i}$. The counts are marginally negative binomial distributed.

Scenario 2. Data are generated similarly to *Scenario 1.* Instead of generating the frailty variable 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 so called mixed Poisson process as studied in Wellner & Zhang (2007) and Lu et al. (2009). 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 (1) still holds.

Scenario 3. Data are generated from a Poisson process with the conditional mean function given by $2t_{ij}^{1/2}e^{\beta_0^T Z_i}$, 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 4. Data are generated from a 'Negative-binomialized' counting process. Conditioning on covariate Z, a random variable M is generated from a Negative binomial distribution, NegBin$(20e^{\beta_0^T Z}, 0.1)$. Given M, a random sample, $X_i, i = 1, 2, \cdots, M$, is generated from distribution function F_x. The count data is defined by

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

F_x is chosen to be $t^{1/2}/90 \cdot I_{[t \leq 8100]} + I_{[t > 8100]}$ such that the proportional mean model in (1) still holds and the baseline mean function $\Lambda_0(t) = 2t^{1/2}$, is the same as those in scenarios 1, 2 and 3 for $t \leq 8100$. With the current formulation of the problem, the data show both over-dispersion and autocorrelation between non-overlapping increments. The covariance matrix has a similar form as matrix V_3, but the true over-dispersion parameter depends on covariates.

For all these scenarios, the monotone cubic B-splines are used to approximate the baseline mean function in the proposed semiparametric GEE method. 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.

For the inference of the regression parameter, we propose an "ad-hoc" approach to estimate the standard error of the estimated regression parameter. We pretend the proposed spline-based semiparametric projected GEE estimate as the ordinary parametric GEE estimate and obtain the standard error as the square-root of asymptotic variance based on the well-known sandwich formula for the ordinary GEE estimate given by Liang & Zeger (1986). Alternatively, the bootstrap method could be implemented for estimating the standard error since the spline approach largely reduces computing time in the estimation.

In our simulation studies, 1000 Monte Carlo samples are generated with sample size of 50 and 100 for each scenario and the results on estimation bias (bias), Monte Carlo standard deviation (M-C sd), average of the estimated standard errors based on either the parametric GEE sandwich formula (SSE) or bootstrap method (BSE), and their 95% coverage probabilities (CP1 with SSE and CP2 with BSE) for the regression parameters are summarized in Tables 1-2 corresponding to the four simulation scenarios with 2 different sample sizes.

When data follow the Gamma-Frailty Poisson process as in *Scenario 1*, all three estimates with the different working-covariance matrices are consistent. The biases are negligible compared to the standard errors. The estimate with the working-covariance matrix $V_3^{(i)}$ apparently outperforms its alternative estimate with the working-covariance matrix $V_1^{(i)}$ or $V_2^{(i)}$ in view of the smaller standard errors. This is expected as the working-covariance matrix $V_3^{(i)}$ correctly specifies the underlying correlations among the cumulative panel counts. The parametric sandwich estimate of the standard error of the estimated regression parameter appears to underestimate the true standard error as compared to the Monte-Carlo standard deviation, which attributes to a lower coverage than the nominal level. The underestimation lessens as sample size increases. Among the three standard error estimates, it seems that the estimate of the proposed GEE method with $V_3^{(i)}$ has the least bias. The bootstrap method provides a less biased estimate of the standard error, especially when the over-dispersion is accounted for in the estimation procedure. The coverage probability based on the bootstrap method when accounting for over-dispersion is close to its nominal level. Figure 1 indicates that the squared bias for the proposed spline-based estimates of Λ_0 under these different working-covariance matrices are negligible relative to their variances. The estimate with $V_3^{(i)}$ behaves the best as it has the smallest variance among the three. Simulation results from *Scenario 2* are similar to the results from *Scenario 1.* The estimate using $V_3^{(i)}$ again behaves better than the estimate with $V_1^{(i)}$ or $V_2^{(i)}$, even though the underlying frailty variable is not Gamma distributed. When data are generated from a Poisson process as in *Scenario 3*, the proposed estimate with $V_3^{(i)}$ behaves very similar to that with $V_2^{(i)}$ which is actually the efficient semiparametric estimate according to Lu et al. (2009). This is mainly due to the fact that the estimate of the over-dispersion parameter is zero most of time in the simulation studies. If the data follow a negative binomial counting process as in *Scenario 4*, the spline based semiparametric GEE estimates using $V_2^{(i)}$ and $V_3^{(i)}$ are similar, both perform slightly better than the estimates using $V_1^{(i)}$.

The simulation results indicate that the proposed spline-based semiparametric projected GEE method with $V_3^{(i)}$ generally produces more efficient estimate of the regression parameter regardless of the distribution of the latent frailty variable and is equally efficient as the semiparametric maximum likelihood estimate when the underlying counting process is indeed Poisson.

