Brazilian Journal of Probability and Statistics
2014, Vol. 28, No. 4, 561–586
DOI: 10.1214/13-BJPS224

Estimation with improved efficiency in semi-parametric linear longitudinal models

Vineetha Warriyar K. V. and Brajendra C. Sutradhar

Memorial University

Abstract. In this article, we revisit the semi-parametric linear models with auto-correlated errors, where the means of the repeated responses of an individual consist of a specified regression function in time dependent covariates as well as a time dependent nonparametric function. The estimation of the regression parameters involved in the specified regression function is of main interest, and most of the existing studies estimate these parameters by using the so-called semi-parametric generalized estimating equations (SGEEs) approach. We offer two main contributions. First, we demonstrate that the existing SGEEs are partly standardized. Second, as opposed to this partly standardized SGEE (PSSGEE) approach, we suggest a fully standardized semi-parametric generalized quasi-likelihood (FSSGQL) approach that provides more efficient regression estimates. This efficiency gain by the FSSGQL approach over the PSSGEE approach is also demonstrated through an empirical study.

1 Introduction

Let t_{ij} denote the time at which the jth ($j = 1, \ldots, n_i$) measurement is made on the ith ($i = 1, \ldots, K$) individual, and y_{ij} denote this measurement which is also referred to as the jth response of the ith individual at time t_{ij}. Next, suppose that $y_i = (y_{i1}, \ldots, y_{ij}, \ldots, y_{in_i})'$ denotes the $n_i \times 1$ vector of repeated responses for the ith ($i = 1, \ldots, K$) individual. Also suppose that these repeated responses are influenced by a smooth nonparametric function $\gamma(t_{ij})$, and a fixed and known $p \times n_i$ covariate matrix $X_i' = (x_i(t_{i1}), \ldots, x_i(t_{ij}), \ldots, x_i(t_{in_i}))$, $x_i(t_{ij})$ being the p-dimensional covariate vector for the ith individual at time point t_{ij}. This type of repeated continuous data measured at time point t_{ij} are usually modeled as

$$\begin{aligned} y_{ij} &= x_i'(t_{ij})\beta + \gamma(t_{ij}) + \epsilon_{ij}(t_{ij}) \\ &= \mu_{ij}(t_{ij}) + \epsilon_{ij}(t_{ij}), \end{aligned} \tag{1.1}$$

or equivalently

$$y_i = X_i\beta + \gamma_i + \epsilon_i, \tag{1.2}$$

Key words and phrases. Auto-correlated errors, improved efficiency, kernel based GQL estimation, nonparametric function, semi-parametric regression model.

Received August 2012; accepted May 2013.

where

$$\gamma_i = \big(\gamma(t_{i1}), \ldots, \gamma(t_{in_i})\big)' \quad \text{and} \quad \epsilon_i = \big(\epsilon_{i1}(t_{i1}), \ldots, \epsilon_{ij}(t_{ij}), \ldots, \epsilon_{in_i}(t_{in_i})\big)'.$$

Note that in (1.2), γ_i is not a subject specific nonparametric function as its construction requires only knowing $\gamma(t)$ at any time t (see Zeger and Diggle (1994) and Sneddon and Sutradhar (2004), e.g.). To be specific, γ_i is used here to represent n_i components, each with the same nonparametric function but evaluated at n_i different time points for the ith individual. Further note that in this longitudinal setup, the components of the error vector ϵ_i must be correlated. But as the correlation structure is unknown in practice, many authors such as Zeger and Diggle (1994) considered that

$$\epsilon_i \sim \big(0, \sigma^2 R_i(\alpha)\big), \tag{1.3}$$

where $\sigma^2 = \text{var}[\epsilon_{ij}(t_{ij})] = \sigma_{ijj}(t_{ij}) = \sigma^2(t_{ij})$, and $R_i(\alpha)$ is a "working" correlation matrix used for unknown true correlation matrix. The commonly used $R_i(\alpha)$ are: (a) the unstructured form $R_i(\alpha) = (r_{i,jk}(\alpha))$ with $r_{i,jk}(\alpha) = \alpha_{|t_{ij}-t_{ik}|}$ (Zeger and Diggle (1994), Lin and Carroll (2001)); (b) equi-correlations form $R_i(\alpha) = \alpha I_{n_i}$, and (c) independence form $R_i(\alpha) = I_{n_i}$ (Lin and Carroll (2001), Severini and Staniswalis (1994)). It then follows from the models (1.1)–(1.3) that the mean response is given by

$$E[Y_{ij}] = \mu_{ij}(t_{ij}) = x_i'(t_{ij})\beta + \gamma(t_{ij}), \tag{1.4}$$

where β is the fixed regression effects, and $\gamma(t_{ij})$ is a nonparametric smooth function of times. It is of main interest to estimate the fixed regression effects β consistently and as efficiently as possible. Note however that this regression estimation requires the consistent estimation of the other secondary functions and parameters, namely the nonparametric smooth function $\gamma(t_{ij})$, and the "working" correlation matrix $R_i(\alpha)$.

Remark that even though β and $\gamma(t)$ together constitute the regression function (1.4), their joint estimation based on cluster correlated data may be difficult. When correlations are ignored, the estimation becomes easier whether using the so-called local polynomial or spline-based methods (Hua (2010)). Thus, in the existing literature dealing with clustered correlated data, they are estimated marginally by using separate estimating equations (Zeger and Diggle (1994), Severini and Staniswallis (1994), and Lin and Carroll (2001)). This makes it simpler, for example, to use "working" independence approach for consistent estimation of $\gamma(t)$ (Zeger and Diggle (1994, Section 3.1)), and a suitable correlation structure based approach for efficient estimation of the main regression parameter β. As far as the con-

sistent estimation of $\gamma(t)$ is concerned, because the $\gamma(t)$ in the mean function (1.4) is a nonparametric function, a kernel approach is used for such an estimation. More specifically, a "working" independence assumption based unbiased estimating function is weighted by using suitable kernel weights and the resulting semi-parametric estimating equation is then solved for $\gamma(t)$. At a given time point t_0, say, under the present linear model with mean regression function as in (1.4), the semi-parametric quasi-likelihood (SQL) estimating equation for $\gamma(t_0)$ has the form

$$\sum_{h=1}^{K}\sum_{u=1}^{n_h} w_{hu}(t_0)\frac{\partial \mu_{hu}(t_{hu})}{\partial \gamma(t_0)}\frac{(y_{hu}-\mu_{hu}(t_{hu}))}{\sigma^2}=0, \tag{1.5}$$

where $w_{hu}(t_0) = p_{hu}(\frac{t_0-t_{hu}}{b})/(\sum_{h=1}^{K}\sum_{u=1}^{n_h} p_{hu}(\frac{t_0-t_{hu}}{b}))$, $p_{hu}(\cdot)$ being a suitable kernel for example, we choose $p_{hu}(\frac{t_0-t_{hu}}{b}) = \frac{1}{\sqrt{2\pi}b}\exp(\frac{-1}{2}(\frac{t_0-t_{hu}}{b})^2)$ with a suitable bandwidth b.

