Spline-based semiparametric projected generalized estimating equation method for panel count data

LEI HUA

Center for Biostatistics in AIDS Research, Department of Biostatistics, Harvard School of Public Health,
FXB 514, 651 Huntington Avenue, Boston, MA 02115, USA
lhua@sdac.harvard.edu

YING ZHANG

Department of Biostatistics, The University of Iowa, C22 GH,
200 Hawkins Drive, Iowa City, IA 52242, USA

SUMMARY

We propose to analyze panel count data using a spline-based semiparametric projected generalized estimating equation (GEE) method with the proportional mean model

\[ E(N(t)|Z) = \Lambda_0(t) \exp(\beta^T_0 Z) \]

The natural logarithm 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 obtained by projecting the GEE estimates into the feasible domain using a weighted isotonic regression (IR). The proposed method avoids assuming any parametric structure of the baseline mean function or any stochastic model for the underlying counting process. Selection of the working covariance matrix that accounts for overdispersion improves the estimation efficiency and leads to less biased variance estimations. Simulation studies are conducted using different working covariance matrices in the GEE to investigate finite sample performance of the proposed method, to compare the estimation efficiency, and to explore the performance of different variance estimates in presence of overdispersion. Finally, the proposed method is applied to a real data set from a bladder tumor clinical trial.

Keywords: Counting process; Generalized estimating equation; Monotone B-splines; Overdispersion; Semiparametric model.

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 first studied by Byar and others (1980). Subjects with superficial bladder tumor were randomized into 1 of 3 treatment groups: placebo, pyridoxine 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 bladder tumor recurrence.

*To whom correspondence should be addressed.

© The Author 2011. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.
Various likelihood-based methods have been studied for the analysis of panel count data, see examples in Lee and Kim (1998) and Sun and Kalbfleisch (1995), among others. Wellner and Zhang (2000, 2007) studied the nonparametric and semiparametric maximum pseudolikelihood estimator (MPLE) and the maximum likelihood estimator (MLE) for panel count data assuming the underlying counting process as a nonhomogeneous Poisson process. Lu and others (2007, 2009) studied the spline-based sieve version of MPLE and MLE by approximating the baseline mean function using monotone B-spline functions. Although these Poisson likelihood-based estimation methods are consistent and robust against the underlying Poisson process assumption, the Poisson process-based likelihood does not take into account the overdispersion problem that often occurs in various applications of longitudinal count data. As acknowledged in the literature of Poisson regression and generalized linear models (GLM) (Cox, 1983; Cox and Snell, 1989; Breslow, 1990; Ganio and Schafer, 1992; Lawless and Nadeau, 1995), neglecting overdispersion not only results in a loss of estimation efficiency, but also often leads to serious underestimation of standard errors of the estimated regression parameters. The latter may further cause inflated test statistics and therefore lead to the inflation of type I error. To correctly estimate the variance, different methods, including ad hoc corrections and those based on resampling, have been discussed in the literature of GLM. However, none of the methods has a uniformly satisfactory result, see examples in Palmer and others (2007) and references therein. Both loss of estimation efficiency and underestimation of the variance are observed in the analysis of panel count data when overdispersion is neglected.

A classical approach to address the overdispersion problem in GLM is to involve a latent variable to account for the extra dispersion. With similar idea, mixed Poisson likelihood with a multiplicative frailty term has been introduced into the analysis of panel count data. Zhang and Jamshidian (2003) introduced a gamma frailty variable to the nonparametric estimation of the mean function of panel count data. Hu and others (2009b) discussed an alternative method based on quasi-score equations with additional quadratic estimation equations to account for the overdispersion. Both methods based on the gamma distribution assumption of the frailty term need to be implemented using self-consistent algorithms. Huang and others (2006) introduced a latent frailty to account for informative observation times and avoided specifying its distribution through a conditional maximum likelihood approach. The robustness of their approach in terms of addressing the overdispersion problem is not specifically addressed.

Instead of specifying the full likelihood of the counting process, Hu and others (2009a) discussed a nonparametric generalized least squares estimation, where the means of the cumulative counts are estimated by projecting sample means of available data to the monotone space. Different weight functions in the generalized least squares estimation lead to estimates closely related to the nonparametric MPLE and MLE by Wellner and Zhang (2000). Hu and others (2009a) showed that in presence of overdispersion, Wellner–Zhang’s MLE is no longer efficient, and it remains unclear how to improve the estimation efficiency.

The Generalized Estimating Equation (GEE) method, developed by Liang and Zeger (1986), has been widely used in parametric regression analysis of longitudinal data. It provides a robust inference with only weak assumptions about the moments of the underlying distribution. The same idea has been generalized to the analysis of recurrent event data by Zeger and Diggle (1994), Hoover and others (1998), Lin and Ying (2001), Wu and Zhang (2002) and Lin and Carroll (2001), among others. Sun and Wei (2000) and Sun and others (2005) studied panel count data based on estimating equation. But the validity of their methods relies on some assumptions of observation times, which may be hard to justify in applications.