4.2 *Application*

The proposed estimating method is applied to the bladder tumor data introduced in 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 tumor number (Z_1) and size (Z_2) at baseline (study entrance), 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 results with the three different working-covariance matrices are shown in Table 3. The tumor number at baseline is positively related to the recurrence of bladder tumor. With one more tumor at baseline, the number of tumors at follow-ups increases by 15.5%, 23.1% and 39.1% on average using the working-covariance matrices $V_1^{(i)}, V_2^{(i)}$ and $V_3^{(i)}$, respectively. Thiotepa instillation effectively decreases the number of recurrent tumors. The number of recurrent tumors in patients with thiotepa instillation is 49.5%, 45.1% and 32.5% of that in placebo group on average using $V_1^{(i)}, V_2^{(i)}$ and $V_3^{(i)}$, respectively. The tumor size and pyridoxin pills are not significantly related to the number of recurrent tumors at follow-up visits. The estimating results using the diagonal working-covariance matrix $V_1^{(i)}$ and the working-covariance matrix based on Poisson process $V_2^{(i)}$ are consistent with the estimating results based on the spline-based semiparametric pseudo-likelihood and the likelihood methods proposed by Lu et al. (2009). The proposed semiparametric projected GEE estimate with 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 panel counts and possible positive correlation among the tumor numbers in non-overlapping time intervals for the underlying tumor progression. The effect of the tumor number at the study entrance and the treatment of thiotepa are more significant when accounting for the correlation between cumulative tumor numbers using the frailty variable. Figure 3 plots the estimated baseline mean function.

5. Final Remark

Modeling panel count data is a challenging task in general. The proposed spline-based semiparametric projected GEE method avoids assuming the underlying count process and borrows the strength from discrete observations within subjects as well as those across subjects to get a spline estimate of the mean function of the counting process. Choosing different working-covariance matrices can accommodate different data structures. The proposed spline-based projected GEE method with the working-covariance matrices $V_3^{(i)}$ accounts for the over-dispersion and inter-correlation between non-overlapping counts. It improves the estimating efficiency and provides a less biased standard error estimation using either the "ad-hoc" parametric GEE sandwich formula or the bootstrap method when over-dispersion is present in data. In our computing algorithm, over-dispersion parameter σ^2 is fixed at its estimate in the first stage and the parameters in the proportional mean function θ are updated in the second stage. Our simulation results (not included in this paper) show that update θ and σ^2 alternately gives similar results.

The proposed model assumes that the observation times are noninformative to the underlying counting process which may be violated in applications. Extension of the proposed method to that scenario requires a further investigation.

References

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

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.

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* 99 710–723.

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* 85 809–822.

Lee, E. W. & Kim, M. Y. (1998). The analysis of correlated panel data using a continuous-time markov model. *Biometrics* 54 1638–1644.

Liang, K.-Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 97 13–22.

Lin, D. & Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association* 96 103–113.

Lin, X. & Carroll, R. (2001). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association* 96 1045–1056.

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.

Schumaker, L. (1981). *Spline Functions: Basic Theory.* New York: Wiley.

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., Park, D.-H., Sun, L. & Zhao, X. (2005). Semiparametric regression analysis of longitudinal data with informative observation times. *Journal of the American Statistical Association* 100 882–889.

Sun, J. & Wei, L.-J. (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *Journal of the Royal Statistical Society, Series B* 62 293–302.

Thall, P. F. (1988). Mixed poisson likelihood regression models for longitudinal interval count data. *Biometrics* 44 197–209.

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.-J., Lin, D. & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distribution. *Journal of the American Statistical Association* 1989 1065–1073.

Wellner, J. A. & Zhang, Y. (2000). Two estimators of the mean of a counting process with anel count data. *The Annals of Statistics* 28 779–814.

Wellner, J. A. & Zhang, Y. (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *The Annals of Statistics* 35 2106–2142.

Wellner, J. A., Zhang, Y. & Liu, H. (2004). A semiparametric regression model for panel count data: when do pseudolikelihood estimators become badly inefficient? In L. DY. & H. PJ, eds., *Proceedings of the second Seattle Biostatistical Symposium: Analysis of Correlated Data*. Springer-Verlag, New York, 143–174.

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* 75 621–629.

Zeger, S. L. & Diggle, P. (1994). Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics* 50 689–699.

Zhang, Y. (2002). A semiparametric pseudo likelihood estimation method for panel count data. *Biometrika* 89 39–48.