Next, because $\mu_{ij}(t_{ij}) = x_i'(t_{ij})\beta + \gamma(t_{ij})$ by (1.4), it is convenient to express the SQL estimating equation for $\gamma(t_{ij})$ (1.5), in terms of known β, as

$$\hat{\gamma}(t_{ij}) = \hat{y}_{ij} - \hat{x}_i'(t_{ij})\beta, \tag{1.6}$$

where

$$\hat{y}_{ij} = \sum_{h=1}^{K}\sum_{u=1}^{n_h} w_{hu}(t_{ij})y_{hu} \quad \text{and} \quad \hat{x}_i'(t_{ij}) = \sum_{h=1}^{K}\sum_{u=1}^{n_h} w_{hu}(t_{ij})x_h'(t_{hu})$$

with $\sum_{h=1}^{K}\sum_{u=1}^{n_h} w_{hu}(t_{ij}) = 1$. Now by using the formulas for $\hat{y}_{ij}$ and $\hat{x}_i'(t_{ij})$, one writes

$$\begin{aligned}\hat{y}_i &= \sum_{h=1}^{K}\big[W_h(t_{i1},\ldots,t_{in_i})\big]y_h,\\ \hat{X}_i &= \sum_{h=1}^{K}\big[W_h(t_{i1},\ldots,t_{in_i})\big]X_h\end{aligned} \tag{1.7}$$

with $W_h(t_{i1},\ldots,t_{in_i})$ as the kernel weights matrix defined as

$$W_h(t_{i1},\ldots,t_{in_i}) = \begin{pmatrix} w_h'(t_{i1})\\ \vdots\\ w_h'(t_{ij})\\ \vdots\\ w_h'(t_{in_i})\end{pmatrix} : n_i \times n_h, \tag{1.8}$$

where $w_h'(t_{ij}) = [w_{h1}(t_{ij}), \ldots, w_{hu}(t_{ij}), \ldots, w_{hn_h}(t_{ij})]$ with $w_{hu}(t)$ at a given time t as given in (1.5), and

$$X_h = \begin{pmatrix} x_h'(t_{h1}) \\ \vdots \\ x_h'(t_{hu}) \\ \vdots \\ x_h'(t_{hn_h}) \end{pmatrix} : n_h \times p. \tag{1.9}$$

The existing studies, such as Severini and Staniswalis (1994, equations (17) and (18)) and You and Chen (2007, Section 4.1) (see also Lin and Carroll (2001)), then obtained the "working" generalized least squared (WGLS) estimator for β as

$$\begin{aligned} \hat{\beta}_{\text{WGLS}} = &\left[\sum_{i=1}^{K}(X_i - \hat{X}_i)'\{\text{var}(Y_i)\}^{-1}(X_i - \hat{X}_i)\right]^{-1} \\ &\times \sum_{i=1}^{K}(X_i - \hat{X}_i)'\{\text{var}(Y_i)\}^{-1}(y_i - \hat{y}_i), \end{aligned} \tag{1.10}$$

with $\text{var}(Y_i) = A_i^{-1/2} R_i(\alpha) A_i^{-1/2}$, $R_i(\alpha)$ being a "working" correlation matrix.

Note that the WGLS estimator in (1.10) may also be referred to as the "working" generalized estimating equations (WGEE) based estimator which uses the so-called "working" correlation matrix $R_i(\alpha)$. This "working" parameter α used to define $R_i(\alpha)$ has, however, a definition problem (Crowder (1995)). Suppose that a "working" correlation estimate $\hat{\alpha}$ under an assumed "working" correlation model is computed. This estimator may not converge to α as the data used for its computation may follow a different model. Thus, $\hat{\alpha}$ converges to α_0, say, which is different than α (Sutradhar and Das (1999)). As far as the formula for $\hat{\alpha}$ is concerned, it is developed based on method of moments following the assumed "working" correlation structure. For example, if an user decides to use an equi-correlation matrix as the "working" correlation structure for all K individuals, then the estimate would satisfy the estimating equation

$$\sum_{i=1}^{K}\sum_{j\neq u}^{n_i}(\tilde{y}_{ij}\tilde{y}_{iu} - \alpha) = 0 \tag{1.11}$$

(Liang and Zeger (1986), Sutradhar (2011, Section 6.4.3)), where

$$\tilde{y}_{ij} = \frac{y_{ij} - x_{ij}'\hat{\beta} - \hat{\gamma}(t_{ij})}{\hat{\sigma}^2},$$

with

$$\hat{\sigma}^2 = \sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(y_{ij} - x_{ij}'\hat{\beta} - \hat{\gamma}(t_{ij})\right)^2 \Big/ \sum_{i=1}^{K} n_i.$$

Similarly, for the estimation of a "working" unstructured correlation matrix, one uses the moment estimating formula

$$\hat{R}_i(\alpha) = \frac{1}{K\hat{\sigma}^2}\sum_{i=1}^{K} r_i r_i' \tag{1.12}$$

(Lin and Carroll (2001)) where $r_i = (r_{i1}, r_{i2}, \ldots, r_{in_i})'$ is the vector of residuals with $r_{ij} = y_{ij} - x_{ij}'\hat{\beta} - \hat{\gamma}(t_{ij})$.

Further note that there are two fold problems with the estimation of β by using the WGEE based formula (1.10). First, because of the aforementioned convergence problems for $R_i(\hat{\alpha})$ to the true correlation structure say $C_i(\rho)$, $\hat{\beta}_{\text{WGEE}}$ obtained by (1.10) may be less efficient sometimes as compared to a β estimate obtained by using $R_i(\hat{\alpha}) = I_{n_i}$ (see Sutradhar and Das (1999), Sutradhar (2011, Chapter 6)). Second, as we demonstrate in the next section, this $\hat{\beta}_{\text{WGEE}}$ is actually obtained from a partly standardized estimating equation (as opposed to a fully standardized equation) which is bound to reduce its efficiency. In the same section, we obtain a fully standardized semi-parametric GLS (FSSGLS), also referred to as the fully standardized semi-parametric generalized quasi-likelihood (FSSGQL), estimator for β which is consistent and highly efficient. The efficiency gain by the FSSGQL estimator over the partly standardized semi-parametric GEE (PSSGEE) estimator is also demonstrated in Section 3 by a simulation based empirical study.