In this manuscript, we discuss a semiparametric analysis of panel count data within the framework of GEE, assuming the proportional mean model as in Wellner and Zhang (2007) and Lu and others (2009), that is,

\[ E(N(t)|Z) = \Lambda_0(t) e^{\beta_0^T Z}, \quad \theta_0 = (\beta_0, \Lambda_0) \in \mathcal{R}^d \times \mathcal{F}^+, \]
where $\mathcal{F}^+ \subset \mathcal{F}$ is the subset of the function space $\mathcal{F}$ with the monotone nondecreasing constraint for the baseline mean function $\Lambda_0$. In the likelihood-based approaches, the estimator is the maximizer of some likelihood $l_n(\theta)$ within the proper parameter space, that is, $\hat{\theta}_n = \arg\max_{\theta \in \mathcal{R}^d \times \mathcal{F}^+} l_n(\theta)$, $\mathcal{F}^+_n \subset \mathcal{F}^+$. Denote $\hat{\theta}_n = \arg\max_{\theta \in \mathcal{R}^d \times \mathcal{F}_n} l_n(\theta)$, $\mathcal{F}_n \subset \mathcal{F}$, the maximizer of the likelihood without the monotone constraint. By second-order Taylor expansion,

$$l_n(\hat{\theta}_n) = l_n(\tilde{\theta}_n) - \frac{1}{2}(\tilde{\theta}_n - \hat{\theta}_n)^T W(\tilde{\theta}_n - \hat{\theta}_n),$$

(1.2)

where $W = -\frac{\partial^2}{\partial \theta^2} l_n(\tilde{\theta}_n)$ and $\tilde{\theta}_n$ is closer to $\hat{\theta}_n$ than $\hat{\theta}_n$ in some metric. The likelihood-based approaches can be considered as a 2-stage estimation problem: (1) obtaining the optimal point $\hat{\theta}_n = (\hat{\beta}_n, \hat{\Lambda}_n)$ by maximizing $l_n(\theta)$ without constraint and (2) projecting $\hat{\theta}_n$ to the proper feasible space by minimizing the second term in (1.2), that is, $\hat{\theta}_n = \arg\min_{\theta \in \mathcal{R}^d \times \mathcal{F}^+_n} (\tilde{\theta}_n - \theta)^T W(\tilde{\theta}_n - \theta)$. We apply similar 2-stage estimation idea in the GEEs: Instead of maximizing a likelihood, the nonconstrained update from the first stage is obtained by solving GEE. By doing so, we only need to specify the proportional mean in (1.1) and a working covariance matrix that may reasonably describe the correlation structure of the cumulative counts; in the second stage, the update of the baseline mean function $\Lambda$ from GEE is projected to the cone depicted by the monotone constraints using a quadratic programming. Such a 2-stage algorithm iterates until convergence to guarantee the estimated mean parameter, $(\hat{\beta}_n, \hat{\Lambda}_n)$ approaches its limiting value within the proper parameter space.

Through extensive simulation studies, we demonstrate that by carefully selecting a covariance matrix, specifically, by accounting for overdispersion driven by the data, the proposed estimator improves the estimation efficiency and the variance estimate is generally satisfactory. 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 discusses variance estimation of the projected GEE estimate of the regression parameters. Section 5 provides numerical results including simulation studies and the analysis of the bladder tumor data. Finally, we give some concluding remarks in Section 6. Some technical details and additional numerical results are given in supplementary material available at Biostatistics online.

## 2. Spline-based Semiparametric Projected GEE Method

Suppose, $\mathbb{N} = \{N(t) : t \geq 0\}$ 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 $\mathcal{T}_{K} \equiv (T_{K,1}, T_{K,2}, \ldots, T_{K,K})$ and $\mathcal{N} \equiv (\mathbb{N}(T_{K,1}), \mathbb{N}(T_{K,2}), \ldots, \mathbb{N}(T_{K,K}))$, the cumulative event counts at these discrete observation times. We assume the number of observations and the observation times, $(K, \mathcal{T}_{K})$, are independent of and noninformative to the point process $\mathbb{N}$, conditional on the covariate vector $Z$. Panel count data are composed of a random sample of $X_1, X_2, \ldots, X_n$, where the observation $X_i$ consists of $(K_i, \mathcal{T}_{K_i}, \mathbb{N}^{(i)}, Z_i)$ with $\mathcal{T}_{K_i} = (T_{K_i,1}^{(i)}, T_{K_i,2}^{(i)}, \ldots, T_{K_i,K_i}^{(i)})$ and $\mathbb{N}^{(i)} = \left(\mathbb{N}^{(i)}(T_{K_i,1}^{(i)}), \mathbb{N}^{(i)}(T_{K_i,2}^{(i)}), \ldots, \mathbb{N}^{(i)}(T_{K_i,K_i}^{(i)})\right)$.

