Robust modified profile estimating function with application to the generalized estimating equation

Molin Wang and John J. Hanfelt

Abstract

We consider methods for reducing the effects of fitting nuisance parameters on a general estimating function, when the estimating function depends on not only a vector of parameters of interest, $\theta$, but also a vector of nuisance parameters, $\lambda$. We propose a class of modified profile estimating functions with plug-in bias reduced by two orders. A robust version of the adjustment term does not require any information about the probability mechanism beyond that required by the original estimating function. An important application of this method is bias correction for the generalized estimating equation in analyzing stratified longitudinal data, where the stratum-specific intercepts are considered as fixed nuisance parameters, the dependence of the expected outcome on the covariates is of interest, and the intracluster correlation structure is unspecified. Furthermore, when the quasi-scores for $\theta$ and $\lambda$ are available, we propose an additional multiplicative adjustment term such that the modified profile estimating function is approximately information unbiased. This multiplicative adjustment term can serve as an optimal weight in the analysis of stratified studies. A brief simulation study shows that the proposed method considerably reduces the impact of the nuisance parameters.

Key words: Bias correction; Generalized estimating equation; Information unbiasedness; Neyman-Scott problem; Nuisance parameter; Profile estimating function; Sparse data
1. Introduction

Suppose \( Y = (y_1, \ldots, y_n) \) is a collection of independent random vectors whose distribution depends upon some unknown parameters including a \( p \)-dimension parameter of interest \( \theta = (\theta_1, \ldots, \theta_p) \) and a vector of nuisance parameters \( \lambda \). Assume that there exist estimating functions

\[
s(\theta, \lambda; Y) = \sum_{i=1}^{n} s_i(\theta, \lambda; y_i) \quad (1)
\]

\[
h(\theta, \lambda; Y) = \sum_{i=1}^{n} h_i(\theta, \lambda; y_i) \quad (2)
\]

for \( \theta \) and \( \lambda \), respectively, where \( s \) and \( h \) are unbiased in the sense that \( E(s) = 0 \) and \( E(h) = 0 \). An estimator of \( \theta \) can be obtained by solving the system of equations \( s = 0 \) and \( h = 0 \). This \( \theta \)-estimator is actually the root of the profile estimating function \( \hat{s} = s(\theta, \hat{\lambda}_\theta) \), where \( \hat{\lambda}_\theta \) is obtained by solving \( h(\theta, \lambda; Y) = 0 \) for given \( \theta \). In the standard asymptotic case in which the dimension of \( \theta \) and \( \lambda \) is held fixed as \( n \to \infty \), the plug-in bias, \( E(\hat{s}) \), is of order \( O(1) \). As noted by Liang & Zeger (1995) and others, in small or medium sized samples, a large plug-in bias can hamper the inference on \( \theta \). In the most severe sparse data case, in which each \( y_i \) has a distinct nuisance parameter \( \lambda_i \), the plug-in bias of \( \sum_{i=1}^{n} s_i(\theta, \hat{\lambda}_\theta; y_i) \) can cause inconsistency of the estimator of \( \theta \). This is a generalization of the Neyman and Scott (1948) problem.

In the parametric setting where the probability model is completely specified, there are many methods for reducing the impact of nuisance parameters; for example, the conditional score (Godambe, 1976; Lindsay, 1982), modified profile log-likelihood (Barndorff-Nielsen, 1980, 1983, 1986), Cox and Reid (1987)’s adjusted profile log-likelihood, McCullagh and Tibshirani (1990)’s adjusted profile
log-likelihood and Waterman and Lindsay (1996a, b)’s projected score method.

By contrast, there are few methods available in the semiparametric setting
where a full probability model is unavailable. Rathouz and Liang (1999, 2001)
present a second-order locally ancillary quasi-score, based on the projected score
method of Waterman and Lindsay (1996a, b), which achieves first-order plug-

in bias correction of a profile quasi-score. Severini (2002) develops a modified
estimating function, motivated by the Barndorff-Nielsen modified profile loglike-
lihood (Barndorff-Nielsen, 1980, 1983), which achieves a plug-in bias of \( O(n^{-1}) \) in
the standard asymptotic case. Wang and Hanfelt (2003) develop a Cox and Reid
type adjusted profile estimating function, inspired by the Cox and Reid adjusted
profile loglikelihood (Cox and Reid, 1987), that achieves the same order of plug-
in bias as Severini’s approach. Both Severini’s and Wang & Hanfelt’s methods,
however, require that the estimating functions \( s \) and \( h \) share certain properties
with score functions. Specifically, Severini’s modified profile estimating function
requires
\[
E(s^{(r)}_\lambda - h_{\theta, \lambda}) = \mathbf{0} \quad \text{and} \quad E\{h^T(s^{(r)}_\lambda - h_{\theta, \lambda})\} = \mathbf{0},
\]
for \( r = 1, \ldots, p \), where we use superscript \( (r) \) to denote the \( r \)th entry of an estimating function, and
subscript to denote derivatives. For example, \( s^{(r)}_\lambda \) and \( s^{(r)}_{\lambda \lambda} \) denote the first and
second derivatives of \( s^{(r)} \) with respect to \( \lambda \), respectively. Without further specifi-
cation, all vectors are row vectors in this paper. Alternatively, the Cox and Reid
type adjusted profile estimating function of Wang and Hanfelt (2003) requires
information unbiasedness (Lindsay, 1982) of \( h \),
\[
E(h^T h + h_\lambda) = \mathbf{0},
\]
and an estimating function analogue of the third Bartlett identity,
\[
E \left( h^T g^{(r)}_\lambda + g^{(r)}_{\lambda \lambda} \right) = \mathbf{0},
\]
where $g^{(r)} = s^{(r)} - E(s^{(r)}_X)E^{-1}(h_X)h^T$, for $r = 1, \ldots, p$. These conditions limit the general applicability of the above approaches.

Of particular interest, consider the following generalized estimating equations (Liang and Zeger, 1986 and Zeger and Liang, 1986):

$$s^T = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \theta} W_i^{-1}(y_i - \mu_i)^T = 0, \quad (5)$$

$$h^T = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \lambda} W_i^{-1}(y_i - \mu_i)^T = 0, \quad (6)$$

where $y_i = (y_{i1}, \ldots, y_{im_i})$ consists of $m_i$ correlated observations, $\mu_i = E(y_i; \theta, \lambda)$, which depends on both $\theta$ and $\lambda$, and $W_i$ is a working variance-covariance matrix of $y_i$ (Liang and Zeger, 1986; Zeger and Liang, 1986), which is not necessary equal to var$(y_i)$. Denote the $(j,k)$th entry of $W_i$ as $w_{ijk}$. Neither Rathouz and Liang’s nor Wang and Hanfelt’s method applies if $W_i \neq \text{var}(y_i)$, since in this situation condition (3) is violated. Consider a simple case in which $\mu_{ij} = E(y_{ij}) = f_i(\lambda + x_{ij}\theta)$ where $f_i$ is an arbitrary real-valued function, $\lambda$ is a scalar nuisance parameter, $x_{ij}$ is a vector of covariates, $w_{ijj}$ is a function of $\theta$ and $\lambda$ only through $\mu_{ij}$, and $w_{ijk}$ is a function of $\theta$ and $\lambda$ only through $\mu_{ij}$ and $\mu_{ik}$. Generally, the conditions required by Severini’s approach fail unless at least one of the following occur(s): (i) $W_i$ is independent of the marginal mean $\mu_i$; or (ii) $x_{ij} = x_{ik}$ for $j \neq k$; or (iii) $w_{ijk} = 0$ for $j \neq k$, that is, a “working independence” model. See the Appendix for more details. In a common situation where var$(y_{ik}) = \phi v(\mu_{ik})$, $v$ is an arbitrary known real-valued function, $\phi$ is an unknown dispersion parameter, and intracluster correlations corr$(y_{ik}, y_{ij})$ are unknown, the working variance-covariance matrix $W_i$ can often be set to $A^{1/2}_i R_i A^{1/2}_i$, where $A_i$ is a diagonal matrix with the $k$th entry given by $v(\mu_{ik})$, and $R_i$ is a working correlation.
matrix (Liang and Zeger, 1986; Zeger and Liang, 1986). It follows that Severini’s conditions typically would hold if one were to use an identity working correlation matrix, but would fail under an exchangeable or unstructured working correlation matrix.