2 PSSGEE versus FSSGQL estimation for the regression effects β

2.1 FSSGQL/FSSGLS estimation

In order to understand that the WGEE/WGLS estimator of β given by (1.10) is in fact a partly standardized GEE estimator, we recall the model (1.1) and replace the nonparametric function $\gamma(t_{ij})$ with its estimate $\hat{\gamma}(t_{ij})$ from (1.6) for known β. This substitution changes the model (1.1) to

$$\begin{aligned} y_{ij} &= x_{ij}'(t_{ij})\beta + \hat{\gamma}(t_{ij}) + \epsilon_{ij}^*(t_{ij}) \\ &= x_{ij}'(t_{ij})\beta + \hat{y}_{ij} - \hat{x}_{ij}'(t_{ij})\beta + \epsilon_{ij}^*(t_{ij}), \end{aligned} \tag{2.1}$$

where $\epsilon_{ij}^*(t_{ij})$ is a new error component different from that of (1.1). Now for all elements of the ith individual we use (2.1) and following the notation in (1.2) write

$$y_i - \hat{y}_i = (X_i - \hat{X}_i)\beta + \epsilon_i^*. \tag{2.2}$$

Now to obtain the GLS estimate of β in (2.2), it is important to examine the mean and variance of the error vector ϵ_i^*.

2.1.1 *Derivation for* $E[\epsilon_i^*]$. By (1.7) and (1.2), we write

$$\begin{aligned} E[\hat{Y}_i] &= \sum_{h=1}^{K} W_h(t_{i1},\ldots,t_{in_i}) E[Y_h] \\ &= \sum_{h=1}^{K} W_h(t_{i1},\ldots,t_{in_i})[X_h\beta + \gamma_h] \\ &= \hat{X}_i\beta + \sum_{h=1}^{K} W_h(t_{i1},\ldots,t_{in_i})\gamma_h, \end{aligned} \tag{2.3}$$

where $\gamma_h = [\gamma(t_{h1}),\ldots,\gamma(t_{hu}),\ldots,\gamma(t_{hn_h})]'$. It then follows that

$$E[Y_i - \hat{Y}_i] = [X_i - \hat{X}_i]\beta + \gamma_i - \sum_{h=1}^{K} W_h(t_{i1},\ldots,t_{in_i})\gamma_h. \tag{2.4}$$

By using (2.4) in (2.2), one obtains

$$\begin{aligned} E[\epsilon_i^*] &= \gamma_i - \sum_{h=1}^{K} W_h(t_{i1},\ldots,t_{in_i})\gamma_h \\ &= \mu_i^* = [\mu_{i1}^*,\ldots,\mu_{ij}^*,\ldots,\mu_{in_i}^*]', \end{aligned} \tag{2.5}$$

with

$$\mu_{ij}^* = \gamma(t_{ij}) - \sum_{h=1}^{K}\sum_{u=1}^{n_h} w_{hu}(t_{ij})\gamma_h(t_{hu}), \tag{2.6}$$

which is free from β.

2.1.2 *Derivation for* $\operatorname{cov}[\epsilon_i^*]$. Let $\Sigma_i(\rho)$ denote the true covariance matrix of the response vector y_i, that is

$$\operatorname{cov}(Y_i) = \Sigma_i(\rho) = A_i^{-1/2} C_i(\rho) A_i^{-1/2}$$

with

$$A_i = \operatorname{diag}[\sigma_{i11}(t_{i1}),\ldots,\sigma_{ijj}(t_{ij}),\ldots,\sigma_{in_in_i}(t_{in_i})]$$

as the $n_i \times n_i$ diagonal matrix with σ_{ijj} as the variance of $\epsilon_{ij}(t_{ij})$, and $C_i(\rho)$ is the true correlation matrix. We consider an auto-correlation class based structure for $C_i(\rho)$ which is discussed in Section 2.1.4. Next because $\hat{Y}_i =$

$\sum_{h=1}^{K}[W_h(t_{i1},\ldots,t_{in_i})]Y_h$, it then follows that

$$\begin{aligned}\Sigma_i^* &= \operatorname{cov}[\epsilon_i^*]\\ &= \operatorname{cov}(Y_i-\hat{Y}_i)\\ &= \operatorname{cov}(Y_i)+\operatorname{cov}(\hat{Y}_i)-2\operatorname{cov}(Y_i,\hat{Y}_i)\\ &= \Sigma_i(\rho)+\left[\sum_{h=1}^{K}W_h(t_{i1},\ldots,t_{in_i})\Sigma_h(\rho)W_h'(t_{i1},\ldots,t_{in_i})\right]\\ &\quad -2W_i(t_{i1},\ldots,t_{in_i})\Sigma_i(\rho).\end{aligned} \tag{2.7}$$

2.1.3 *FSSGQL estimating equation*. Now by (2.5) and (2.7), it follows from (2.2) that

$$E[Y_i-\hat{Y}_i]=[X_i-\hat{X}_i]\beta+\mu_i^*,\qquad \operatorname{cov}[Y_i-\hat{Y}_i]=\Sigma_i^*(\rho). \tag{2.8}$$

Consequently, following Sutradhar (2003, Section 3), one writes the GQL estimating equation for β as

$$\sum_{i=1}^{K}\frac{\partial[(X_i-\hat{X}_i)\beta+\mu_i^*]'}{\partial\beta}\big[\Sigma_i^*\big]^{-1}\big\{(y_i-\hat{y}_i)-(X_i-\hat{X}_i)\beta-\mu_i^*\big\}=0. \tag{2.9}$$

Note that this semi-parametric GQL estimating equation (2.9) is fully standardized, because it uses the correct longitudinal weight matrix $\Sigma_i^*(\rho)$, and the gradient function is computed for the correct mean vector $[X_i-\hat{X}_i]\beta+\mu_i^*$. Thus this equation may be referred to as the fully standardized semi-parametric GQL (FSSGQL) estimating equation, which is also the same as the FSS generalized least squared (FSSGLS) estimating equation. It is clear from (2.9) that the FSSGQL estimator has the closed form formula, which is given by

$$\begin{aligned}\hat{\beta}_{\mathrm{FSSGQL}} &= \left[\sum_{i=1}^{K}(X_i-\hat{X}_i)'\big(\Sigma_i^*\big)^{-1}(X_i-\hat{X}_i)\right]^{-1}\\ &\quad\times\sum_{i=1}^{K}(X_i-\hat{X}_i)'\big(\Sigma_i^*\big)^{-1}\big(y_i-\hat{y}_i-\hat{\mu}_i^*\big),\end{aligned} \tag{2.10}$$

where $\hat{\mu}_i^*$ is obtained by using $\hat{\gamma}_i$ from (1.6) for γ_i in (2.5), for all $i=1,\ldots,K$. Further note that in the present case, the FSSGQL estimating equation is equivalent to FSSGLS estimating equation.