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 $t = \{L = t_0 = t_1 = \cdots = t_l < t_{l+1} < \cdots < t_{l+m_n} = t_{l+m_n+1} = \cdots = t_{m_n+2l} = U\}$ partition $[L, U]$ into $m_n + 1$ subintervals, where $m_n \approx n^\nu$ is a positive integer such that $\max_{1 \leq k \leq m_n+1} (t_{l+k} - t_{l+k-1}) = O(n^{-\nu})$. Denote $\phi_{l,t}$ a class of B-spline functions of order $l, l \geq 1$. $\phi_{l,t}$ is spanned by a series of B-spline basis functions $B(t) = (B_1(t), B_2(t), \ldots, B_{q_n}(t))^T$, where $q_n = m_n + l$, that is, $\phi_{l,t} = \{a^T B(t) : a \in J_n\}$ with $J_n = \{a = (a_1, a_2, \ldots, a_{q_n})^T : a_i \in R, i = 1, \ldots, q_n\}$. A subclass of $\phi_{l,t}$, $\psi_{l,t} = \{a^T B(t), a \in J_n^+\}$ with $J_n^+ = \{a \in J_n : a_1 \leq a_2 \leq \cdots \leq a_{q_n}\}$ 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 $\Lambda_0(t)$ can be found. With this approximation, the proportional mean in (1.1) can be rewritten as $E(N(t)|Z) = \Lambda_0(t)e^{\beta Z} = \exp \{ \sum_{i=1}^{n_i} \alpha_i B_1(t) + \beta Z \}$. To estimate the mean parameter, a 2-stage iterative algorithm is implemented. In the first stage, a Newton–Raphson (NR) type of algorithm that intends to solve for the generalized estimation equations of the form

$$U(\theta) = \sum_{i=1}^{n} \left( \frac{\partial \mu_i}{\partial \theta} \right)^T V^{-1}(\theta)(\mu_i(\theta) - \mu(\theta)) = 0 \tag{2.1}$$

is implemented to obtain the NR update

$$\tilde{\theta} = (\tilde{\beta}^T, \tilde{\alpha}^T)^T = (\tilde{\beta}_1, \ldots, \tilde{\beta}_d, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_q)^T,$$

where $\mu_i(\theta) = \left( \mu_{K_1,1}(\theta), \mu_{K_1,2}(\theta), \ldots, \mu_{K_i,1}(\theta) \right)^T$ with $\mu_{K_i,j}(\theta) = \exp \left( \sum_{l=1}^{q_n} \alpha_l B_l(t) \right)^j + \beta^T Z_i$ for $j = 1, 2, \ldots, K_i$. In the second stage, the NR update, $\tilde{\alpha}$ is projected into the feasible space $J_n^+$ by a quadratic programming:

$$\hat{\alpha} = \text{Proj}_W \left[ \tilde{\alpha}, J_n^+ \right] = \arg \min_{\alpha \in J_n^+} (\alpha - \tilde{\alpha})^T W(\alpha - \tilde{\alpha}), \tag{2.2}$$

where $W$ is a positive definite matrix. Instead of specifying $W$ to be the negative Hessian matrix as in the likelihood-based approach, expectation of the negative derivative of GEE could be used for numerical convenience. A different choice of $W$ is to use a diagonal matrix with the elements chosen as the diagonal elements of the expectation of the negative derivative of GEE. With such a choice, the second-step projection can be easily implemented using isotonic regression (IR) and is used in our numerical studies. Section 3 provides further details on the implementation of the algorithm. After the estimates of the spline coefficients $\hat{\alpha} = \{ \hat{\alpha}_l, l = 1, 2, \ldots, q_n \}$ are obtained subject to the constraints, the spline-based semiparametric projected GEE estimate of $\Lambda_0$ is taken to be $\hat{\Lambda}(t) = \exp \left( \sum_{i=1}^{n_i} \hat{\alpha}_l B_l(t) \right)$.

At the first GEE-based updating stage, $V^{(i)}$ is the working covariance matrix for the panel counts from the $i$th process and plays a pivotal role in determining the estimation efficiency. Different choices of this working 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 Poisson process, that is, $V^{(i)}_1 = (\sigma_{jl})_{K_i \times K_i}$ with $\sigma_{jl} = \mu^{(i)}_{K_j,j}$ for $j = l$ and $\sigma_{jl} = 0$ for $j \neq l$. 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 pseudolikelihood studied by Lu and others (2009), and the proof is given in Appendix A.1 of the supplementary material available at Biostatistics online. Alternative to using the diagonal matrix that ignores the correlation among the cumulative counts, using a working covariance matrix that accommodates such correlation will intuitively produce more efficient estimate. The covariance function based on the Poisson counting process $\text{Cov}(N(t_1), N(t_2)) = E(N(t_1))$ for $t_1 \leq t_2$ leads to the selection of the working covariance matrix $V^{(i)}_2 = (\sigma_{jl})_{K_i \times K_i}$ with $\sigma_{jl} = \mu^{(i)}_{K_i,\text{min}(j,l)}$. 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 and others (2009), and the proof is given in Appendix A.2 of the supplementary material available at Biostatistics online. Despite the improved estimation efficiency using $V^{(i)}_2$ compared to the one using $V^{(i)}_1$, it still imposes unrealistic assumptions to the covariance structure of the data: First, it assumes the variance of the cumulative counts equal to the mean, that is, no overdispersion is accounted for the data and second, it assumes independence of the counts between nonoverlapping intervals. When either of these assumptions is violated, the estimate based on $V^{(i)}_2$ may
not be efficient. More importantly, neglecting the overdispersion may severely bias the variance estimation of the estimated regression parameter and inflate type I error in inference.