Zhang, Y. & Jamshidian, M. (2003). The gamma-frailty poisson model for the nonparametric estimation of panel count data. *Biometrics* 59 1099–1106.

6. Appendix

In this section, we show that spline-based semiparametric GEE with $V_1^{(i)}, V_2^{(i)}$ and $V_3^{(i)}$ coincide with the score equations under the different models. First, we define the following notations

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

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

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

6.1 *Agreement between the GEE with $V_1^{(i)}$ and the score equation of the spline-based pseudo-likelihood*

Using $V_1^{(i)}$ as the working-covariance matrix, the U function of Equation (3) can be rewritten as

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

This is exactly the score function of the spline-based pseudo-likelihood derived by Lu et al. (2009)

6.2 *Agreement between the GEE with $V_2^{(i)}$ and the score equation of the spline-based likelihood*

When using $V_2^{(i)}$ as the working-covariance matrix, the U function of Equation (3) can be rewritten as

$$U(\theta) = \sum_{i=1}^{n} \left(1_{K_i} Z_i^T, B^{(i)}\right)^T \operatorname{diag}\left(\mu_{K_i,1}^{(i)}, \cdots, \mu_{K_i,K_i}^{(i)}\right) V_2^{(i)^{-1}} \left(\mathbb{N}^{(i)} - \mu^{(i)}\right).$$

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

$$\tilde{l}_n(\theta; D) = \sum_{i=1}^{n} \sum_{j=1}^{K_i} \left[\Delta\mathbb{N}_{K_i,j}^{(i)} \log \Delta\tilde{\Lambda}_{K_i,j}^{(i)} + \Delta\mathbb{N}_{K_i,j}^{(i)} \beta^T Z_i - e^{\beta^T Z_i} \Delta\tilde{\Lambda}_{K_i,j}^{(i)}\right] \tag{6}$$

where

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

A careful examination of this likelihood shows that its score function can be rewritten in a matrix form,

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

Since

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

$$= \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & -1 & 1 \end{pmatrix} \operatorname{diag}\left(\mu^{(i)}_{K_i,1}, \cdots, \mu^{(i)}_{K_i,K_i}\right)\left(1_{k_i} Z_i^T, B^{(i)}\right)$$

The score function can be further written as

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

where

$$\Sigma = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & -1 & 1 \end{pmatrix}^T \operatorname{diag}\left(\Delta\mu^{(i)}_{K_i,1}, \cdots, \Delta\mu^{(i)}_{K_i,K_i}\right)^{-1} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & -1 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{\mu^{(i)}_{K_i,1}} & -\frac{1}{\mu^{(i)}_{K_i,2}-\mu^{(i)}_{K_i,1}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\mu^{(i)}_{K_i,2}-\mu^{(i)}_{K_i,1}} & -\frac{1}{\mu^{(i)}_{K_i,3}-\mu^{(i)}_{K_i,2}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & -\frac{1}{\mu^{(i)}_{K_i,K_i}-\mu^{(i)}_{K_i,K_i}} \\ 0 & 0 & 0 & \cdots & \frac{1}{\mu^{(i)}_{K_i,K_i}-\mu^{(i)}_{K_i,K_i}} \end{pmatrix} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & -1 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{\mu^{(i)}_{K_i,1}} + \frac{1}{\mu^{(i)}_{K_i,2}-\mu^{(i)}_{K_i,1}} & -\frac{1}{\mu^{(i)}_{K_i,2}-\mu^{(i)}_{K_i,1}} & \cdots & \cdots & 0 \\ -\frac{1}{\mu^{(i)}_{K_i,2}-\mu^{(i)}_{K_i,1}} & \frac{1}{\mu^{(i)}_{K_i,2}-\mu^{(i)}_{K_i,1}} + \frac{1}{\mu^{(i)}_{K_i,3}-\mu^{(i)}_{K_i,2}} & -\frac{1}{\mu^{(i)}_{K_i,3}-\mu^{(i)}_{K_i,2}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & \frac{1}{\mu^{(i)}_{K_i,K_i}-\mu^{(i)}_{K_i,K_i-1}} \end{pmatrix}$$

It is a straightforward algebra to verify that $\Sigma = \left(V_2^{(i)}\right)^{-1}$, so the GEE with the working-covariance matrix $V_2^{(i)}$ is the same as the score equation of the likelihood given in (6).