The major goal of this article is to propose a class of modified estimating functions of form $s_m(\theta, \lambda; Y) = s(\theta, \lambda; Y) + d(\theta, \lambda; Y)$, where $d$ is an additive adjustment term and $s$ is an arbitrary unbiased estimating function defined in (1), such that $\hat{s}_m = s_m(\theta, \hat{\lambda}_\theta)$, where $\hat{\lambda}_\theta$ is the root of $h$ in (2), is biased at order $O(n^{-1})$ without additional requirements on $s$ and $h$ except mild regularity conditions. In stratified studies, with a specific nuisance parameter, $\lambda_k$ say, per stratum $k$, $k = 1, \ldots, q$, the modified estimating function would take the form $s_m = \sum_k \{s_k(\theta, \lambda_k) + d_k(\theta, \lambda_k)\}$. Here it would be advantageous to consider optimal weights for combining the stratum-specific modified estimating functions, a topic considered in Section 3.

The main results are developed in Section 2. Section 2.1 proposes a class of modified profile estimating functions, and the robust version of the adjustment term is developed in Section 2.2. In Section 2.3, we present a variance estimator for the $\theta$-estimator obtained from the proposed modified estimating functions. In Section 4 we illustrate our method using some examples and simulations. Finally, we make brief remarks in Section 5.

2. Main Results

2.1. Modified profile estimating function

For notational simplicity, throughout Section 2, we assume that $\lambda$ is a scalar. The approach can easily be extended to the case in which $\lambda$ is a vector with
fixed length. Let unbiased estimating functions $s$ and $h$ be defined in (1) and (2), respectively. Let the parameter spaces for $\theta$ and $\lambda$ be $\Omega_1$ and $\Omega_2$, respectively, $\beta = (\theta, \lambda)$, and $\Omega = \Omega_1 \times \Omega_2$. We define the following regularity conditions for $s_i(y_i; \theta, \lambda)$ and $h_i(y_i; \theta, \lambda)$, where the density function $f(y_i; \theta, \lambda)$ is defined on an abstract measurable sample space with measure $v$ for every value of the parameter $\beta \in \Omega$:

(A1) for almost all $y_i(v)$, $\partial^4 s_i / \partial \lambda^4$, $\partial^4 h_i / \partial \lambda^4$, $\partial s_i / \partial \theta$ and $\partial h_i / \partial \theta$ exist for all $\beta \in \Omega$;

(A2) $\int r^t f d\mu$, $r = s_i$ or $h_i$, is four times differentiable with respect to $\lambda$;

(A3) For all $\beta \in \Omega$, $0 < \{E_\beta(\partial h_i / \partial \lambda)\}^2 < \infty$, and the following expectations are finite: $E_\beta(\partial s_i / \partial \lambda)$, $E_\beta(s_i h_i)$, $E_\beta(\partial r / \partial \theta)$, $E_\beta(\partial r / \partial \lambda h_i)$, $E_\beta(\partial^2 r / \partial \lambda^2)$ and $E_\beta(\partial^3 r / \partial \lambda^3)$, $r = s_i$ or $h_i$.

Our first goal is to evaluate the bias of $s(\theta, \hat{\lambda}_\theta)$, which is equivalent to evaluating the bias of $g(\theta, \lambda) - g(\theta, \hat{\lambda}_\theta)$, where (Wang and Hanfelt, 2003)

$$g = s - \frac{E(s_\lambda)}{E(h_\lambda)} h.$$  

Note that $g(\theta, \hat{\lambda}_\theta) = s(\theta, \hat{\lambda}_\theta)$, $E(g) = 0$, and $E(g_\lambda) = 0$; this last orthogonality property simplifies the asymptotic bias calculation.

We use the standard result (e.g., McCullagh & Tibshirani, 1990)

$$\hat{\lambda}_\theta - \lambda = -E^{-1}(h_\lambda) h + l + O_p(n^{-3/2}),$$  

where

$$l = \left\{h_\lambda - E(h_\lambda)\right\} h \frac{E^2(h_\lambda)}{E(h_\lambda)} - \frac{E(h_{\lambda\lambda})h^2}{2E^3(h_\lambda)}.$$  

By Taylor expansion,

$$g^{(r)}(\theta, \lambda) - g^{(r)}(\theta, \hat{\lambda}_\theta) = g^{(r)}_\lambda(\theta, \lambda)(\lambda - \hat{\lambda}_\theta) - \frac{1}{2} g^{(r)}_{\lambda\lambda}(\theta, \lambda)(\lambda - \hat{\lambda}_\theta)^2.$$
and random variable. We also have $\hat{E}$

\[
\hat{E}\text{ of the above formula, and taking advantage of the property } E(\theta, \lambda) = 0, \text{ we obtain }
\]

\[
g^{(r)}(\theta, \lambda) = g^{(r)}(\theta, \lambda) - g^{(r)}(\theta, \hat{\lambda}_\theta) = d^{(r)} + u_1^{(r)} + u_2^{(r)} + O_p(n^{-1}),
\]

where

\[
d^{(r)} = \frac{E(g^{(r)}_\lambda h)}{E(h_\lambda)} - \frac{E(h^2)E(g^{(r)}_\lambda \lambda)}{2E^2(h_\lambda)},
\]

\[
u_1^{(r)} = \frac{g^{(r)}_\lambda h - E(g^{(r)}_\lambda h)}{E(h_\lambda)} - \frac{\{h^2 - E(h^2)\}E(g^{(r)}_\lambda \lambda)}{2E^2(h_\lambda)},
\]

and

\[
u_2^{(r)} = -g^{(r)}_\lambda l + \frac{E(g^{(r)}_\lambda h l h)}{6E^2(h_\lambda)} - \frac{h^2\{g^{(r)}_\lambda \lambda - E(g^{(r)}_\lambda \lambda)\}}{2E^2(h_\lambda)}.
\]

Note that $E(u_1^{(r)}) = 0$ and $u_2^{(r)} = z_1 n^{-1/2} + O_p(n^{-1})$, where $z_1$ is a zero-mean random variable. We also have $\hat{d}^{(r)} = d^{(r)} + z_2 n^{-1/2} + O_p(n^{-1})$, where $z_2$ is a zero-mean random variable. It follows that, under uniform integrability (Serfling 1980, P13), $E(\hat{s}^{(r)} + \hat{d}^{(r)}) = O(n^{-1})$. We summarize the result in the following theorem.

**Theorem 1.** For unbiased estimating functions $s$ and $h$ defined in (1) and (2), respectively, let $\hat{s}_m = \hat{s} + \hat{d}$, where $d = (d^{(1)}, \ldots, d^{(p)})$ and $d^{(r)}$ is defined by (10). Under the regularity conditions (A1-A3) and uniform integrability (Serfling 1980, P13), $E(\hat{s}_m) = O(n^{-1})$.

After plugging the expression of $g$ given by (7) into formula (10), $d^{(r)}$ can be written in terms of the original estimating functions $s$ and $h$ as

\[
d^{(r)} = \frac{E(s^{(r)}_\lambda)}{E(h_\lambda)} \left\{ \frac{E(\sum_{i=1}^n (h_{i,\lambda} h_i))}{E(h_\lambda)} - \frac{E(\sum_{i=1}^n (h_i^2))E(h_{\lambda,\lambda})}{2E^2(h_\lambda)} \right\}
\]

\[
+ \frac{E(\sum_{i=1}^n (s^{(r)}_\lambda h_i))}{E(h_\lambda)} - \frac{E(\sum_{i=1}^n (h_i^2))E(s^{(r)}_{\lambda,\lambda})}{2E^2(h_\lambda)}.
\]
Under the special conditions (3) and (4) required by Wang and Hanfelt (2003)’s approach, \( d^{(r)} \) can be simplified as \(-\frac{1}{2}E(g^{(r)}_\lambda)/E(h_\lambda)\), which coincides with the adjustment term of Wang and Hanfelt (2003)’s Cox and Reid type adjusted profile estimating function. The advantage here is that Theorem 1 holds for general choices of estimating functions \( s \) and \( h \), even if \( s \) and \( h \) fail to share properties with score functions, such as the properties specified in (3) and (4).

An attractive feature of the modified profile estimating function, \( \hat{s}_m \), is that it is invariant under reparameterization of the nuisance parameters. See the Appendix for the technical arguments.

2.2. Robust modified profile estimating function

In general, the adjustment term \( d^{(r)} \) requires knowledge of higher moments of the data than what is needed to specify the original estimating functions \( s \) and \( h \). We can obtain a robust version of the adjustment term, however, by substituting empirical estimates for the expectations that depend on higher moments, without affecting the asymptotic results.