2.1.4 *A general auto-correlation model*. To model the correlations of the repeated linear data, we follow Sutradhar (2010) (see also Sutradhar (2011,

Chapter 2)) and assume that the repeated data follow a class of auto-correlation structures that accommodates Gaussian type all possible autoregressive moving average of order r, s (ARMA(r, s)) correlation models with AR(1), MA(1), AR(2), MA(2), EQC (equi-correlations), as some special cases. Note that the AR(1), MA(1), and EQC structures for repeated data were also alluded in Liang and Zeger (1986), and subsequently these structures such as EQC correlation structure was used by Severini and Staniswallis (1994), for example, in the semi-parametric longitudinal setup. To be specific, we suggest that the error vector ϵ_i in (1.2) has the correlation matrix $C_i(\rho)$ given by

$$\begin{aligned} C_i(\rho) &= \begin{pmatrix} 1 & \rho_1 & \rho_2 \cdots & \rho_{n_i-1} \\ \rho_1 & 1 & \rho_1 \cdots & \rho_{n_i-2} \\ \vdots & & \cdots & \vdots \\ \rho_{n_i-1} & \rho_{n_i-2} & \cdots & 1 \end{pmatrix} \qquad \text{for all } i = 1, 2, \ldots, K; \\ \Sigma_i(\rho) &= \operatorname{var}(Y_i) = A_i^{1/2} C_i(\rho) A_i^{1/2}, \end{aligned} \tag{2.11}$$

where for $\ell = 1, \ldots, n_i - 1$, ρ_ℓ denotes the lag ℓ correlation between $\epsilon_{ij}(t_{ij})$ and $\epsilon_{i,j+\ell}(t_{i,j+\ell})$. Note that when variances are nonstationary, that is, the responses are heteroscedastic, one writes $\sigma^2(t_{ij})$ for $\sigma_{ijj}(t_{ij})$ (Fan et al. (2007, Section 2.1)). We however assume that the variances are stationary and hence write $A_i = \sigma^2 I_{n_i}$, where σ^2 is an unknown scalar constant, and I_{n_i} is the $n_i \times n_i$ identity matrix. Following examples demonstrate the correlation models that produce the $C_i(\rho)$ as in (2.11) in the linear model setup.

Examples of models:

(i) AR(1) *model*:

$$\begin{aligned} &\epsilon_{ij}(t_{ij}) = \phi \epsilon_{i,j-1}(t_{i,j-1}) + a_{ij}(t_{ij}), \\ &\quad |\phi| < 1, a_{ij}(t_{ij}) \overset{\text{i.i.d.}}{\sim} N\big(0, \sigma_a^2\big) \; \forall i = 1, 2, \ldots, K; \; j = 1, \ldots, n_i, \end{aligned}$$

(ii) MA(1) *model*:

$$\begin{aligned} &\epsilon_{ij}(t_{ij}) = \theta a_{i,j-1}(t_{i,j-1}) + a_{ij}(t_{ij}), \\ &\quad |\theta| < 1, a_{ij}(t_{ij}) \overset{\text{i.i.d.}}{\sim} N\big(0, \sigma_a^2\big) \; \forall i = 1, 2, \ldots, K; \; j = 1, \ldots, n_i, \end{aligned}$$

and

(iii) *EQC model*:

$$\begin{aligned} &\varepsilon_{ij}(t_{ij}) = \varepsilon_{i0}(t_{i0}) + a_{ij}(t_{ij}), \\ &\qquad a_{ij}(t_{ij}) \overset{\text{i.i.d.}}{\sim} \big(0, \sigma_a^2\big), \varepsilon_{i0}(t_{i0}) \sim N\big(0, \tilde{\sigma}^2\big), \end{aligned}$$

yield ρ_ℓ, the lag ℓ correlations between $\epsilon_{ij}(t_{ij})$ and $\epsilon_{i,j+\ell}(t_{i,j+\ell})$, as

$$\rho_\ell = \phi^\ell;$$
$$\rho_\ell = \begin{cases} \dfrac{\theta}{1+\theta^2}, & \text{for } \ell = 1, \\ 0, & \text{for } \ell = 2, 3, \ldots, n_i - 1, \end{cases} \quad \text{and} \quad \rho_\ell = \zeta = \frac{\tilde{\sigma}^2}{\tilde{\sigma}^2 + \sigma_a^2},$$

respectively, and they satisfy the auto-correlation structure $C_i(\rho)$ in (2.11).

2.1.5 *Estimation of the correlation matrix.* For $n = \max_{1 \le i \le K} n_i$, and

$$\delta_{iu} = \begin{cases} 1, & \text{if } u \le n_i, \\ 0, & \text{if } n_i < u \le n, \end{cases}$$

the auto-correlation matrix $C_i(\rho)$ (2.11) is estimated by using the estimates of lag correlation ρ_ℓ given by

$$\hat{\rho}_\ell = \frac{\sum_{i=1}^K \sum_{u=1}^{n-\ell} \delta_{iu}\delta_{i,u+\ell}\tilde{y}_{iu}\tilde{y}_{i,u+l} / \sum_{i=1}^K \sum_{u=1}^{n-\ell} \delta_{iu}\delta_{i,u+\ell}}{\sum_{i=1}^K \sum_{u=1}^{n_i} \delta_{iu}\tilde{y}_{iu}^2 / \sum_{i=1}^K \sum_{u=1}^{n_i} \delta_{iu}}, \quad \ell = 1, 2, \ldots, n-1 \tag{2.12}$$

(Sutradhar (2011, Section 2.2.2)) with $\tilde{y}_{iu} = \frac{y_{iu} - x_{iu}'\hat{\beta} - \hat{\gamma}(t_{iu})}{\hat{\sigma}}$, where $\hat{\beta}$ and $\hat{\gamma}(t)$ are the FSSGQL estimates of β and $\gamma(t)$, respectively, and σ^2 for the A_i matrix in (2.11) is estimated as

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^K \sum_{j=1}^{n_i} (y_{ij} - x_{ij}'\hat{\beta} - \hat{\gamma}(t_{ij}))^2}{\sum_{i=1}^K n_i}. \tag{2.13}$$

2.1.6 *Estimation steps.* Note that the moment estimators for lag correlations (2.12) and variance component (2.13) are primarily developed by assuming that β and $\gamma(\cdot)$ are known, but the estimates are obtained by using $\hat{\beta}_{\text{FSSGQL}}$ for β from (2.10), and $\hat{\gamma}(\cdot)$ from (1.6). For convenience of application of the proposed fully SSGQL approach, we now summarize this approach in the following four steps.