To account for overdispersion in the proposed projected GEE framework, we adopt the method developed by Zeger (1988) to consider a latent frailty process and only specify the first moment of the frailty term $\gamma$, that is, $E(\gamma) = 1$, which guarantees the identifiability of the model and does not violate the proportional mean model specified in (1.1). Denote $\text{Var}(\gamma) = \sigma^2$, the marginal variance function based on the Frailty Poisson process is $\text{Var}(N(i)) = \mu_t + \sigma^2 \mu_t^2$, where $\mu_t = E(N(i))$. The correlation between successive counts is accounted for by the frailty parameter as well, namely $\text{Cov}(N(t_1), N(t_2)) = \mu_{t_1} + \sigma^2 \mu_{t_1} \mu_{t_2}$, for $t_1 < t_2$. This leads to a working covariance matrix $V_3^{(i)} = (\sigma j_l)_{K_i \times K_i}$ with $\sigma j_l = \mu_{K_i, \min(j, l)} + \sigma^2 \mu_{K_i, j} \mu_{K_i, l}$, and it can be rewritten as $V_3^{(i)} = V_2^{(i)} + \sigma^2 (\mu(i))\otimes^2$, where $a\otimes^2 = aa^T$. $V_2^{(i)}$ 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 frailty variable, $\gamma \sim \Gamma(1/\sigma^2, 1/\sigma^2)$, the cumulative count follows a nonhomogeneous Poisson process with mean $\gamma \Lambda(i) e^{\theta Z}$. The proof is given in Appendix A.3 of the supplementary material available at Biostatistics online.

3. NUMERICAL ALGORITHM

To compute the proposed projected GEE estimate of $(\beta, \Lambda)$, an estimate of the overdispersion parameter $\sigma^2$ is needed. It is possible to create an extra estimating equation using the second moment to jointly solve for $(\beta, \alpha, \sigma^2)$. But this approach is numerically cumbersome and hence not recommended here. We propose to estimate $\sigma^2$ externally to the GEE using Zeger’s method of moments estimate, that is,

$$
\hat{\sigma}^2_n = \frac{\sum_{i=1}^{n} \sum_{j=1}^{K_i} (N_{ij} - \hat{\mu}_{ij})^2 - \sum_{i=1}^{n} \sum_{j=1}^{K_i} \hat{\mu}_{ij}^2}{\sum_{i=1}^{n} \sum_{j=1}^{K_i} 2 \hat{\mu}_{ij}},
$$

(3.1)

where $N_{ij} = N(i)(T_{K_i,j}^{(i)})$, $\hat{\mu}_{ij}$ is a consistent estimate of $E(N_{ij})$, $j = 1, \ldots, K_i$, $i = 1, \ldots, n$. Zeger’s method could potentially underestimate the overdispersion parameter and even end up with a negative $\hat{\sigma}^2_n$ when sample size is small. If that happens, $\hat{\sigma}^2_n$ is forced to be zero. In our spline-based semiparametric projected GEE method, this overdispersion parameter is a nuisance parameter and has little impact on the consistency of the estimate of $(\beta, \Lambda)$. Due to its computational convenience, the spline-based semiparametric projected GEE method with $V_1^{(i)}$, or equivalently, the spline-based semiparametric maximum pseudolikelihood estimate studied by Lu and others (2009) is implemented to get an initial consistent estimate of $\theta = (\beta, \alpha)$, $\theta^{(0)} = (\beta^{(0)}, \alpha^{(0)})$. Then, an estimate of $\sigma^2$, $\hat{\sigma}^2_n$ is obtained using Zeger’s method in (3.1), where $\hat{\mu}_{ij} = \mu_{K_i, j}(\theta^{(0)})$. Replacing $\sigma^2$ by the estimate, $\hat{\sigma}^2_n$, the estimate of $\theta = (\beta, \alpha)$ is obtained by iteratively projecting the GEE update of

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

into the feasible space $\Theta = R^d \times J_n^+$ until convergence.

A hybrid algorithm of NR type method and IR is used to compute the spline-based projected GEE estimate. NR algorithm is used to update the parameters in the mean function first. After each NR
iteration, the projection step is made by an easy-to-implement IR (Robertson and others, 1988; Best and Chakravarti, 1990). At the current estimate \( \theta^{(k)} = (\beta^{(k)}, \alpha^{(k)}) \), denote

\[
H \left( \theta^{(k)}, \sigma_n^2 \right) = -E \left\{ V_{\theta} U \left( \theta^{(k)}, \sigma_n^2 \right) \right\} = \sum_{i=1}^{n} \left( \frac{\partial \mu^{(i)}(\theta^{(k)})}{\partial \theta} \right) V_{3}^{(i)-1}(\theta^{(k)}, \sigma_n^2) \left( \frac{\partial \mu^{(i)}(\theta^{(k)})}{\partial \theta} \right)^T
= \begin{pmatrix}
H_{\beta \beta}(\theta^{(k)}, \sigma_n^2) & H_{\beta \alpha}(\theta^{(k)}, \sigma_n^2) \\
H_{\alpha \beta}(\theta^{(k)}, \sigma_n^2) & H_{\alpha \alpha}(\theta^{(k)}, \sigma_n^2)
\end{pmatrix},
\]