6.3 *Agreement between the GEE with $V_3^{(i)}$ and the score equation of the likelihood of Gamma-Frailty Poisson model*

By the derivation of the equivalence between the GEE with $V_2^{(i)}$ and the score equation of likelihood in (6), we have

$$\left(\frac{\partial\mu^{(i)}}{\partial\theta}\right)^T V_2^{(i)^{-1}}\left(\mathbb{N}^{(i)} - \mu^{(i)}\right) = \sum_{j=1}^{K_i}\left(\frac{\partial\Delta\mu^{(i)}_{K_i,j}}{\partial\theta}\right)\left(\frac{\Delta\mathbb{N}^{(i)}_{K_i,j}}{\Delta\mu^{(i)}_{K_i,j}} - 1\right) \tag{7}$$

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

$$\left(\frac{\partial \mu^{(i)}}{\partial \theta}\right)^T V_2^{(i)^{-1}} \mu^{(i)} = \sum_{j=1}^{K_i} \frac{\partial \Delta \mu_{K_i,j}^{(i)}}{\partial \theta} = \frac{\partial \mu_{K_i,K_i}^{(i)}}{\partial \theta} = \left(Z_i^T, B_{K_i,K_i}^{(i)T}\right)^T \mu_{K_i,K_i}^{(i)} \tag{8}$$

Taking the β part of (7), we have

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

The left hand side of (8) can be rewritten as

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

and the right hand side of (8) can also be rewritten as

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

This implies that

$$\mu^{(i)^T} V_2^{(i)^{-1}} \left(\mathbb{N}^{(i)} - \mu^{(i)}\right) = \mathbb{N}_{K_i,K_i}^{(i)} - \mu_{K_i,K_i}^{(i)}. \tag{9}$$

Again letting $\mathbb{N}^{(i)} = 2\mu^{(i)}$, we obtain

$$\mu^{(i)} V_2^{(i)^{-1}} \mu^{(i)} = \mu_{K_i,K_i}^{(i)} \tag{10}$$

The U function of Equation with $V_3^{(i)}$ as the working-covariance matrix can then be rewritten as,

$$\begin{aligned} U(\theta) &= \sum_{i=1}^{n} \left(\frac{\partial \mu^{(i)}}{\partial \theta}\right)^T \left(V_2^{(i)} + \sigma^2 \mu^{(i)} \mu^{(i)^T}\right)^{-1} \left(\mathbb{N}^{(i)} - \mu^{(i)}\right) \\ &= \sum_{i=1}^{n} \left(\frac{\partial \mu^{(i)}}{\partial \theta}\right)^T \left(V_2^{(i)} - \frac{\sigma^2}{1 + \sigma^2 \mu^{(i)^T} \left(V_2^{(i)}\right)^{-1} \mu^{(i)}} \left(V_2^{(i)}\right)^{-1} \mu^{(i)} \mu^{(i)^T} V_2^{-1}\right) \left(\mathbb{N}(T_i) - \mu^{(i)}\right) \\ &= \sum_{i=1}^{n} \left\{ \left(\frac{\partial \mu^{(i)}}{\partial \theta}\right)^T V_2^{(i)^{-1}} \left(\mathbb{N}^{(i)} - \mu^{(i)}\right) - \frac{\sigma^2}{1 + \sigma^2 \mu^{(i)^T} V_2^{-1} \mu^{(i)}} \left(\frac{\partial \mu^{(i)}}{\partial \theta}\right)^T V_2^{(i)^{-1}} \mu^{(i)} \right. \end{aligned}$$

$$
\begin{aligned}
&\quad \times \mu^{(i)^T} V_2^{(i)^{-1}} \left(\mathbb{N}^{(i)} - \mu^{(i)}\right)\Big\} \\
&= \sum_{i=1}^{n} \left\{ \sum_{j=1}^{K_i} \left(\mu_{K_i,j}^{(i)} \left(Z_i^T, B_{K_i,j}^{(i)^T}\right)^T - \mu_{K_i,j-1}^{(i)} \left(Z_i^T, B_{K_i,j-1}^{(i)^T}\right)^T \right) \left(\frac{\Delta \mathbb{N}_{K_i,j}^{(i)}}{\Delta \mu_{K_i,j}^{(i)}} - 1 \right) - \right. \\
&\quad \left. \frac{\sigma^2}{1 + \sigma^2 \mu_{K_i,K_i}^{(i)}} \left(Z_i^T, B_{K_i,K_i}^{(i)^T}\right)^T \mu_{K_i,K_i}^{(i)} \left(\mathbb{N}_{K_i,K_i}^{(i)} - \mu_{K_i,K_i}^{(i)}\right) \right\} \qquad \text{(by Equations (8)-(10))} \\
&= \sum_{i=1}^{n} \left\{ \sum_{j=1}^{K_i} \left(\mu_{K_i,j}^{(i)} \left(Z_i^T, B_{K_i,j}^{(i)^T}\right)^T - \mu_{K_i,j-1}^{(i)} \left(Z_i^T, B_{K_i,j-1}^{(i)^T}\right)^T \right) \frac{\Delta \mathbb{N}_{K_i,j}^{(i)}}{\Delta \mu_{K_i,j}^{(i)}} \right. \\
&\quad \left. - \frac{1 + \sigma^2 \mathbb{N}_{K_i,K_i}^{(i)}}{1 + \sigma^2 \mu_{K_i,K_i}^{(i)}} \left(Z_i^T, B_{K_i,K_i}^{(i)^T}\right)^T \mu_{K_i,K_i}^{(i)} \right\}
\end{aligned}
$$

This is exactly the score function of the Gamma-frailty Poisson likelihood.