Step F1. For an initial value of β, we solve the "working" independence assumption based semi-parametric equation (1.5) to estimate the nonparametric function $\gamma(\cdot)$.

Step F2. The estimate of $\gamma(\cdot)$ from Step F1 along with the initial value of β are used in (2.13) to obtain first an initial estimate of the variance component σ^2, and then initial estimates of lag correlations by (2.12).

Step F3. In this step, the estimates of auto-correlations from Step F2 are used to compute first the kernel weights based covariance matrix $\Sigma_i^* = \text{cov}[Y_i - \hat{Y}_i]$ in (2.7), which is then used in (2.10) to obtain the FSSGQL estimate of β.

Step F4. Next, the first step estimate of β from Step F3 is applied to Step F1 to obtain an improved estimate for the nonparametric function $\gamma(\cdot)$. This constitute a cycle and the cycles of iterations continue until convergence.

2.2 PSSGEE estimation

When the FSSGQL estimator (2.10) is compared to the WGLS (or WGEE) estimator of β from (1.10), it is clear that WGEE estimator (1.10) was developed using an incorrect weight matrix $\text{var}(Y_i)$, whereas the FSSGQL estimator (2.10) is developed using the correct weight matrix $\Sigma_i^* = \text{var}(Y_i - \hat{Y}_i)$. This makes the WGEE estimator (1.10) a partly standardized estimator, because it uses correct derivative matrix but an incorrect weight matrix. For this reason, when compared to (2.10), the WGEE estimator in (1.10) may be referred to as the partly standardized GEE (PSSGEE) estimator. Note that this anomaly in PSSGEE arises because of the fact that the existing approaches developed PSSGEE ignoring unknown β in the formula for nonparametric function estimate (1.6). This incorrect use of the weight matrix is bound to produce less efficient estimate as compared to the FSSGQL estimator (2.10).

In the present semi-parametric linear longitudinal setup, the aforementioned efficiency loss can be verified in theory for the PSSGEE estimators by comparing their asymptotic variances with that of the FSSGQL estimator. To demonstrate this, consider the approximate asymptotic ($K \to \infty$) covariance matrix of the FSSGQL (2.10) and PSSGEE (1.10) estimators, given by

$$\text{var}[\hat{\beta}_{\text{FSSGQL}}] = \left[\sum_{i=1}^{K}(X_i - \hat{X}_i)'\big(\Sigma_i^*(\rho)\big)^{-1}(X_i - \hat{X}_i)\right]^{-1}$$

and

$$\begin{aligned}\text{var}[\hat{\beta}_{\text{PSSGEE}}] = &\left[\sum_{i=1}^{K}(X_i - \hat{X}_i)'V_i^{-1}(\alpha_0)(X_i - \hat{X}_i)\right]^{-1} \\ &\times \left[\sum_{i=1}^{K}(X_i - \hat{X}_i)'V_i^{-1}(\alpha_0)\Sigma_i^*(\rho)V_i^{-1}(\alpha_0)(X_i - \hat{X}_i)\right] \\ &\times \left[\sum_{i=1}^{K}(X_i - \hat{X}_i)'V_i^{-1}(\alpha_0)(X_i - \hat{X}_i)\right]^{-1},\end{aligned}$$

where α_0 is the converged value for the "working" correlation estimator $\hat{\alpha}$ (see Sutradhar and Das (1999)). Now assuming that the "working" covariance exists and computable such as in the "working" independence case when $\alpha_0 = 0$, $V_i(\alpha_0) = I_{n_i}$, by following Sutradhar (2011, Theorem 2.1, p. 13), for example, one may show that

$$\text{var}[\hat{\beta}_{u,\text{FSSGQL}}] \leq \text{var}[\hat{\beta}_{u,\text{PSSGEE}}] \qquad \text{for all } u = 1, \ldots, p.$$

However, this approach does not help to compare the PSSGEE estimators among themselves. For this, and also to examine the finite sample performances between the FSSGQL and all possible PSSGEE estimators, we conduct a simulation study in Section 3.

2.2.1 *Some remarks on heteroscedasticity based PSSHGEE(I) and PSSHGEE approaches*. Some authors, for example, Fan et al. (2007, Section 2.1), and Fan and Wu (2008, eqnuation (1)) have estimated the nonparametric function $\gamma(t)$ by using similar formula as in (1.5) but by using time dependent variances denoted by $\sigma^2(t)$ at a given time t. For the estimation of the regression effects β, they have used different "working" correlation structures for $R_i(\alpha)$ in the PSSGEE based estimate given by (1.10). Fan and Wu (2008, equation (6)) used the ordinary least squares (OLS) technique which is the same as using (1.10) with correlation matrix $R_i(\alpha) = I_{n_i}$, ignoring correlations. For a given t, the heteroscedasticity, that is, the time dependent variances were computed by

$$\sigma^2(t) = \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i} r_{ij}^2(t) w_{ij}(t)}{\sum_{i=1}^{K}\sum_{j=1}^{n_i} w_{ij}(t)}, \tag{2.14}$$

where $r_{ij}(t) = y_{ij} - x_{ij}'(t)\hat{\beta} - \hat{\gamma}(t)$, and $w_{ij}(t)$ are defined as in (1.5). Thus, for the estimation of β by (1.10), Fan and Wu (2008) use $\Sigma_i(\alpha) = A_i = \text{diag}[\sigma^2(t_{i1}), \ldots, \sigma^2(t_{in_i})]$. This partly standardized semi-parametric heteroscedastic GEE (I) estimator may be denoted by $\hat{\beta}_{\text{PSSHGEE(I)}}$.

The estimation of $\gamma(t)$ and $\sigma^2(t)$ is similar in both Fan et al. (2007) and Fan and Wu (2008). However, for β estimation by (1.10), Fan et al. (2007) have assumed that the error vector ϵ_i in (1.2) follow a multivariate normal distribution with a "working" correlation matrix $R_i(\alpha)$, and estimated the "working" correlation parameter α by maximizing the normal likelihood (Fan et al. (2007, equations (2)–(3))). This estimator may be referred to as the PSSHGEE based estimator. The normality assumption for the error vector is a further limitation in addition to the "working" correlation based limitation to define the correlations of the data. Thus, this PSSHGEE approach will be of limited use in practice. Nevertheless, we include this approach in the empirical efficiency comparison in Section 3, but

compute the lag correlations by moment approach (see (1.11)–(1.12)) which does not require any normality assumption.

3 A simulation study

The purpose of this section is to conduct a simulation study to examine the finite sample performance of the FSSGQL and various versions of the existing PSSGEE approaches in estimating the main regression parameters as well as the nuisance nonparametric function. As far as the longitudinal sample, regression parameters, nonparametric functions, and true correlation models are concerned, we consider the following simulation design.