the negative expectation of the derivative of the estimating function, which is the Fisher information if the underlying stochastic model is indeed Gamma–Frailty Poisson model. We choose \( W = \text{diag}(w_1, w_2, \ldots, w_q) = \text{diag} \left( H_{\alpha \alpha}(\theta^{(k)}, \sigma_n^2) \right) \) for the weight matrix in (2.2). The projection is actually the weighted IR problem and the solution has a nice interpretation: It is the left derivative of the greatest minorant of the cumulative sum diagram \( \{ P_i, i = 0, 1, \ldots, n \} \) (Groeneboom and Wellner, 1992), where \( P_0 = (0, 0) \) and \( P_i = \left( \sum_{l=1}^{i} w_l, \sum_{l=1}^{i} w_l \alpha^{(k)}_i \right) \) and can be expressed as

\[
\hat{a}_i = \max_{j < l > l} \frac{\sum_{m=j}^{l} w_m \alpha^{(k)}_m}{\sum_{m=j}^{l} w_m}.
\]

The step-by-step NR/IR algorithm tailored to the spline-based projected GEE estimation using \( V_3^{(i)} \) as the working covariance matrix is included in the Appendix B of the supplementary material available at Biostatistics online.

For the proposed method, this algorithm appears to be very efficient and converges much faster than the algorithm adopted in Lu and others (2009). Note that if the NR update from the \( k \)th iteration is within the cone depicted by the monotonicity constraint, that is, \( \alpha^{(k)} \in J_n^+ \), no projection is needed, \( \hat{\alpha}^{(k)} = \tilde{\alpha}^{(k)} \). In our numerical studies presented in Section 5, the projected step is usually only needed in the first several iterations. Cheng and others (2011) extended the result of Jongbloed (1998) to study the sufficient conditions for the numerical convergence of the projection algorithm when \( W \) is specified as the full negative Hessian matrix. In this paper, the diagonal matrix of the expectation of the Hessian matrix is chosen due to its easy geometry interpretation, and our simulation studies showed that the proposed algorithm with such chosen weight matrix \( W \) converges numerically as well.

4. INFERENCE PROCEDURE

With their coincidence to the score equations of the likelihood-based methods, that is, sieve semiparametric MPLE and MLE, the estimates based on the proposed projected GEE using \( V_1^{(i)} \) or \( V_2^{(i)} \), as working covariance matrices are consistent and the estimates of the regression parameters are asymptotically normal due to the theorem developed in Lu and others (2009). When the initial estimates of the parameters in the mean function in (3.1) are consistent, as is the case with projected GEE using \( V_1^{(i)} \), Zeger’s method of moments estimate of the overdispersion converges in probability to \( \sigma_0^2 \) that satisfies \( \text{Var}(\bar{N}(T)) = E(\bar{N}(t)) + \sigma_0^2 (E(N(t)))^2 \). The outline of the proof is included in the Appendix C of the supplementary material available at Biostatistics online. With the consistency of the estimated overdispersion parameter and the fact that GEE with \( V_3^{(i)} \) coincides with the score equation of the Gamma–Frailty Poisson model, we anticipate that the projected GEE estimation with \( V_3^{(i)} \) is also
consistent, and the estimated regression parameters are asymptotically normal with smaller standard errors compared to the corresponding estimates with $V_2^{(i)}$ when overdispersion is present in the data. However, a rigorous proof is not straightforward and requires a substantial effort with empirical process theory. The complete development of the asymptotic theorems with $V_3^{(i)}$ is our ongoing research and will be reported separately once available.

With the asymptotical normality of the estimates with $V_1^{(i)}$ or $V_2^{(i)}$, Lu and others (2009) estimated the standard errors of the estimated regression parameters using bootstrap method since the asymptotic variance matrix is in a complicated form and a direct estimation of this matrix is impossible. We also implemented the bootstrap method to estimate the standard errors of the estimated regression parameters under the projected GEE model with different variance covariance matrices. As pointed out in Section 1, with presence of severe overdispersion, bootstrap may still underestimate the variance of the estimated regression parameter. The performance of the bootstrap variance estimation when over-dispersion is neglected is further evaluated in simulation studies.

Though the estimating equation (2.1) may not hold for the estimated spline coefficients, the estimated regression parameters still satisfy the estimating equation. In light of the asymptotic normality of the proposed estimator regardless of its covariance matrix, we conjecture that the variance of the estimated regression parameter could be estimated by the corresponding components from the well-known sandwich formula given by Liang and Zeger (1986). The speculation of the variance estimate of the projected GEE estimator is "ad hoc" in that we neglect the dimension of the proposed spline-based projected GEE estimate of the baseline mean function increasing along with sample size and pretend the proposed spline-based projected GEE estimate as an ordinary parametric GEE estimate. Its performance is examined in our simulation studies included in Section 5.

5. Numerical Results

5.1 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 $X_i = (K_i, T_{K_i}, N_i, Z_i)$ 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 prescheduled at $T^* = \{T^*: T^*_j = 2j, j = 1, \ldots, 6\}$. The actual observation times $T_{ij}$ are generated from a normal distribution, $\mathcal{N}(T_{ij}^*, 1/3)$. Let $\zeta_{ij} = 1[T_{ij-1} < T_{ij}]$, for $i = 1, \ldots, 6$ and $T_{i0} = 0$. Let $\delta_{ij} = 1$ if the $j$th visit actually happens and 0 otherwise with $P(\delta_{ij} = 1) = \frac{1}{1 + e^{-\beta_0 T_{ij}}}$.