Table 1

Simulations results of the splines-based sieve semiparametric GEE estimators with three different covariance matrices for data from scenario 1 and scenario 2

	Scenario 1: Frailty Poisson Data						Scenario 2: Mixture Poisson Data					
	N=50			N=100			N=50			N=100		
	$V_1^{(i)}$	$V_2^{(i)}$	$V_3^{(i)}$	$V_1^{(i)}$	$V_2^{(i)}$	$V_3^{(i)}$	$V_1^{(i)}$	$V_2^{(i)}$	$V_3^{(i)}$	$V_1^{(i)}$	$V_2^{(i)}$	$V_3^{(i)}$
β_1												
Bias	0.0519	0.0528	0.0162	0.0376	0.0370	-0.0001	0.0043	0.0049	0.0020	-0.0023	-0.0026	-0.0019
MC-sd	0.3640	0.3647	0.2476	0.2657	0.2639	0.1810	0.0810	0.0809	0.0640	0.0601	0.0606	0.0425
SSE	0.2208	0.2221	0.1807	0.1765	0.1772	0.1382	0.0564	0.0559	0.0546	0.0454	0.0452	0.0405
BSE	0.3063	0.3053	0.2739	0.2145	0.2137	0.1683	0.0732	0.0728	0.0649	0.0529	0.0528	0.0434
CP1	0.7089	0.7059	0.8383	0.7590	0.7500	0.8776	0.7990	0.7980	0.8980	0.8320	0.8299	0.9305
CP2	0.8544	0.8584	0.9452	0.8540	0.8480	0.9330	0.9050	0.9080	0.9370	0.9035	0.8994	0.9429
β_2												
Bias	-0.0090	-0.0192	-0.000	-0.0712	-0.0691	-0.0455	0.0017	0.0010	-0.0020	0.0016	-0.0001	0.0007
MC-sd	1.1374	1.1281	0.7979	0.8326	0.8196	0.5477	0.2540	0.2504	0.2015	0.1840	0.1812	0.1384
SSE	0.7964	0.8011	0.6345	0.6405	0.6407	0.4715	0.1999	0.1990	0.1800	0.1565	0.1556	0.1314
BSE	0.9846	0.9799	0.8617	0.7089	0.7048	0.5347	0.2312	0.2307	0.2045	0.1675	0.1668	0.1384
CP1	0.8235	0.8185	0.8936	0.8540	0.8660	0.9087	0.8640	0.8630	0.9150	0.8973	0.8973	0.9367
CP2	0.9073	0.9103	0.9432	0.8950	0.8960	0.9510	0.9190	0.9260	0.9390	0.9212	0.9315	0.9419
β_3												
Bias	0.0391	0.0375	0.0236	-0.0023	-0.0031	-0.0082	0.0012	0.0015	0.0029	0.0022	0.0027	0.0035
MC-sd	0.6126	0.6042	0.4499	0.4177	0.4145	0.2939	0.1443	0.1407	0.1232	0.1015	0.0981	0.0805
SSE	0.4491	0.4487	0.3802	0.3488	0.3479	0.2801	0.1179	0.1156	0.1098	0.0908	0.0892	0.0796
BSE	0.5426	0.5378	0.4827	0.3775	0.3751	0.3053	0.1321	0.1297	0.1208	0.0954	0.0939	0.0830
CP1	0.8534	0.8594	0.9140	0.8860	0.8900	0.9509	0.8750	0.8740	0.9070	0.9170	0.9232	0.9471
CP2	0.9123	0.9232	0.9531	0.9130	0.9170	0.9570	0.9150	0.9170	0.9340	0.9274	0.9336	0.9585

MC-sd: Monte-Carlo standard deviation; SSE: Ad-hoc parametric sandwich standard error estimation; BSE: Bootstrap standard error estimation; CP1: 95% Coverage based on SSE; CP2: 95% Coverage based on BSE;

Table 2

Simulations results of the splines-based sieve semiparametric GEE estimators with three different covariance matrices for data from scenario 3 and scenario 4