For instance, when \( s \) and \( h \) are generalized estimating equations (5) and (6), \( E(s^{(r)}_\lambda), E(h_\lambda), E(s^{(r)}_\lambda h_\lambda) \) and \( E(h_{\lambda\lambda}) \) in \( d^{(r)} \) depend on only the first joint moment of the responses, which is also required to specify \( s \) and \( h \). However, \( E\{\sum_{i=1}^n (h_i^2)\}, E\{\sum_{i=1}^n (h_i h_i)\} \) and \( E\{\sum_{i=1}^n (s_i^{(r)} h_i)\} \) in (13) require knowledge of the second joint moments. In this case, we can use the following robust version of the adjustment term:

\[
\begin{align*}
\frac{d^{(r)}}{E(h_\lambda)} &= \frac{\sum_i (h_i g^{(r)}_\lambda)}{E(h_\lambda)} - \frac{\sum_i (h_i^2) E\left( g^{(r)}_\lambda \right)}{2E^2(h_\lambda)} \\
&= -\frac{E(s^{(r)}_\lambda)}{E(h_\lambda)} \left\{ \frac{\sum_{i=1}^n (h_i h_i)}{E(h_\lambda)} - \frac{\sum_{i=1}^n (h_i^2) E(h_{\lambda\lambda})}{2E^2(h_\lambda)} \right\}
\end{align*}
\]
\[
\sum_{i=1}^{n}(s^{(r)}_ih_i) - \frac{\sum_{i=1}^{n}(h^2_i)E(s^{(r)}_{\lambda\lambda})}{2E^2(h_\lambda)},
\]

where we have replaced those expectations depending on second moments with their empirical equivalents. If desired, one could also replace the expectations \(E(s_{\lambda}), E(h_{\lambda}), E(h_{\lambda\lambda}),\) and \(E(s_{\lambda\lambda})\) with their empirical equivalents \(s_{\lambda}, h_{\lambda}, h_{\lambda\lambda},\) and \(s_{\lambda\lambda}\), respectively, provided that the observed denominator \(h_{\lambda} \neq 0\) in (14).

Denote the modified estimating function using a robust version of the adjustment term as \(\hat{s}_{m,e}\); that is, \(\hat{s}_{m,e} = \hat{s} + \hat{d}_e\), where \(d_e = (d_e^{(1)}, \ldots, d_e^{(p)})\) and \(d_e^{(r)}\) is given by (14). We show in Appendix C that \(E(\hat{d}_e^{(r)}) = E(\hat{d}^{(r)}) + O(n^{-1})\), and thus Theorem 1 applies. It follows that \(E(\hat{s}_{m,e}) = O(n^{-1})\). Note that \(\hat{s}_{m,e}\), however, does not have the invariance property that \(\hat{s}_m\) has.

### 2.3. Precision

Denote the \(\theta\)-estimator obtained from the modified profile estimating function \(\hat{s}_m\) or \(\hat{s}_{m,e}\) as \(\hat{\theta}\). From the arguments in Sections 2.1 and 2.2, \(\hat{s}_m = g + O_p(1)\) and \(\hat{s}_{m,e} = g + O_p(1)\), where \(g\) is defined in (7). It follows that (Wang and Hanfelt, 2003)

\[
n^{1/2}(\hat{\theta} - \theta) = \left\{-\frac{1}{n}E(g_\theta^T)\right\}^{-1}(n^{-1/2}g) + O_p(n^{-1/2}),
\]

where

\[
E(g_\theta) = E(s_\theta) - E(s^T_\theta)E(h_\theta)/E(h_{\lambda});
\]

\[
\text{var}(g) = E(\sum_i s_i^T s_i) - 2E^T(s_\lambda)E(\sum_i s_i h_i)/E(h_{\lambda}) + E^T(s_\lambda)E(s_\lambda)E(\sum_i h_i^2)/E^2(h_{\lambda}).
\]

Therefore, \(\hat{\theta}\) is asymptotically distributed as normal with asymptotic variance \(E^{-1}(g_\theta)\text{var}(g)\{E^{-1}(g_\theta)\}^T\), which can be estimated at \(\theta = \hat{\theta}\) and \(\lambda = \hat{\lambda}_g\) in...
practice. A similar variance approximation can be used for the estimator obtained from the profile estimating function, \( \hat{s} \). It follows that the variances of the \( \hat{s} \)-based estimator and the \( \hat{s}_m \) (or \( \hat{s}_{m.e} \))-based estimator are asymptotically equivalent. Therefore, we do not lose much information about \( \theta \) by doing the proposed bias reduction adjustment to the profile estimating function.

3. Adjustments for Stratified Studies

Suppose there are \( q \) independent strata with stratum \( k \) consisting of \( n_k \) independent observations, denoted as \( y_k = \{y_{k1}, \ldots, y_{kn_k}\} \), where \( y_{kj} \) can be either scalar or vector. Let \( Y = \{y_1, \ldots, y_q\} \). Suppose the parameters of interest, \( \theta \), are common to all the strata, the nuisance parameters, \( \lambda_1, \ldots, \lambda_q \), are stratum-specific, and \( s(\theta, \lambda_1, \ldots, \lambda_q; Y) = \sum_k s_k(\theta, \lambda_k; y_k) = \sum_k \sum_j s_{kj}(\theta, \lambda_k; y_{kj}) \) and \( h_k(\theta, \lambda_k; y_k) = \sum_j h_{kj}(\theta, \lambda_k; y_{kj}) \) are estimating functions for \( \theta \) and \( \lambda_k \), respectively. It follows that the modified estimating function can be written as \( s_m(\theta, \lambda_1, \ldots, \lambda_q; Y) = \sum_k s_{km} = \sum_k \{s_k(\theta, \lambda_k; y_k) + d_k(\theta, \lambda_k; y_k)\} \), where \( d_k \) is the adjustment term for \( s_k \). Assuming that each \( n_k \) can be written in the form \( n_k = M_k n \), with \( A \leq M_k \leq B \) and where \( A \) and \( B \) are positive finite numbers, an appropriate asymptotic index for the stratified studies is \( q \to \infty \) and \( n \to \infty \).

Since \( \hat{s}_{km} \), defined as \( s_{km}(\theta, \hat{\lambda}_{k\theta}) \), where \( \hat{\lambda}_{k\theta} \) is the root to \( h_k = 0 \), is biased at order \( O(n^{-1}) \), by similar arguments to those in Sartori (2003), \( \hat{s}_m = \sum_k \hat{s}_{km} \) has the usual asymptotic distribution provided that \( 1/n = o(q^{-1/3}) \), while the analogous condition for a profile estimating function is \( 1/n = o(q^{-1}) \).

In this section, hereinafter we consider the important scenario in which knowledge of the first two moments of the data is available, and the stratum-specific \( s \)
and $h$ are respectively $\theta$ and $\lambda$-quasi-scores:

\begin{align*}
    s^T &= \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \theta} \text{var}^{-1}(y_i)(y_i - \mu_i)^T = 0, \quad (15) \\
    h &= \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \lambda} \text{var}^{-1}(y_i)(y_i - \mu_i)^T = 0, \quad (16)
\end{align*}

with $\mu_i = E(y_i; \theta, \lambda)$. For notational convenience, we omit the subscript $k$ denoting a stratum, $k = 1, \ldots, q$. In this scenario, Rathouz and Liang (1999, 2001)’s second-order locally ancillary quasi-score reduces the plug-in bias of the $\theta$-quasi-score by one order. Severini (2002)’s modified estimating function reduces the plug-in bias by two orders, but is applicable only if $\text{var}(y_i)$ is a function of $\theta$ and $\lambda$ only through $E(y_i; \theta, \lambda)$ (Wang and Hanfelt, 2003). Here, the proposed modified estimating function, $\hat{s}_m$, which depends on only the first two moments of the data, can be simplified to Wang and Hanfelt’s (2003) Cox and Reid type adjusted profile estimating function. None of these methods, however, typically leads to an estimating function that has an information bias of $o(1)$ as $n \to \infty$.

This is an important consideration, since the presence of information bias leads to a sub-optimally weighted estimating function for a stratified design.