Each subject has $K_i = \sum_{j=1}^{6} \zeta_{ij} \delta_{ij}$ observations at $T_{K_i} = (T_{K_i,1}, T_{K_i,2}, \ldots, T_{K_i,K_i})$, where $T_{K_i,j}$ is the $j$th ordered observation time of $\{T_{ij}^* : \xi_{ij} \delta_{ij} = 1, j = 1, \ldots, 6\}$; (ii) the covariate vector $Z_i = (Z_{i1}, Z_{i2}, Z_{i3})$ is simulated by $Z_{i1} \sim \text{Uniform}(0, 1)$, $Z_{i2} \sim \mathcal{N}(0, 1)$, and $Z_{i3} \sim \text{Bernoulli}(0.5)$; and (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 $(Z_i, K_i, T_{K_i})$, 2 different scenarios are used to generate the panel counts $N_i = \left(\begin{array}{c} N_{i1}^{(i)}(T_{K_i,1}) \\ N_{i2}^{(i)}(T_{K_i,2}) \\ \vdots \\ N_{iK_i}^{(i)}(T_{K_i,K_i}) \end{array} \right)$.

Scenario 1. Data are generated from a Gamma–Frailty Poisson process. The frailty variables $\gamma_1, \gamma_2, \ldots, \gamma_n$ constitute a random sample from Gamma distribution, $\Gamma(0.5, 0.5)$ that results in the overdispersion parameter being 2. Conditioning on the frailty variable $\gamma_i$ as well as the covariates $Z_i$, the panel counts for each subject are drawn from a Poisson process, that is,

$$N_i^{(i)}(T_{K_i,j}) - N_i^{(i)}(T_{K_i,j-1}) \sim \text{Poisson} \left[ 2\gamma_i \left( T_{K_i,j}^{(i)} \right)^{1/2} \left( T_{K_i,j-1}^{(i)} \right)^{1/2} e^{\beta_0 Z_i} \right].$$
for \( j = 1, 2, \ldots, K_i \). 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 (1.1) and \( E(N(t)|Z) = 2t^{1/2}e^{\beta_0^T Z t} \). The counts are marginally negative binomial distributed.

Scenario 2. Data are generated from a Poisson process with the conditional mean function given by \( 2t_{ij}^{1/2}e^{\beta_0^T Z t} \), that is,

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

for \( j = 1, 2, \ldots, K_i \).

For both scenarios, the monotone cubic B-splines are used to approximate the baseline mean function in the proposed semiparametric projected 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. The same choice of \( m_n \) and the position of these knots are adopted in Lu and others (2009) and is related to the asymptotic properties of B-spline-based estimator for the baseline mean function.

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, relative efficiency compared to estimates based on \( V_3^{(i)} \) (inversed ratio of mean squared errors, relative efficiency), average of the estimated standard errors based on either the parametric GEE sandwich standard error formula (SSE) or bootstrap standard error method (BSE), and 95% coverage probability (CP1 with SSE and CP2 with BSE) for the regression parameter are summarized in Table 1. The square of the biases and the Monte Carlo standard deviation, relative efficiencies. This is expected as the working covariance matrix \( V_3 \) remains with an increased sample size. In contrast, after accounting for the data-driven overdispersion, the bootstrap provides a reasonable variance estimate and the coverage probability is close to its nominal level.

When data follow the Gamma–Frailty Poisson process as in Scenario 1, all 3 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 and the relative efficiencies. This is expected as the working covariance matrix \( V_3^{(i)} \) correctly specifies the underlying correlations among the cumulative panel counts. Neglecting overdispersion, that is, projected GEE with \( V_1^{(i)} \) or \( V_2^{(i)} \), generally leads to an underestimated variance of the estimated regression parameters, regardless of whether the ad hoc sandwich formula or the bootstrap method is used. The underestimation remains with an increased sample size. In contrast, after accounting for the data-driven overdispersion, the bootstrap provides a reasonable standard error estimate and the coverage probability is close to its nominal level. Although the sandwich formula tends to underestimate the standard error, the underestimation lessens with an increased sample size. Our additional simulations (result not shown here) indicate that when sample size increases to 600, the sandwich variance estimation, in projected GEE with \( V_3^{(i)} \), provides a reasonable variance estimate and the coverage probability is approximate to its nominal level. Figure 1 shows that the squared biases for the proposed spline-based estimates of \( \lambda_0 \) 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 3.