	Scenario 3: Poisson Data						Scenario 4: Negative Binomial Count Data					
	N=50			N=100			N=50			N=100		
	$V_1^{(i)}$	$V_2^{(i)}$	$V_3^{(i)}$	$V_1^{(i)}$	$V_2^{(i)}$	$V_3^{(i)}$	$V_1^{(i)}$	$V_2^{(i)}$	$V_3^{(i)}$	$V_1^{(i)}$	$V_2^{(i)}$	$V_3^{(i)}$
β_1												
Bias	-0.0001	0.0003	0.0002	0.0012	0.0010	0.0009	-0.0004	-0.0004	-0.0008	0.0008	0.0009	0.0009
MC-sd	0.0314	0.0291	0.0292	0.0209	0.0195	0.0196	0.0357	0.0344	0.0350	0.0248	0.0231	0.0236
SSE	0.0276	0.0259	0.0260	0.0189	0.0176	0.0177	0.0312	0.0296	0.0302	0.0212	0.0200	0.0206
BSE	0.0346	0.0324	0.0329	0.0218	0.0203	0.0206	0.0390	0.0369	0.0383	0.0246	0.0232	0.0242
CP1	0.9180	0.9100	0.9090	0.9080	0.9220	0.9230	0.9070	0.9000	0.9040	0.9120	0.9080	0.9100
CP2	0.9680	0.9660	0.9680	0.9550	0.9630	0.9650	0.9630	0.9590	0.9630	0.9480	0.9490	0.9600
β_2												
Bias	0.0023	0.0016	0.0015	0.0000	0.0000	-0.0002	0.0014	0.0020	0.0017	-0.0019	-0.0024	-0.0026
MC-sd	0.1016	0.0969	0.0972	0.0687	0.0636	0.0641	0.1180	0.1100	0.1120	0.0791	0.0744	0.0752
SSE	0.0879	0.0826	0.0828	0.0625	0.0584	0.0584	0.0991	0.0947	0.0957	0.0707	0.0671	0.0679
BSE	0.1045	0.0976	0.0986	0.0683	0.0637	0.0643	0.1178	0.1122	0.1150	0.0771	0.0730	0.0747
CP1	0.9130	0.8960	0.8960	0.9190	0.9320	0.9300	0.8950	0.9080	0.9100	0.9200	0.9280	0.9270
CP2	0.9500	0.9410	0.9420	0.9400	0.9510	0.9540	0.9380	0.9480	0.9490	0.9360	0.9420	0.9490
β_3												
Bias	-0.0002	-0.0008	-0.0008	0.0014	0.0014	0.0015	-0.0038	-0.0038	-0.0037	0.0014	0.0016	0.0017
MC-sd	0.0694	0.0657	0.0657	0.0495	0.0449	0.0450	0.0799	0.0743	0.0749	0.0568	0.0538	0.0540
SSE	0.0651	0.0612	0.0613	0.0462	0.0431	0.0432	0.0738	0.0701	0.0706	0.0525	0.0496	0.0499
BSE	0.0728	0.0680	0.0684	0.0488	0.0454	0.0456	0.0822	0.0780	0.0791	0.0553	0.0522	0.0527
CP1	0.9180	0.9160	0.9170	0.9240	0.9290	0.9290	0.9190	0.9230	0.9220	0.9200	0.9090	0.9160
CP2	0.9480	0.9420	0.9430	0.9370	0.9430	0.9430	0.9480	0.9570	0.9570	0.9330	0.9290	0.9360