Consider $g$ in (7) with $s$ and $h$ defined by (15) and (16), respectively. Since $E(g^Tg)$ can be written as

\begin{align*}
    E(s^Ts) - E(s^Th)E(s_\lambda)E(h_\lambda) - E(s_\lambda^T)E(s_\lambda)E(h_\lambda) + E(s_\lambda^T)E(s_\lambda)E(h_\lambda^2)E^2(h_\lambda),
\end{align*}

the following equalities are a set of sufficient conditions for information unbiasedness of $g$:

\begin{align*}
    E(s^Ts) + E(s_\theta) = 0, E(sh) + E(h_\theta) = 0, E(sh) - E(s_\lambda)E(h_\lambda^2)/E(h_\lambda) = 0. \quad (17)
\end{align*}
Since quasi-scores $s$ and $h$ satisfy the conditions in (17), $g$ is information unbiased; that is, $E(g^T g) + E(g_0) = 0$. It follows that both the information bias of $\hat{s}$ and $\hat{s}_m$ is $O(1)$, as $n \to \infty$.

To reduce the information bias, consider a modified profile estimating function taking the following form:

$$\hat{s}_\delta = \hat{\delta}^T \hat{s}_m^T = \hat{\delta}^T (\hat{s} + \hat{d})^T,$$

where the $r$th entry of $d$ is defined in (10) and $\delta$ is a $p \times p$ weight matrix. Our goal is to obtain the multiplicative adjustment term $\delta(\theta, \lambda)$ such that $\hat{s}_\delta$ is not only approximately unbiased but also approximately information unbiased, i.e.

$$E(d\hat{s}_\delta/d\theta) + \text{var}(\hat{s}_\delta) = o(1), \ n \to \infty.$$  \hfill (18)

A natural guess is that $\delta$ should be close to $\{(\text{var}(\hat{s} + \hat{d}))^{-1}E(-\hat{s}_\delta - \hat{d}_\delta)\}$, where the expectation and variance are, however, difficult to compute. In fact,

$$\{(\text{var}(\hat{s} + \hat{d}))^{-1}E(\hat{s}_\delta + \hat{d}_\delta) = \{E(g^T g - g^T u_2 - u_2^T g + u_1^T u_1)\}^{-1}E\{g_0 + u_{3\theta} - u_{1\theta}\} + O(n^{-2})$$

where $u_{1}^{(r)} = -u_{2}^{(r)} - d_{1}^{(r)} h/E(h_\lambda)$, $u_3 = (u_3^{(1)}, \ldots, u_3^{(p)})$, $u_2 = (u_2^{(1)}, \ldots, u_2^{(p)})$, $u_1 = (u_1^{(1)}, \ldots, u_1^{(p)})$, and $d_{1}^{(r)}$, $u_{1}^{(r)}$ and $u_{2}^{(r)}$ are defined in (10), (11) and (12), respectively. Note that both $u_{2}^{(r)}$ and $u_{3}^{(r)}$ are $O_p(n^{-1/2})$, and both $E(u_{3}^{(r)})$ and $E(u_{2}^{(r)})$ are $O(n^{-1})$. We define

$$\delta(\theta, \lambda) = -\{E(g^T g - g^T u_2 - u_2^T g + u_1^T u_1)\}^{-1}E\{g_0 + u_{3\theta} - u_{1\theta}\}. \hfill (19)$$

Since $E(g^T g) + E(g_0) = 0$, we have $\hat{\delta} = \delta + O_p(n^{-3/2})$. It follows that $E(\hat{s}_\delta) = O(n^{-1})$ and the information bias of $\hat{s}_\delta$, $E(d\hat{s}_\delta/d\theta) + \text{var}(\hat{s}_\delta)$, is typically $o(1)$, as $n \to \infty$. The technical details are presented in the Appendix. We summarize this result in the following theorem.
Theorem 2. For the $\theta$ and $\lambda$-quasi-scores $s$ and $h$ in (15) and (16), respectively, let $\hat{s}_\delta = \delta^T \hat{s}_m^T = \delta^T (\hat{s} + \hat{d})^T$, where $\delta$ is defined by (19) and $d = (d^{(1)}, \ldots, d^{(p)})$ with $d^{(r)}$ defined by (10). Under mild regularity conditions and uniform integrability, $E(\hat{s}_\delta) = O(n^{-1})$ and $E(d\hat{s}_\delta/d\theta) + \text{var}(\hat{s}_\delta) = o(1)$.

The multiplicative adjustment term $\delta$ in (19) is actually valid to achieve approximate information unbiasedness of an arbitrary modified profile estimating function in class $F$, where

$$F = \{\hat{s}_m : \hat{s}_m = \hat{s} + \hat{c}; \ E(\hat{c}) = E(\hat{d}) + O(n^{-1})\}.$$  

That is, if $\hat{s}_m$ is in $F$, the information bias of $\delta^T \hat{s}_m^T$ is $o(1)$ as $n \rightarrow \infty$. As an example, if $\text{var}(y_i)$ in (15-16) is a function of $\theta$ and $\lambda$ only through $E(y_i)$, Severini’s (2002) modified profile estimating function is in class $F$ and thus after multiplying by $\hat{\delta}$ it achieves the approximate information unbiasedness.

Note that, when computing $\delta$ in (19), after plugging in the expression of $u_1$ in (11), $E(u_1^T u_1)$ involves some fourth-moment terms, such as $v_1 E(g_\lambda^T g_\lambda h^2)$, $v_2 E(g_\lambda h^3)$ and $v_3 E(h^4)$, where $v_1 = \{E(h_\lambda)\}^{-2}$, $v_2 = -E(g_\lambda E(h_\lambda))^{-3}$ and $v_3 = \{E(g_\lambda E(h_\lambda))\}^{-4}$. Since $g_\lambda$ and $h$ are sums of independent zero-mean random variables $g_\lambda$ and $h_i$, $i = 1, \ldots, n$, respectively, $v_1 E(g_\lambda^T g_\lambda h^2)$ is equal to

$$v_1 \sum_i E(g_\lambda^T g_\lambda h_i^2) + v_1 \sum_{i \neq j} E(g_\lambda^T g_\lambda) E(h_j^2) + v_1 \sum_{i \neq j} 2E(g_\lambda^T h_i) E(g_\lambda h_j),$$

where the first term is of $O(n^{-1})$. After ignoring this lower order term, $v_1 E(g_\lambda^T g_\lambda h^2)$ can be approximated by sum of the last two terms, which depend on only the first two moments of the data. Similarly, $v_2 E(g_\lambda h^3)$ and $v_3 E(h^4)$, as well as $E(g^T u_2)$ and $E(u_{3\theta})$, can also be approximated using only the first two moments of the data, after omitting $O(n^{-1})$ terms. With these approximations, the asymptotic information unbiasedness property (18) still holds. Therefore, remarkably, the
proposed weight $\delta$ can be approximated using the knowledge of only the first two moments of the data, which is also required by the quasi-scores.

If using subscript $k$ to denote a stratum, our modified estimating function for stratified study can be written as $s^\prime_\delta = \sum_k \delta_k(s_k + d_k)$, where the multiplicative adjustment term $\delta_k$ serves as an optimal weight. By similar arguments to those presented in section 2.3, as the number of observations in each stratum goes to infinity and the number of strata is fixed, the $\theta$-estimator, $\hat{\theta}$, obtained from solving $s^\prime_\delta$, is asymptotically distributed as normal with asymptotic variance $E^{-1}(\sum_k g_k \theta)\text{var}(\sum_k g_k \theta)^T$, where $g_k = s_k - h_k E(s_k \lambda)/E(h_k \lambda)$.

4. Applications and Examples

4.1. Bias reduction to generalized estimating equations

We consider a longitudinal data study in which there are $n$ independent clusters with the correlated observations in the $i$th cluster given by $y_i = (y_{i1}, \ldots, y_{im})$. Suppose $E(y_i) = \mu_i(\theta, \lambda; x_i)$, where $\theta$ and $\lambda$ are $p$-dimension parameter of interest and $q$-dimension nuisance parameter, respectively, and $x_i$ is a vector of covariates. It follows that the generalized estimating equations for $\theta$ and $\lambda$ are given by (5) and (6), respectively. Consider a common situation for longitudinal studies in which the information about the intracluster correlation structure is unavailable. We apply the robust version of the adjustment term, $d(r)$ in (14), to reduce the plug-in bias of the profile generalized estimating function.