When data are generated from a Poisson process as in Scenario 2, 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 and others (2009). This is mainly due to the fact that the estimate of the overdispersion parameter is zero most of times in the simulation studies. Without overdispersion, the sandwich formula still underestimates the variance, while the underestimation lessens with an increased sample size. The bootstrap...
Table 1. Simulations results of the splines-based sieve semiparametric GEE estimators with 3 different covariance matrices for Gamma–Frailty Poisson Data and Poisson Process Data

<table>
<thead>
<tr>
<th>Scenario 1: Gamma–Frailty Poisson Data</th>
<th>Scenario 2: Poisson Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N = 50$</td>
<td>$N = 100$</td>
</tr>
<tr>
<td>$N = 50$</td>
<td>$N = 100$</td>
</tr>
<tr>
<td>$V_1^{(i)}$</td>
<td>$V_2^{(i)}$</td>
</tr>
<tr>
<td>$V_1^{(i)}$</td>
<td>$V_2^{(i)}$</td>
</tr>
<tr>
<td>$V_1^{(i)}$</td>
<td>$V_2^{(i)}$</td>
</tr>
<tr>
<td>$V_1^{(i)}$</td>
<td>$V_2^{(i)}$</td>
</tr>
<tr>
<td>$V_1^{(i)}$</td>
<td>$V_2^{(i)}$</td>
</tr>
<tr>
<td>$V_1^{(i)}$</td>
<td>$V_2^{(i)}$</td>
</tr>
<tr>
<td>$V_1^{(i)}$</td>
<td>$V_2^{(i)}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias</td>
<td>Bias</td>
<td>Bias</td>
</tr>
<tr>
<td>MC-sd</td>
<td>MC-sd</td>
<td>MC-sd</td>
</tr>
<tr>
<td>RE</td>
<td>RE</td>
<td>RE</td>
</tr>
<tr>
<td>SSE</td>
<td>SSE</td>
<td>SSE</td>
</tr>
<tr>
<td>BSE</td>
<td>BSE</td>
<td>BSE</td>
</tr>
<tr>
<td>CP1</td>
<td>CP1</td>
<td>CP1</td>
</tr>
<tr>
<td>CP2</td>
<td>CP2</td>
<td>CP2</td>
</tr>
</tbody>
</table>

MC-sd, Monte Carlo standard deviation; RE, relative efficiency (inversed ratio of mean square errors) compared to estimates based on $V_3^{(i)}$; 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.
Fig. 1. Simulation results for estimations of the baseline mean function, $\Lambda_0(t) = 2t^{1/2}$. 

Spline-based semiparametric projected generalized estimating equation method
method provides more reasonable standard error estimations in small sample settings, regardless of working covariance matrices used in the projected GEE.

Additional simulation studies are implemented to examine the performance of the proposed estimators in Mixture Poisson data discussed in Wellner and Zhang (2007) and Lu and others (2009) and “negative binomial counting process” discussed in Wellner and others (2004). The results can be found in the Appendix D of the supplementary material available at Biostatistics online. In summary, our 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 as efficient as the semiparametric maximum likelihood estimate when the underlying counting process is indeed a Poisson process. Accounting for data-driven overdispersion leads to less biased variance estimates. The bootstrap provides a better estimate than the \textit{ad hoc} sandwich formula in small sample settings.

5.2 Application

The proposed estimating method is applied to the bladder tumor data introduced in Section 1. A total of 116 patients were randomized into 3 treatment groups, with 31 using pyridoxine pills, 38 instilled with thiotepa, and 47 in placebo group. Their follow-up times vary from 1 to 64 weeks. Four variables, including the tumor number (\( Z_1 \)) and size (\( Z_2 \)) at baseline (study entrance), and 2 indicator variables: 1 for pyridoxine (\( Z_3 \)) and 1 for thiotepa (\( Z_4 \)), are included in the proportional mean model, that is,

\[
E(N(t)|Z_1, Z_2, Z_3, Z_4) = \Lambda_0(t) \exp(\beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4).
\]

Analysis results with the 3 different working covariance matrices are shown in Table 2. The tumor number at baseline is positively related to the recurrence of bladder tumor. With per tumor increase 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 49.8%, 45.1%, and 34.3% of that in placebo group on average using \( V_1^{(i)} \), \( V_2^{(i)} \), and \( V_3^{(i)} \), respectively.