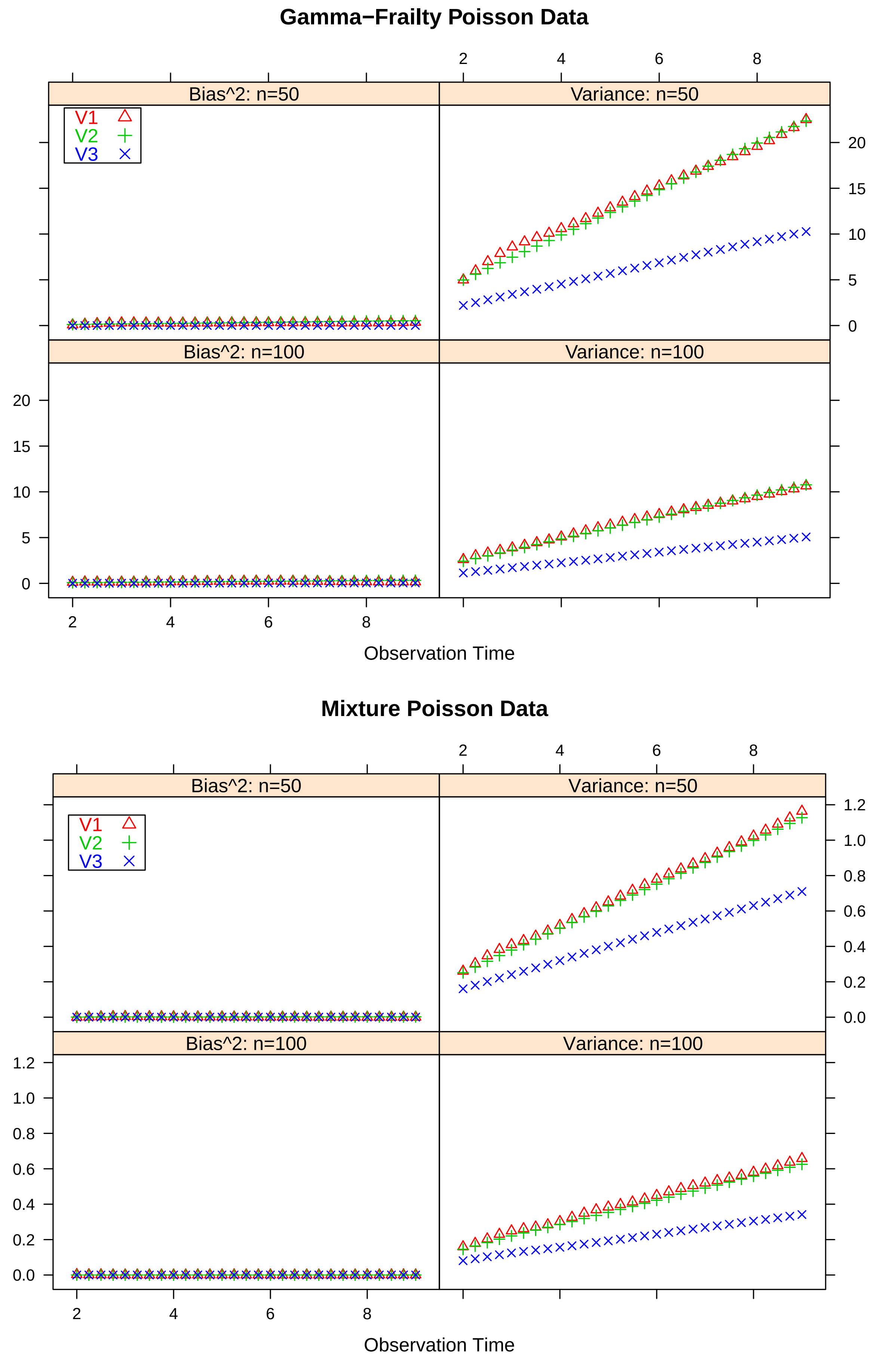

Figure 1. *Simulation results for estimations of the baseline mean function,* $\Lambda_0(t) = 2t^{1/2}$