First, we derive the expression of $g$ in (7). We obtain $g^T = s^T - ah^T$, where

$$a = E(s_\lambda)^T E^{-1}(h_\lambda)$$

with

$$E(s_\lambda)^T = \sum_i \mu_{i\theta} W_i^{-1} \mu_{i\lambda}^T, \quad E(h_\lambda) = \sum_i \mu_{i\lambda} W_i^{-1} \mu_{i\lambda}^T.$$  (20)
Next, we obtain

$$\sum_i (h_i^T h_i) = \sum_i \mu_i \lambda_i \mathbf{W}_i^{-1} (y_i - \mu_i)^T (y_i - \mu_i) \mathbf{W}_i^{-1} \mu_i^T; \quad (21)$$

$$\sum_i (h_i^T g_i^{(r)}) = \alpha - e_i b_i^T - \beta, \quad (22)$$

where $b_i$ is the $r$th row of $a$,

$$\alpha = \sum_i \mu_i \lambda_i \mathbf{W}_i^{-1} (y_i - \mu_i)^T (y_i - \mu_i) (\mu_i \lambda_i \mathbf{W}_i^{-1} \mu_i^T),$$

and $\beta = (\beta_1, \ldots, \beta_q)$ with

$$\beta_j = \sum_i \mu_i \lambda_i \mathbf{W}_i^{-1} (y_i - \mu_i)^T (y_i - \mu_i) (\mathbf{W}_i^{-1} \mu_i^T \lambda_j) b_i^T.$$

Lastly, in (14), $E \left( g_i^{(r)} \right)$ is a $q \times q$ matrix with the $j \times k$th entry:

$$- \sum_i (\mu_i \lambda_i \mathbf{W}_i^{-1} \lambda_j) - \sum_i (\mu_i \lambda_i \mathbf{W}_i^{-1} \mu_i^T \lambda_j) \lambda_k - b_i \lambda_k (\sum_i \mu_i \lambda_i \mathbf{W}_i^{-1} \mu_i^T \lambda_j) + b_i \{ \sum_i (\mu_i \lambda_i \mathbf{W}_i^{-1} \lambda_j) \mu_i^T \lambda_k + \sum_i (\mu_i \lambda_i \mathbf{W}_i^{-1} \mu_i^T \lambda_j) \lambda_k \}.$$

(23)

Note again that, remarkably, the adjustment term requires information about only the first moment, $\mu_i$, which is also required by the generalized estimating equations (5) and (6).

**Example 1. Analysis of stratified longitudinal studies.** Suppose there are $q$ independent strata with stratum $k$ consisting of $n_k$ independent clusters. The correlated observations in the $j$th cluster of the $k$th stratum are denoted by $y_{kj1}, \ldots, y_{kjm_k}$ with $E(y_{kj}) = \mu(\lambda_k + \theta x_{kj})$ and $\text{var}(y_{kj}) = \phi v \{ E(y_{kj}) \}$, where $\theta$ is a $p$-dimensional vector of main effects, $x_{kj}$ is a $p$-dimensional vector of covariates, $\mu()$ and $v()$ are arbitrary functions, and $\phi$ is an unknown dispersion parameter. The main effects $\theta$ are of interest while the stratum-specific intercepts...
\( \lambda_k, k=1, \ldots, q, \) are treated as nuisance parameters. Note that the intracluster correlation and the third- and higher-order moments of the data remain unspecified in this example.

Generalized estimating equations (Liang & Zeger, 1986) for \( \theta \) and \( \lambda_k, k=1, \ldots, q, \) are respectively

\[
\begin{align*}
\mathbf{s}^T(\theta, \lambda) &= \sum_{k=1}^q \mathbf{s}_k^T = \sum_{k=1}^q \sum_{j=1}^{n_k} \frac{\partial E(y_{kj})}{\partial \theta} \left( A_{kj}^{1/2} R_{kj} A_{kj}^{1/2} \right)^{-1} \{y_{kj} - E(y_{kj})\}^T = \mathbf{0}, \\
h_k(\theta, \lambda_k) &= \sum_{j=1}^{n_k} \frac{\partial E(y_{kj})}{\partial \lambda_k} \left( A_{kj}^{1/2} R_{kj} A_{kj}^{1/2} \right)^{-1} \{y_{kj} - E(y_{kj})\}^T = \mathbf{0},
\end{align*}
\]

for \( k = 1, \ldots, q, \) where \( A_{kj} \) is an \( m_{kj} \times m_{kj} \) diagonal matrix with \( v\{E(y_{kjl})\} \) as the \( l \)th diagonal element and \( R_{kj} \) is a \( m_{kj} \times m_{kj} \) working correlation matrix for the \( j \)th cluster in the \( k \)th stratum (Liang and Zeger, 1986; Zeger and Liang, 1986).

Applying the robust adjusted profile estimating function, we obtain

\[
\hat{s}_{m.e} = \sum_{k=1}^q \{\mathbf{s}_k(\theta, \hat{\lambda}_{k\theta}) + \mathbf{d}_{ke}(\theta, \hat{\lambda}_{k\theta})\}.
\]

Here, \( \hat{\lambda}_{k\theta} \) is obtained by solving \( h_k = 0 \) for fixed \( \theta, \) \( \mathbf{d}_{ke} \) is a \( p \)-dimension vector with the \( r \)th entry defined by (14), where \( E(h_{\lambda}), \sum(h_i^2), \sum(h_i g_{i\lambda}^{(r)}) \) and \( E(g_{i\lambda}^{(r)}) \) are derived in (20-23) with \( y_{kj} \) replacing the notation \( y_i \) in the formulas, \( \mu_{kj} = E(y_{kj}) \) replacing \( \mu_i, \) and \( A_{kj}^{1/2} R_{kj} A_{kj}^{1/2} \) replacing \( W_i. \) Note that correct specification of the intracluster correlation matrix is not required.

We conducted a brief simulation study for a simple case where \( E(y_{kjl}) = 1/(\lambda_k + x_{kjl}\theta) \) and \( \text{var}(y_{kjl}) = \phi E^2(y_{kjl}), \) for \( k = 1, \ldots, q, j = 1, \ldots, n \) and \( l = 1, \ldots, m. \) This model has the canonical link and the mean-variance relationship of a gamma regression model. We simulated stratified longitudinal data with
$q=500$ strata, $n=4$ clusters per stratum, and each cluster size is given by $m=3$. The stratum-specific nuisance parameter $\lambda_k$ was generated from uniform$(6.0, 7.0)$ and covariate $x_{kj}$ was a scalar and generated from uniform$(3.0, 5.0)$. To show that our method is robust to higher moment assumptions, the response $y_{kj}$ was generated from multivariate normal with mean $E(y_{kj})$ and variance-covariance matrix $\phi A_{kj}^{-1/2} R_{kj} A_{kj}^{-1/2}$ with

$$R_{kj} = \begin{pmatrix}
1.0 & 0.4 & 0.3 \\
0.4 & 1.0 & 0.4 \\
0.3 & 0.4 & 1.0
\end{pmatrix},$$

and $\phi$ being 0.2 or 0.5.

Table 1 presents the simulation results for the point estimator of $\theta$ using the proposed modified profile estimating function, compared with that obtained by generalized estimating equation with different working correlation matrices: independent and exchangeable. As seen in Table 1, the estimator obtained from $s_{m,e}$ has dramatically smaller bias, and comparable standard error, than that from the generalized estimating equation. The analytical method in Section 2.3 for estimating the standard error of the estimator of $\theta$ performed well when the dispersion parameter $\phi$ is small and is biased when $\phi$ is large, in which case we can use bootstrap method to estimate the standard error.

4.2. Matched study

Rathouz and Liang (1999) and Wang and Hanfelt (2003) have applied their methods to a matched study where the nuisance parameters are stratum-specific intercepts, and there is a dispersion parameter involved in the second moment of the response variable. Both the estimating functions proposed in Rathouz and
Liang (1999) and Wang and Hanfelt (2003) depend on the dispersion parameter, which is assumed known in their examples. Here, we apply the proposed robust modified estimating function with an additive adjustment term to the same application, but assume the dispersion parameter unknown. We also consider applying a multiplicative adjustment to the estimating function with the aim of reducing information bias, using the theory developed in Section 3, under the additional assumption that the dispersion parameter is known or can be estimated precisely.