Table 2. The spline-based sieve semiparametric inference for bladder tumor data

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SSE</th>
<th>( p ) value</th>
<th>BSE</th>
<th>( p ) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( V_1 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( Z_1 )</td>
<td>0.1444</td>
<td>0.0518</td>
<td>0.0053</td>
<td>0.0660</td>
<td>0.0286</td>
</tr>
<tr>
<td>( Z_2 )</td>
<td>-0.0447</td>
<td>0.0488</td>
<td>0.3595</td>
<td>0.0449</td>
<td>0.3189</td>
</tr>
<tr>
<td>( Z_3 )</td>
<td>0.1776</td>
<td>0.2246</td>
<td>0.4292</td>
<td>0.2894</td>
<td>0.5395</td>
</tr>
<tr>
<td>( Z_4 )</td>
<td>-0.6966</td>
<td>0.2397</td>
<td>0.0037</td>
<td>0.3250</td>
<td>0.0321</td>
</tr>
<tr>
<td>( V_2 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( Z_1 )</td>
<td>0.2075</td>
<td>0.0677</td>
<td>0.0022</td>
<td>0.0905</td>
<td>0.0218</td>
</tr>
<tr>
<td>( Z_2 )</td>
<td>-0.0353</td>
<td>0.0732</td>
<td>0.6299</td>
<td>0.0691</td>
<td>0.6098</td>
</tr>
<tr>
<td>( Z_3 )</td>
<td>0.0637</td>
<td>0.3502</td>
<td>0.8556</td>
<td>0.3891</td>
<td>0.8699</td>
</tr>
<tr>
<td>( Z_4 )</td>
<td>-0.7960</td>
<td>0.2952</td>
<td>0.0070</td>
<td>0.3780</td>
<td>0.0352</td>
</tr>
<tr>
<td>( V_3 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( Z_1 )</td>
<td>0.3289</td>
<td>0.0702</td>
<td>&lt;0.0001</td>
<td>0.0994</td>
<td>0.0009</td>
</tr>
<tr>
<td>( Z_2 )</td>
<td>0.0054</td>
<td>0.0767</td>
<td>0.9437</td>
<td>0.0809</td>
<td>0.9466</td>
</tr>
<tr>
<td>( Z_3 )</td>
<td>0.0213</td>
<td>0.4069</td>
<td>0.9583</td>
<td>0.4081</td>
<td>0.9584</td>
</tr>
<tr>
<td>( Z_4 )</td>
<td>-1.0692</td>
<td>0.3389</td>
<td>0.0016</td>
<td>0.3944</td>
<td>0.0067</td>
</tr>
</tbody>
</table>
The tumor size and pyridoxine pills are not significantly related to the number of recurrent tumors at follow-up visits. The analysis 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 analysis results based on the spline-based semiparametric pseudolikelihood and the likelihood methods proposed by Lu and others (2009). The proposed semiparametric projected GEE estimate with the frailty Poisson covariance matrix $V_3^{(i)}$ provides an estimate of the overdispersion parameter as 1.32. This is smaller than the estimate based on a nonparametric analysis from Hu and others (2009b), which may be partly explained by the inclusion of the baseline bladder tumor size and number in the semiparametric regression model. In addition, based on the bootstrap variance estimate, the estimated overdispersion parameter is significantly different from zero (standard deviation = 0.49, $p = 0.007$). It implies the presence of overdispersion in the recurrent tumor counts and possibly positive correlation among the tumor numbers in nonoverlapping time intervals for the underlying tumor progression. Accounting for overdispersion better presents these data and the new analysis considering the overdispersion provides more efficient inference for the regression parameters: The effects of the tumor number at the study entrance and the treatment of thiotepa are significantly amplified both quantitatively and statistically; the effects of nonsignificant factors are shrunk more towards zero. Figure 2 plots the estimated baseline mean functions for the control arm. Based on the estimated regression parameter for $Z_4$, the estimated cumulative counts for thiotepa arm are also plotted. Both estimates are comparable to the results shown in Hu and others (2009b). In addition, the proposed method provides a smooth estimate of the cumulative mean function. With the semiparametric mean assumption, the effect of the thiotepa instillation has a better interpretation and can be tested as shown in Table 2.

![Fig. 2. Point estimates of the baseline mean function.](image-url)
Modeling panel count data is a challenging task in general. The proposed spline-based semiparametric projected GEE method avoids assuming a stochastic model for the underlying counting process and borrows the strength from discrete observations within subjects as well as those across subjects to obtain a spline-based smooth estimate of the mean function of the underlying counting process. This approach provides a general framework for the analysis of panel count data that facilitates a selection of working covariance matrix to accommodate different data structures to improve the estimation efficiency.

The proposed spline-based projected GEE method with the working covariance matrix $V^{(i)}_3$ accounts for overdispersion and intercorrelation between nonoverlapping counts. Numerical results summarized in the manuscript shows the projected GEE with $V^{(i)}_3$ generally improves the estimation efficiency. More importantly, accounting for overdispersion is necessary to provide an unbiased inference procedure using either the \textit{ad hoc} parametric GEE sandwich formula or the bootstrap method for variance estimation when count data are overdispersed. The bootstrap method provides a more reliable variance estimation in a small sample size setting. The asymptotic properties of the proposed estimator, including its consistency, convergence rate, and asymptotic normality of the estimated regression parameter, will provide further theoretical support for the proposed method. They are currently under development and will be reported separately once available.

In our computing algorithm, $\hat{\sigma}^2_n$ is fixed with $\hat{\mu}_{ij}$ calculated from projected GEE using $V^{(i)}_1$ in the first stage, and the parameters in the proportional mean function $\theta$ are updated in the second stage. Our additional simulations show that updating $\theta$ and $\sigma^2$ alternately gives similar results. With the consistency of the mean estimates based on $V^{(i)}_1$, the estimated overdispersion parameter $\hat{\sigma}^2_n$ also converges to its limiting value in probability. In spite of the underestimation of the method of moments estimate of the overdispersion in a finite sample size setting as pointed out by Zeger (1988) and shown in our simulation studies, accounting for overdispersion still improves the estimation efficiency of the mean parameters. Different approaches on testing the overdispersion have been discussed in the parametric analysis of count data, for example, by Lawless (1987). Further research is needed to investigate their performance in the nonparametric/semiparametric analysis of panel count data.

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.

**Supplementary material**

Supplementary material is available at http://biostatistics.oxfordjournals.org.

**Acknowledgments**

We owe thanks to coeditor Professor Anastasios A. Tsiatis, an Associated Editor, and 2 anonymous referees. Their useful comments and suggestions largely helped improve this manuscript from an earlier version. \textit{Conflict of Interest:} None declared.

**References**


[Received February 18, 2011; revised June 20, 2011; accepted for publication July 28, 2011]