3.1 Simulation design

(a) *Sample size*: $K = 100$; $n_i = 4$ for $i = 1, \ldots, K$; and $t_{ij} = j$ for all $i = 1, \ldots, K$, and $j = 1, \ldots, n_i$.

(b) *Covariate selection*: We consider $p = 2$ time dependent covariates with their values as

$$x_{ij1}(t_{ij}) = \begin{cases} \frac{1}{2}, & j = 1, 2, \\ 0, & j = 3, 4, \end{cases} \qquad i = 1, 2, \ldots, 50,$$

$$x_{ij1}(t_{ij}) = \begin{cases} -\frac{1}{2}, & j = 1, \\ 0, & j = 2, 3, \\ \frac{1}{2}, & j = 4, \end{cases} \qquad i = 51, 52, \ldots, 100,$$

$$x_{ij2}(t_{ij}) = \left\{ \frac{j - 2.5}{2j}, \qquad j = 1, 2, 3, 4, \qquad i = 1, 2, \ldots, 50, \right.$$

$$x_{ij2}(t_{ij}) = \begin{cases} 0, & j = 1, 2, \\ \frac{1}{2}, & j = 3, 4, \end{cases} \qquad i = 51, 52, \ldots, 100.$$

For the effects of these covariates, we consider $\beta_1 = 1.0$ and $\beta_2 = 0.5$.

(c) *Nonparametric function*: By using $t_{ij} = j$, we consider a quadratic as well as a harmonic function for $\gamma(t_{ij})$ given by

(i) $\gamma(t_{ij}) = 3 + 2(t_{ij} - \frac{n_i+1}{2}) + (t_{ij} - \frac{n_i+1}{2})^2$; $t_{ij} = 1, 2, 3, 4$,

(ii) $\gamma(t_{ij}) = \sin(2t_{ij})$.

(d) *True correlation structure*: We consider three correlation structures from Section 2.1.4, with selected values of parameters as indicated below.

(i) AR(1) *model*: $\phi = 0.5, 0.8$; $\sigma_a^2 = 1.0$,

(ii) MA(1) *model*: $\theta = 0.1, 0.4$; $\sigma_a^2 = 1.0$,

(iii) *EQC model*: $\zeta = \frac{\tilde{\sigma}^2}{\tilde{\sigma}^2 + \sigma_a^2} = 0.5, 0.8$; $\sigma_a^2 = 1.0$.

3.2 Data generation and simulation results

We use the above selected design parameters in (1.2) and simulate y_{ij} for $i = 1, 2, \ldots, 100$ and $j = 1, 2, 3, 4$ for 1000 times, following a true correlation structure that belongs to a class of auto-correlation models with correlation matrix (2.11). As the true correlation structure, we consider all three correlation models, namely AR(1), MA(1), and EQC models considered in Section 2.1.4.

Under each simulation, applying the four steps procedure of Section 2.1.5, we obtain the FSSGQL estimates of β, $\gamma(t)$, σ^2, and $\rho(\ell)$. Note that in this approach, irrespective of the true correlation models, AR(1), MA(1), or EQC, the correlation matrix is estimated by using the estimate of the general correlation matrix $C_i(\rho)$. Moreover, this approach uses corrected weight matrix in the estimating formula (2.10) for β.

The "working" correlations based PSSGEE estimates are obtained by following (1.10). Because the "working" correlations approach does not have any guidance for the selection of correlation model, one may choose any of the low order commonly used structures such as AR(1), MA(1), EQC, or "working" independence models (Liang and Zeger (1986)). Thus, if data are generated from the true AR(1) model, we examine through efficiency comparison whether one can use any of the other low order correlation models such as MA(1), EQC, or "working" independence models. Note that if one uses $C_i(\rho)$ as a "working" correlation structure, it is obvious that such a "working" correlations based estimates will be fully efficient because the true correlation models such as AR(1) is represented by the $C_i(\rho)$ structure. This is however not a good illustration of the use of "working" correlations matrix as because in the "working" correlations approach one also has to know how $C_i(\rho)$ matrix based estimate will perform even if the true correlation structure does not belong to auto-correlation class. Turning back to the low order correlations based estimates, if MA(1) structure is used as a "working" correlation structure, then MA(1) based correlations are estimated by method of moments, and so on. We also use the unstructured (UNS) (see Lin and Carroll (2001), e.g.) correlation model as a "working" correlation model. Further, the PSSHGEE(I) and PSSHGEE based estimates discussed in Section 2.2.1 are also computed.

For presentational simplicity in tabular and graphics form, we rename the FSSGQL estimates as semi-parametric GQL (SGQL) estimates, and similarly all PSSGEE and PSSHGEE estimates as SGEE and SHGEE estimates, respectively. The efficiency of these estimates are computed by comparing their simulations based variance with the variance of the known correlation structure based estimates, where the known correlation structure based estimates were computed by replacing the $C_i(\rho)$ matrix in the FSSGQL approach with the true correlation such as AR(1) correlation matrix.

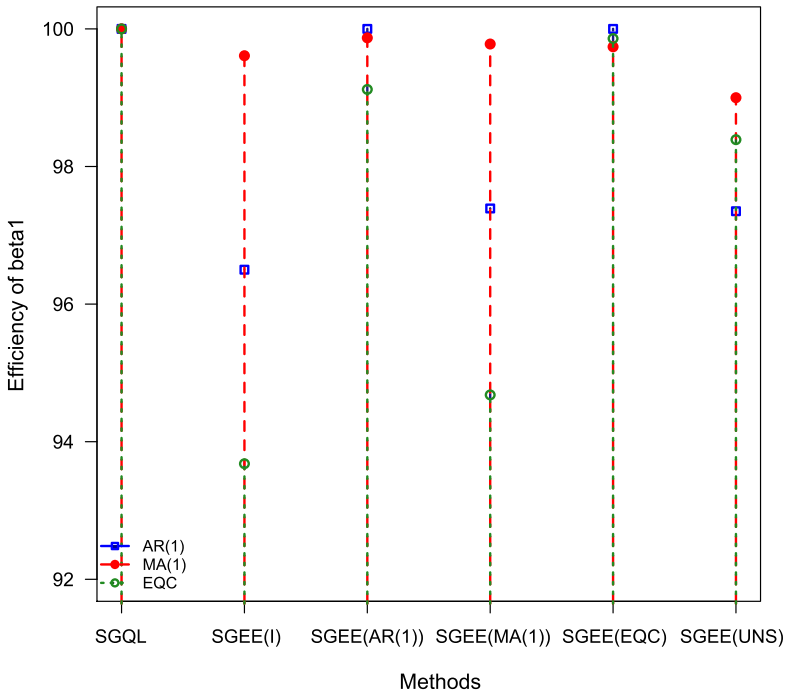

Figure 1 *Efficiency comparisons of various semi-parametric methods for the estimates of* β_1 *with* $\gamma(t) = 3 + 2(t - \frac{n+1}{2}) + (t - \frac{n+1}{2})^2$, *under selected correlation processes*: AR(1) *with* $\phi = 0.8$, MA(1) *with* $\theta = 0.4$ *and EQC with* $\zeta = 0.8$.