**Example 2.** Suppose there are \( q \) independent strata, and stratum \( k \) consists of \( m_k \) independently distributed observations \( y_{kj} \), \( j = 1, \ldots, m_k \), with explanatory variable \( x_{kj} \). Let the mean of \( y_{kj} \) be given by \( \mu_{kj} = (\lambda_k + \theta x_{kj})^{-1} \), with variance \( \phi \mu_{kj}^2 \), where \( \phi \) is a dispersion parameter, \( \theta \) is a vector of parameters of interest and the stratum-specific \( \lambda_k \), \( k = 1, \ldots, q \), are nuisance parameters. The \( \theta \)- and \( \lambda_k \)-quasi-scores are

\[
\begin{align*}
\mathbf{s}_q(\theta, \lambda_1, \ldots, \lambda_q) &= \sum_{k=1}^{q} \sum_{j=1}^{m_k} \frac{\partial \mu_{kj}}{\partial \theta} \frac{y_{kj} - \mu_{kj}}{\text{var}(y_{kj})} = -\sum_{k=1}^{q} \sum_{j=1}^{m_k} x_{kj} (y_{kj} - \mu_{kj}) / \phi \\
\mathbf{h}_q(\lambda_k, \theta) &= \sum_{j=1}^{m_k} \frac{\partial \mu_{kj}}{\partial \lambda_k} \frac{y_{kj} - \mu_{kj}}{\text{var}(y_{kj})} = -\sum_{j=1}^{m_k} (y_{kj} - \mu_{kj}) / \phi,
\end{align*}
\]

respectively. To avoid the unknown dispersion parameter \( \phi \) in our estimating functions, we simply set

\[
\mathbf{s}(\theta, \lambda_1, \ldots, \lambda_q) = \sum_{k=1}^{q} \sum_{j=1}^{m_k} x_{kj} (y_{kj} - \mu_{kj}),
\]

\[
\mathbf{h}_k(\lambda_k, \theta) = \sum_{j=1}^{m_k} (y_{kj} - \mu_{kj}).
\]

Using the robust adjustment term in (14), we obtain

\[
\hat{s}_{m.e} = \hat{s} + \sum_{k=1}^{q} \left\{ \left( \frac{m_k}{\sum_{j=1}^{m_k} \mu_{kj}^3} - \frac{a_k}{\sum_{k=1}^{k} \mu_{kj}^3} \right) \frac{m_k}{\sum_{j=1}^{m_k} \mu_{kj}^2} \right\} (y_{kj} - \mu_{kj})^2 |_{\lambda_k = \hat{\lambda}_k},
\]

18
where $a_k = \sum_{j=1}^{m_k} x_{kj} \mu_{kj}^2 / \sum_{j=1}^{m_k} \mu_{kj}^2$ and $\hat{\lambda}_{k\theta}$ is obtained by solving $h_k = 0$. Here $s_{m.e}$ does not involve the dispersion parameter $\phi$.

We conducted a brief simulation study. As in Rathouz and Liang (2002) and Wang and Hanfelt (2003), we considered a matched pair design where $j=1, 2$ and $x_{kj} = 1$ for $j=1$ and $x_{kj} = 0$ otherwise. We generated data from a log-normal distribution with the correctly specified mean and variance. In the simulated samples the number of strata is 100, 50 or 20, the dispersion parameter is 0.2 or 0.5, $\theta$ is 0.2 or 1.0 and the stratum-specific nuisance parameters $\lambda_k$ were set to $\lambda_0 + \lambda_{1i}$, where $\lambda_{1i}$ were generated from a beta distribution with mean $\gamma$ and variance $\psi \gamma (1 - \gamma)$, $\gamma$ and $\psi$ were set to 0.2, and $\lambda_0$ was set to 2.0. In this simulation study, the more appropriate asymptotics for this sparse data setting would be as $q \to \infty$ and $m_k$ are uniformly bounded. An aim of this simulation study is to evaluate the performance of the proposed method in the extremely sparse data situation.

Table 2 presents the comparison between the point estimators of $\theta$ using the profile quasi-score, $s$, and the modified profile estimating function, $s_{m.e}$, based on 1000 simulations. As seen in Table 2, the relative bias of the modified profile estimating function estimator is much smaller than the relative bias of the profile quasi-score estimator. As an alternative to the analytical variance in Section 2.3, we adopted a bootstrap method to estimate the variance. Specifically, $N$ strata are randomly selected from the $q$ strata to obtain an estimate. Repeat this step for $B$ times and denote the estimate obtained using the $b$th resample as $\hat{\theta}(b)$; the standard error of $\hat{\theta}$ can be estimated by the following formula:

$$
\hat{se} = \left[ \frac{\sum_{b=1}^{B} (\hat{\theta}(b) - \hat{\theta}(.))^2}{B - 1} \right]^{1/2},
$$
where \( \hat{\theta}(.) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}(b) \).

We also conducted a simulation study to test the performance of a robust version of the modified profile estimating function, \( \hat{s}_\delta \), in Section 3 where we assumed that the dispersion parameter \( \phi \) is known and used only mean and variance of \( y_{kj} \). The stratum-specific nuisance parameters \( \lambda_k \) were generated from a beta distribution with mean \( \gamma \) and variance \( \psi \gamma (1 - \gamma) \) where \( \gamma \) and \( \psi \) were set to 0.2. The independent variable \( x_{kj} \) was simulated from Uniform(7, 9). We set \( \phi = 0.6 \), the number of stratum, \( q \), 50 or 100, and the number of observations in each stratum, \( m \), 8 or 15. Shown in Table 3 is comparison among point estimators of \( \theta \) using the profile quasi-score, the robust modified profile estimating function \( \hat{s}_{m.e} \), and the robust modified profile estimating function \( \hat{s}_\delta \) based on 1000 simulations. Again, the relative biases of the two modified profile estimating function estimators are much smaller than the relative bias of the profile quasi-score estimator. The mean square error of the \( \hat{s}_\delta \) estimator is consistently smaller than the \( \hat{s}_{m.e} \) estimator; the reduction in mean square error was not very substantial, however, as seen in Table 3.

5. Concluding Remarks

This paper proposes an additive adjustment term for second-order bias reduction of profile estimating functions. There are two key advantages of the proposed approach: (i) it does not require that the original estimating functions share special properties with score functions, such as those properties mentioned in Section 1 for the Severini (2002) and Wang and Hanfelt (2003) approaches; and (ii) the robust modified profile estimating function does not require any information about the probability mechanism that generated the data, except what is
needed to specify the original estimating functions. For example, when applied in a
generalized estimating equations context, the proposed method requires correct
specification of only the regression model for the marginal univariate means.

In addition, in the context of a stratified study, this paper proposes a mul-
tiplicative adjustment term to achieve information bias reduction when the es-
timating functions for the parameters of interest and the nuisance parameters
are quasi-scores. This adjustment term serves as an optimal weight in stratified
studies. Remarkably, this multiplicative adjustment term can be implemented
using only the specification of first two moments of the observation, which is also
required by the quasi-scores.

Appendix A. Conditions under which the method of Severini (2002)
applies to GEE.

For simplicity, assume both $\theta$ and $\lambda$ are scalars. It follows from (5) and (6)
that $s_{\lambda} - h_{\theta} = \sum_i r_i (y_i - \mu_i)^T$, where

$$r_i = \frac{\partial \mu_i}{\partial \theta} \frac{\partial W_i^{-1}}{\partial \lambda} - \frac{\partial \mu_i}{\partial \lambda} \frac{\partial W_i^{-1}}{\partial \theta}.$$  

The conditions required by Severini (2002) are: (i) $E(s_{\lambda\lambda} - h_{\theta\lambda}) = \sum_i r_i (-\partial \mu_i / \partial \lambda)^T = 0$, (ii) $E\{h(s_{\lambda} - h_{\theta})\} = \sum_i \partial \mu_i / \partial \lambda W_i^{-1} \text{var}(y_i) r_i^T = 0$. If $W_i = \text{var}(y_i)$, these two
conditions are identical. Generally, conditions (i) and (ii) do not hold unless
$r_i = 0$. A trivial case in which $r_i = 0$ is when $W_i$ does not depend on $\theta$ and $\lambda$.