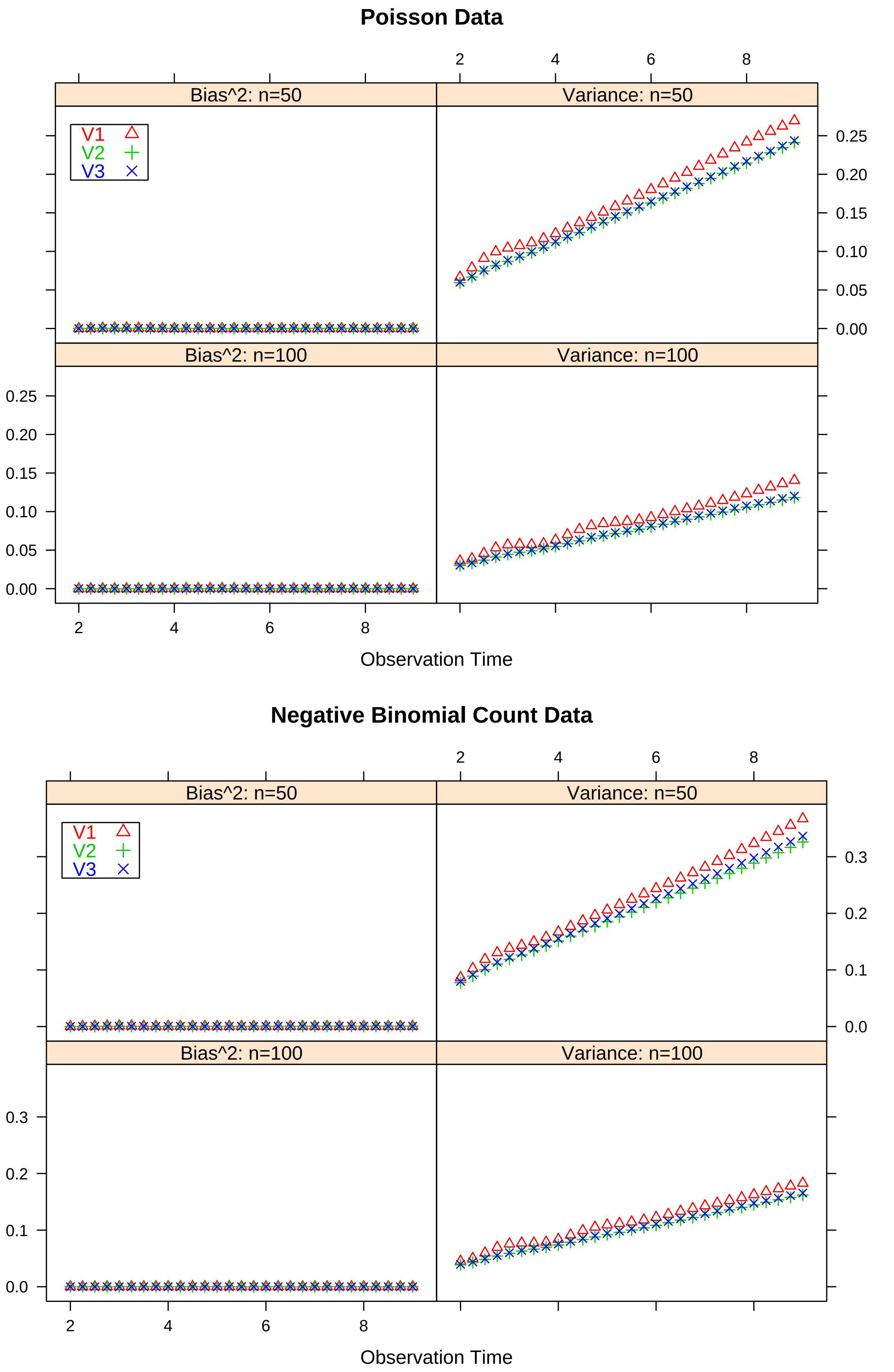

Figure 2. *Simulation results for estimations of the baseline mean function,* $\Lambda_0(t) = 2t^{1/2}$

Table 3

The spline-based sieve semiparametric inference for bladder tumor data

$\underline{V_1}$

	Est.	Sandwich Std.	p-value	Bootstrap Std.	p-value
Z1	0.1444	0.0518	0.0053	0.0660	0.0286
Z2	−0.0447	0.0488	0.3595	0.0449	0.3189
Z3	0.1776	0.2246	0.4292	0.2894	0.5395
Z4	−0.6966	0.2397	0.0037	0.3250	0.0321

$\underline{V_2}$

	Est.	Sandwich Std.	p-value	Bootstrap Std.	p-value
Z1	0.2075	0.0677	0.0022	0.0905	0.0499
Z2	−0.0353	0.0732	0.6299	0.0691	0.0972
Z3	0.0637	0.3502	0.8556	0.3891	0.9730
Z4	−0.7960	0.2952	0.0070	0.3780	<0.0001

$\underline{V_3}$

	Est.	Sandwich Std.	p-value	Bootstrap Std.	p-value
Z1	0.3289	0.0702	0.0000	0.0994	0.0009
Z2	0.0054	0.0767	0.9437	0.0809	0.9484
Z3	0.0213	0.4069	0.9583	0.4081	0.9782
Z4	−1.0692	0.3389	0.0016	0.3944	0.0044

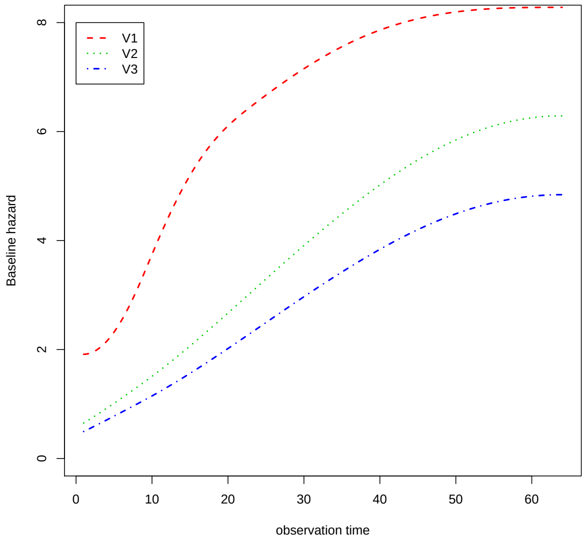

Figure 3. *Point estimates of the baseline mean function*