Because, the regression parameters β_1 and β_2 are of main interest, we mainly concentrate on the efficiency performance of the estimation methods for these two parameters. More specifically, we display the efficiencies for a selected correlation parameter value in Figures 1 and 2 for the estimation of β_1 and β_2 respectively, when $\gamma(t)$ is chosen as $3 + 2(t - \frac{n+1}{2}) + (t - \frac{n+1}{2})^2$ and in Figures 3 and 4 when $\gamma(t) = \sin(2t)$. When various methods of estimation for β_1 and β_2 are compared, all methods appear to produce unbiased and hence consistent estimates for both of the regression parameters. However, it is clear from Figures 1 and 2 that the proposed SGQL approach always yields the same or more efficient estimates than the other SGEE approaches including the unstructured correlations based SGEE (UNS) approach. For example, for the estimation of β_1 (Figure 1), under the true AR(1) correlation structure with $\phi = 0.8$ ($\rho = 0.8$) the SGQL and SGEE (EQC) provide almost equally efficient estimate whereas the other SGEE approaches including SGEE (UNS) provide less efficient estimate. Under the true MA(1) correlation model with $\theta = 0.4$ ($\rho = 0.35$), all approaches appear to produce the almost equal efficient estimate for β_1, the SGEE (UNS) being slightly inferior. Similarly under the EQC process with $\zeta = 0.8$ ($\rho = 0.8$) all SGEE approaches are less efficient than the SGQL approach. Note that SGEE(I) performs the worst among all

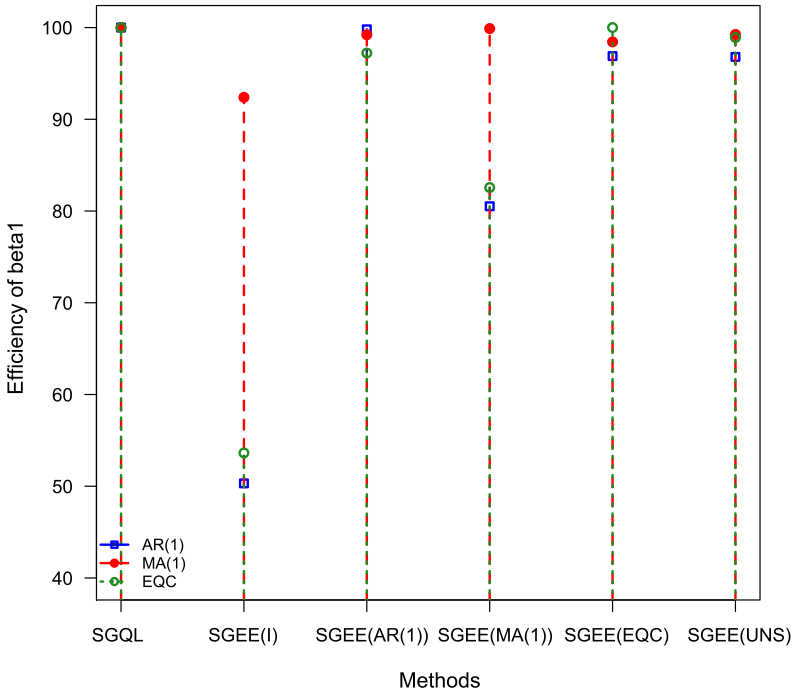

Figure 2 *Efficiency comparisons of various semi-parametric methods for the estimates of* β_2 $\gamma(t) = 3 + 2(t - \frac{n+1}{2}) + (t - \frac{n+1}{2})^2$, *under selected correlation processes*: AR(1) *with* $\phi = 0.8$, MA(1) *with* $\theta = 0.4$ *and EQC with* $\zeta = 0.8$.

'working' correlation approaches. Figure 2 shows that for the estimation of β_2, all SGEE approaches are in general inferior to the SGQL approach, the SGEE(I) being the worst followed by SGEE (MA(1)). The efficiency performances of 'working' correlation methods reported through Figures 3 and 4 are the same as explained in Figures 1 and 2. Thus, the SGQL approach uniformly produce same or higher efficient estimates for both β_1 and β_2 irrespective of the true correlation structures as well as nonparametric functional forms.

The efficiency of SHGEE(I) and SHGEE approaches (Fan et al. (2007), Fan and Wu (2008)) discussed in Section 2.2.1 are displayed in Tables 1(a), 2 and 3, along with other PSSGEE estimates. It is clear that similar to other SGEE approaches they also produce the regression estimates with larger variances as compared to the FSSGQL estimates.

Note that the estimation of $\beta = (\beta_1, \beta_2)'$ require the estimating formula of $\gamma(t)$ which is estimated by using the semi-parametric QL (SQL) estimating equation (1.5) under all SGQL and SGEE approaches. For the bandwidth b involved in the Gaussian kernel in (1.5), we have chosen $b = \frac{1}{(Kn)^{1/5}}$ (Pagan and Ullah (1999, p. 25)). For selected values of the correlation parameter, the estimates of $\gamma(t) = 3 + 2(t - \frac{n+1}{2}) + (t - \frac{n+1}{2})^2$ for all possible values of t are shown in

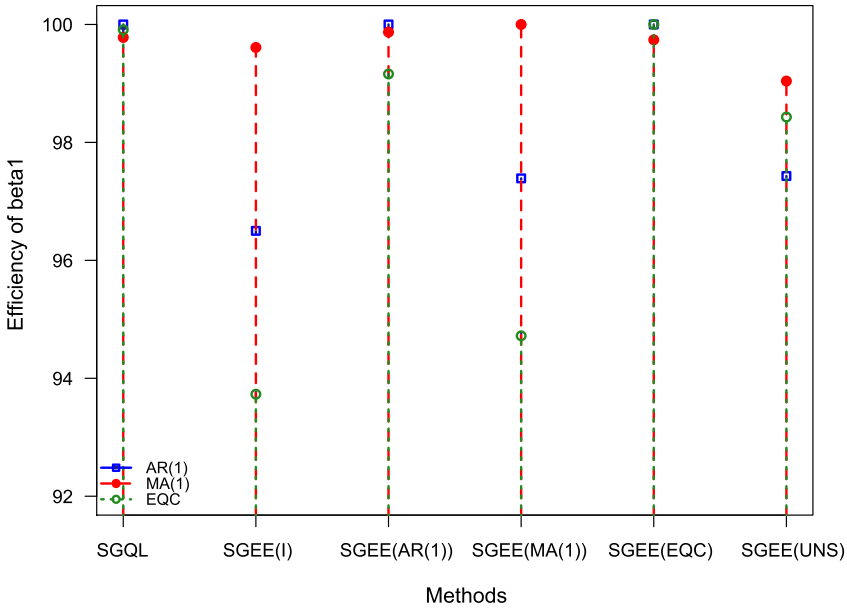

Figure 3 *Efficiency comparisons of various semi-parametric methods for the estimates of* β_1 *with* $\gamma(t) = \sin 2t$, *under selected correlation processes*: AR(1) *with* $\phi = 0.8$, MA(1) *with* $\theta = 0.4$ *and EQC with* $\zeta = 0.8$.