Assume the size of each cluster is 2. Denote the $jk$th entry of $W_i^{-1}$ as $a_{ijk}$, $j = 1, 2$, $k = 1, 2$. Letting $r_i = (r_{i1}, r_{i2})$, $r_i = 0$ leads to

$$r_{ik} = \frac{\partial \mu_{i1}}{\partial \theta} \frac{\partial a_{11k}}{\partial \lambda} + \frac{\partial \mu_{i2}}{\partial \theta} \frac{\partial a_{22k}}{\partial \lambda} - \frac{\partial \mu_{i1}}{\partial \lambda} \frac{\partial a_{11k}}{\partial \theta} - \frac{\partial \mu_{i2}}{\partial \lambda} \frac{\partial a_{22k}}{\partial \theta} = 0.$$
for $k = 1, 2$. In the simple case where $w_{ijk}(\theta, \lambda) = w_{ijk}(\mu_{i1}(\theta, \lambda), \mu_{i2}(\theta, \lambda))$, for $j \neq k$; and $w_{ijj}(\theta, \lambda) = w_{ijj}(\mu_j(\theta, \lambda))$, we obtain

$$r_{ik} = \left( \frac{\partial a_{i1k}}{\partial \mu_{i2}} - \frac{\partial a_{i2k}}{\partial \mu_{i1}} \right) \left( \frac{\partial \mu_{i1}}{\partial \theta} \frac{\partial \mu_{i2}}{\partial \lambda} - \frac{\partial \mu_{i2}}{\partial \theta} \frac{\partial \mu_{i1}}{\partial \lambda} \right).$$

Given $\mu_{ik} = f_i(\lambda + x_{ik}\theta)$, the term in the second bracket is non-zero unless $x_{i1} = x_{i2}$. The term in the first bracket is generally not zero unless $w_{ijk} = 0$ for $j \neq k$, in which case $\partial a_{i1k}/\partial \mu_{i2} = \partial a_{i2k}/\partial \mu_{i1} = 0$. We have similar results if the size of cluster $m_i$ exceeds 2.

Appendix B. Technical arguments for invariance of $\hat{s}_m$.

This proof is similar to the proof for invariance in Wang and Hanfelt (2003). Consider a reparameterization of the nuisance parameter, such as $(\theta, \lambda) \mapsto (\theta, \phi)$, where $\phi = \phi(\theta, \lambda)$ and $|\phi_\lambda| \neq 0$. Here $\theta$ is held fixed. For notationally simplicity, we assume $\theta$ and $\lambda$ are scalars. Let $h_1(\theta, \lambda)$ be the estimating function for $\lambda$ under parameterization $(\theta, \lambda)$ and let $h_2(\theta, \phi)$ refer to the same estimating function after taking the reparameterization $\phi = \phi(\theta, \lambda)$. Let $g_1$, $g_2$, $d_1$ and $d_2$ be defined similarly. Suppose $g_1(\theta, \lambda) = g_2(\theta, \phi)$ and $h_1(\theta, \lambda) = h_2(\theta, \phi)\phi_\lambda$, as would occur if the estimating functions are generalized estimating functions in (5) and (6). We have $g_{1\lambda} = g_{2\phi}\phi_\lambda$ and $g_{1\lambda\lambda} = g_{2\phi\phi}\phi_\lambda^2 + g_{2\phi}\phi_\lambda$. It follows that $d_1 = d_2$. Hence, $\hat{s}_m$ is invariant under this reparameterization.

Appendix C. Technical arguments for the main result in Section 2.2.

Suppose the regularity conditions (A1-A3) are satisfied, and all the involved expectations in the arguments are finite. Assume that the uniform integrability (Serfling 1980, P13) required in the arguments is satisfied. When replacing the
Applying similar arguments to the other terms in \( d \), we prove expectations \( E(s_\lambda), \ E(h_\lambda), \ E(h_{\lambda\lambda}), \) and \( E(s_{\lambda\lambda}) \) with their empirical equivalents \( s_\lambda, h_\lambda, h_{\lambda\lambda}, \) and \( s_{\lambda\lambda}, d_{e}^{(r)} \) can be written as

\[
d_{e}^{(r)} = -\frac{s_\lambda^{(r)}}{h_\lambda} \left\{ \sum_{i=1}^{n}(h_{i\lambda}h_i) - \frac{\sum_{i=1}^{n}(h_i^2)h_{\lambda\lambda}}{2h_\lambda^3} \right\} + \frac{\sum_{i=1}^{n}(s_{i\lambda}^{(r)}h_i)}{h_\lambda} - \frac{\sum_{i=1}^{n}(h_i^2)s_{\lambda\lambda}^{(r)}}{2h_\lambda^3}.
\]

We prove \( E(\hat{d}_{e2}^{(r)}) = E(d^{(r)}) + O(n^{-1}) \). The result for \( \hat{d}_{e}^{(r)} \) can be proved using similar arguments. We first show \( E(d_{e2}^{(r)}) - d^{(r)} = O(n^{-1}) \). Note that

\[
h_\lambda^{-m} = E^{-m}(h_\lambda)[1 - m\{h_\lambda - E(h_\lambda)\}/E(h_\lambda) + O_p(n^{-1})],
\]

for \( m = 1, \ldots, 4 \). It follows that expectation of the second term in \( d_{e2}^{(r)} \) can be written as

\[
-\sum_{i=1}^{n}(h_{i\lambda}h_i)/h_\lambda^2
\]

\[
= -E^{2}(h_\lambda)E\{s_\lambda^{(r)} \sum_{i=1}^{n}(h_{i\lambda}h_i)\} + 2E^{3}(h_\lambda)E[s_\lambda^{(r)} \sum_{i=1}^{n}(h_{i\lambda}h_i)\{h_\lambda - E(h_\lambda)\}] + O(n^{-2})
\]

\[
= -E^{2}(h_\lambda)E\{s_\lambda^{(r)} \sum_{i=1}^{n}(h_{i\lambda}h_i)\} + O(n^{-1})
\]

\[
= -E^{2}(h_\lambda)E[s_\lambda^{(r)} \sum_{i=1}^{n}(h_{i\lambda}h_i)\} - E^{2}(h_\lambda)E(s_{\lambda\lambda}^{(r)})E[s_\lambda^{(r)} \sum_{i=1}^{n}(h_{i\lambda}h_i)\} + O(n^{-1})
\]

\[
= -E^{2}(h_\lambda)E(s_{\lambda\lambda}^{(r)})E[s_\lambda^{(r)} \sum_{i=1}^{n}(h_{i\lambda}h_i)\} + O(n^{-1}).
\]

Applying similar arguments to the other terms in \( d_{e2}^{(r)} \), we have \( E(d_{e2}^{(r)}) = O(n^{-1}) \). Next, we show \( E(\hat{d}_{e2}^{(r)}) = E(d^{(r)}) + O(n^{-1}) \). We have

\[
E(\hat{d}_{e2}^{(r)}) = E(d_{e2}^{(r)}) - E(h_\lambda d_{e2}^{(r)}/\partial\lambda)/E(h_\lambda) + O(n^{-1})
\]

\[
= d^{(r)} - E(h_\lambda d_{e2}^{(r)}/\partial\lambda)/E(h_\lambda) + O(n^{-1})
\]

\[
= E(\hat{d}^{(r)}) - E(h_\lambda d_{e2}^{(r)}/\partial\lambda)/E(h_\lambda) + O(n^{-1}),
\]

23
where the first equality follows from Taylor expansion. In order to show $E(\hat{d}^{(r)}_2) = E(\hat{d}^{(r)}) + O(n^{-1})$, we need to prove $E(h\partial d^{(r)}_2 / \partial \lambda) = O(1)$. We have

$$h \frac{\partial d^{(r)}_2}{\partial \lambda} = \frac{h \sum_{i=1}^{n} (s^{(r)}_{i\lambda} h_i)}{h \lambda} + \frac{h \sum_{i=1}^{n} (s^{(r)}_{i\lambda}) h_{i\lambda}}{h \lambda} - \frac{h \sum_{i=1}^{n} (s^{(r)}_{i\lambda}) h_{i\lambda}}{h \lambda}$$

$$- \frac{hs^{(r)}_{\lambda\lambda} \sum_{i=1}^{n} (h_{i\lambda} h_i)}{h \lambda} - \frac{hs^{(r)}_{\lambda\lambda} \sum_{i=1}^{n} (h_{i\lambda} h_i)}{h \lambda} + 2hs^{(r)}_{\lambda\lambda} \sum_{i=1}^{n} (h_{i\lambda} h_i) h_{\lambda\lambda}$$

$$- \frac{h \sum_{i=1}^{n} (2h_i h_{i\lambda}) s^{(r)}_{\lambda\lambda}}{2h \lambda} - \frac{h \sum_{i=1}^{n} (h_i^2) s^{(r)}_{\lambda\lambda}}{2h \lambda} + \frac{h \sum_{i=1}^{n} (h_i^2) s^{(r)}_{i\lambda\lambda} h_{\lambda\lambda}}{2h \lambda}$$

$$+ \frac{hs^{(r)}_{\lambda\lambda} \sum_{i=1}^{n} (h_i^2) h_{\lambda\lambda}}{2h \lambda} + \frac{hs^{(r)}_{\lambda\lambda} \sum_{i=1}^{n} (2h_i h_{i\lambda}) h_{\lambda\lambda}}{2h \lambda} + \frac{hs^{(r)}_{\lambda\lambda} \sum_{i=1}^{n} (h_i^2) h_{\lambda\lambda}}{2h \lambda} - \frac{3hs^{(r)}_{\lambda\lambda} \sum_{i=1}^{n} (h_i^2) h_{\lambda\lambda}}{2h \lambda}.$$

By (24), we obtain

$$E\left\{ \frac{h \sum_{i=1}^{n} (s^{(r)}_{i\lambda} h_i)}{h \lambda} \right\} = \frac{E \left[ h \sum_{i=1}^{n} (s^{(r)}_{i\lambda} h_i) [1 - \{ h_{i\lambda} - E(h_{i\lambda}) \} / E(h_{i\lambda}) + O_p(n^{-1})] \right]}{E(h_{i\lambda})} = O(1)$$

and

$$E\left\{ \frac{h \sum_{i=1}^{n} (s^{(r)}_{i\lambda} h_i) h_{\lambda\lambda}}{h \lambda} \right\} = \frac{E \left[ h \sum_{i=1}^{n} (s^{(r)}_{i\lambda} h_i) h_{\lambda\lambda} [1 - 2 \{ h_{i\lambda} - E(h_{i\lambda}) \} / E(h_{i\lambda}) + O_p(n^{-1})] \right]}{E^2(h_{i\lambda})} = O(1)$$

By similar arguments, the expectation of each other term in $h\partial d^{(r)}_2 / \partial \lambda$ is also $O(1)$. Therefore $E(h\partial d^{(r)}_2 / \partial \lambda) = O(1)$.

**Appendix D. Technical arguments for the main result in Section 3.**

Suppose the regularity conditions (A1-A3) are satisfied, and all the involved expectations in the arguments are finite. Assume that the uniform integrability (Serfling 1980, P13) required in the arguments is satisfied. By (9) and since $\hat{d}^{(r)} - d^{(r)} = -d^{(r)}_\lambda h / E(h_{i\lambda}) + O_p(n^{-1})$, we have $s^{(r)} + \hat{d}^{(r)} = g^{(r)}(\theta, \lambda) - u^{(r)}_1 + u^{(r)}_3 + O_p(n^{-1})$, where $u^{(r)}_3 = -u^{(r)}_3 - d^{(r)}_\lambda h / E(h_{i\lambda}) = O_p(n^{-1/2})$ and $E(u^{(r)}_3) = O(1/n)$.

Since $E(g^T u_1) = O(1/n)$, we have

$$\text{var}(\hat{s} + \hat{d}) = E\{(\hat{s} + \hat{d})^T(\hat{s} + \hat{d})\} + O(n^{-2})$$

$$= E(g^T g + g^T u_3 + u_3^T g + u_1^T u_1) + O(n^{-1}) = E(g^T g) + O(1),$$
and
\[ E(\hat{s}_\theta + \hat{d}_\theta) = E(g_\theta + u_{3\theta} - u_{1\theta}) + O(n^{-1}) \]
\[ = E(g_\theta) + O(1). \]

It follows that \( \delta = t + O(n^{-2}) \), where
\[ t = -\{\text{var}(\hat{s} + \hat{c})\}^{-1}E\left(\frac{d\hat{s}}{d\theta} + \frac{d\hat{c}}{d\theta}\right). \]

If \( E(g^T g) + E(g_\theta) = 0 \), we have
\[ \delta = I_p - E(g^T g + g^T u_3 + u_3^T g + u_1^T u_1)^{-1}E(g^T u_3 + u_3^T g + u_1^T u_1 + u_{3\theta} - u_{1\theta}), \]
where \( I_p \) is a \( p \times p \) identity matrix. It follows that \( \hat{\delta} = \delta + O_p(n^{-3/2}) \), and
\[ t = \delta + O_p(n^{-2}) = \hat{\delta} + O_p(n^{-3/2}). \] Therefore
\[ E(d\hat{s}_\delta/d\theta) + \text{var}(\hat{s}_\delta) = t^T E\{(\hat{s} + \hat{d})^T(\hat{s} + \hat{d})\}t + t^T E\left\{\left(\frac{d\hat{s}}{d\theta} + \frac{d\hat{d}}{d\theta}\right)\right\} + o(1) = o(1), \ n \to \infty. \]

Also, if \( s \) and \( h \) are \( \theta \) and \( \lambda \)-quasi-scores, respectively,
\[ E(g^T u_3) = -E(g^T u_2). \]

References


Table 1: Simulation results for stratified longitudinal data with $q = 500$, $n = 4$, $m = 3$ and $\lambda_k \sim \text{Un}(6, 7)$ based on 1000 replicates

<table>
<thead>
<tr>
<th>Models</th>
<th>Modified GEE</th>
<th>Profile GEE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\phi$</td>
<td>$\theta$</td>
</tr>
<tr>
<td>Independent working correlation matrix</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5 2.0</td>
<td>-1.3</td>
<td>0.234</td>
</tr>
<tr>
<td>1.0 2.0</td>
<td>-2.0</td>
<td>0.170</td>
</tr>
<tr>
<td>0.5 -1.8</td>
<td>-1.8</td>
<td>0.130</td>
</tr>
<tr>
<td>0.2 2.0</td>
<td>-0.5</td>
<td>0.143</td>
</tr>
<tr>
<td>1.0 -1.1</td>
<td>-1.1</td>
<td>0.106</td>
</tr>
<tr>
<td>0.5 -1.3</td>
<td>-1.3</td>
<td>0.083</td>
</tr>
<tr>
<td>Exchangeable working correlation matrix</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5 2.0</td>
<td>0.1</td>
<td>0.253</td>
</tr>
<tr>
<td>1.0 -1.6</td>
<td>-1.6</td>
<td>0.167</td>
</tr>
<tr>
<td>0.5 -2.0</td>
<td>-2.0</td>
<td>0.124</td>
</tr>
<tr>
<td>0.2 2.0</td>
<td>-0.4</td>
<td>0.141</td>
</tr>
<tr>
<td>1.0 -1.0</td>
<td>-1.0</td>
<td>0.104</td>
</tr>
<tr>
<td>0.5 -1.3</td>
<td>-1.3</td>
<td>0.081</td>
</tr>
</tbody>
</table>

%bias, relative bias calculated by the mean difference between the $\theta$-estimator and the true $\theta$ relative to the true $\theta$; SE$_e$, empirical standard error; SE$_a$, average of analytical standard error.
Table 2: Simulation results for matched-pairs data based on 1000 replicates

<table>
<thead>
<tr>
<th>Model</th>
<th>Modified PEF</th>
<th>Profile quasi-score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%bias</td>
<td>SE</td>
</tr>
<tr>
<td>100</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>-11.4</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>-5.7</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>-5.3</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>-9.6</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>-7.2</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>-3.9</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>-9.3</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>-3.2</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>-4.3</td>
</tr>
</tbody>
</table>

%bias, relative bias calculated by the mean difference between the \( \hat{\theta} \)-estimator and the true \( \theta \) relative to the true \( \theta \); SE, empirical standard error; SEa, average of bootstrap standard error with \( B=100, N=95, 48 \) and \( 18 \) for \( q = 100, 50, 20, \) respectively.
### Table 3: Simulation results for matched study based on 1000 replicates

<table>
<thead>
<tr>
<th>Model</th>
<th>Modified PEF ( \hat{s}_q )</th>
<th>Modified PEF ( \hat{s}_{m,e} )</th>
<th>Profile quasi-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q )</td>
<td>( m )</td>
<td>( \theta )</td>
<td>%bias</td>
</tr>
<tr>
<td>50</td>
<td>8</td>
<td>5.0</td>
<td>-1.053</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>5.0</td>
<td>0.893</td>
</tr>
<tr>
<td></td>
<td>10.0</td>
<td>-0.992</td>
<td>3.146</td>
</tr>
<tr>
<td>100</td>
<td>8</td>
<td>5.0</td>
<td>-1.754</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>5.0</td>
<td>0.749</td>
</tr>
<tr>
<td></td>
<td>10.0</td>
<td>0.435</td>
<td>2.149</td>
</tr>
</tbody>
</table>

%bias, relative bias calculated by the mean difference between the \( \theta \)-estimator and the true \( \theta \) relative to the true \( \theta \); SE, empirical standard error; MSE, mean square error.