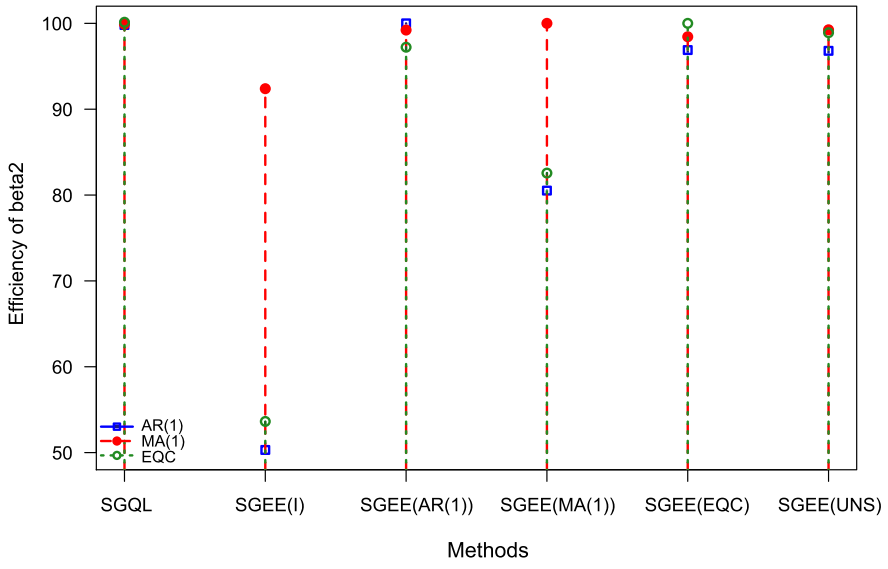

Figure 4 *Efficiency comparisons of various semi-parametric methods for the estimates of* β_2 *with* $\gamma(t) = \sin 2t$, *under selected correlation processes*: AR(1) *with* $\phi = 0.8$, MA(1) *with* $\theta = 0.4$ *and EQC with* $\zeta = 0.8$.

Table 1(a) *Estimates under the true* AR(1) *model. Simulated means* (*SMs*), *simulated standard errors* (*SSEs*) *and estimated standard errors* (*ESEs*) *of the estimates of regression parameters* $\beta_1 = 1$ *and* $\beta_2 = 0.5$, *under* AR(1) *correlation model for selected values of the model parameters* ϕ *and* σ^2; *with* $K = 100$; $n = 4$; *and* 1000 *simulations*

ϕ (σ^2)	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
0.5 (1.33)	SGEE (AR(1))	SM	0.9997	0.5082		0.4974		
		SSE	0.2339	0.3073		0.0580		
		ESE	0.2228	0.2916				
	SGQL	SM	0.9993	0.5072		0.4987	0.2489	0.1277
		SSE	0.2340	0.3073		0.0504	0.0728	0.0973
		ESE	0.2228	0.2910				
	SGEE (UNS)	SM	0.9999	0.5077				
		SSE	0.2365	0.3105				
		ESE	0.2233	0.2910				
	SGEE(I)	SM	0.9999	0.5094				
		SSE	0.2343	0.3715				
		ESE	0.2279	0.3029				
	SGEE (MA(1))	SM	0.9996	0.5086	0.4692			
		SSE	0.2349	0.3099	0.0251			
		ESE	0.2546	0.2864				
	SGEE (EQC)	SM	0.9998	0.5087	0.3529			
		SSE	0.2339	0.3112	0.0549			
		ESE	0.1850	0.3333				
	SHGEE(I)	SM	0.9999	0.5093				
		SSE	0.2343	0. 3722				
		ESE	0.2290	0.3044				
	SHGEE	SM	0.9991	0.5074		0.4983	0.2500	0.1292
		SSE	0.2337	0.3077		0.0477	0.0732	0.0981
		ESE	0.2236	0.2930				
0.8 (2.78)	SGEE (AR(1))	SM	1.0005	0.5066		0.7998		
		SSE	0.2425	0.3149		0.0298		
		ESE	0.2354	0.3002				
	SGQL	SM	1.0003	0.5057		0.8001	0.6400	0.5140
		SSE	0.2425	0.3155		0.0316	0.0504	0.0730
		ESE	0.2353	0.2981				
	SGEE (UNS)	SM	1.0013	0.5047				
		SSE	0.2491	0.3253				
		ESE	0.2331	0.2907				
	SGEE(I)	SM	1.0022	0.5181				
		SSE	0.2513	0.6259				
		ESE	0.3291	0.4374				

Table 1(a) (*Continued*)

ϕ (σ^2)	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
	SGEE (MA(1))	SM	1.0018	0.5111	0.4800			
		SSE	0.2490	0.3911	0.0000			
		ESE	0.3683	0.4106				
	SGEE (EQC)	SM	1.0011	0.5083	0.6987			
		SSE	0.2425	0.3250	0.0400			
		ESE	0.1839	0.4004				
	SHGEE(I)	SM	1.0023	0.5181				
		SSE	0.2524	0.6275				
		ESE	0.3313	0.4401				
	SHGEE	SM	1.001	0.5062		0.8001	0.6418	0.5180
		SSE	0.2450	0.3178		0.0263	0.0503	0.0735
		ESE	0.2360	0.3021				

Table 1(b) *Estimates under the true* AR(1) *model. Simulated means* (*SMs*) *and simulated standard errors* (*SSEs*) *of the estimates of regression parameters* $\beta_1 = 1$ *and* $\beta_2 = 0.5$, *under* AR(1) *correlation model for selected values of the model parameters* ϕ *and* σ^2; *with* $\gamma(t) = 3 + 2(t - \frac{n+1}{2}) + (t - \frac{n+1}{2})^2$; $K = 100$; $n = 4$; *and* 1000 *simulations* (*Epanechnikov Kernel*)

ϕ (σ^2)	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
0.5 (1.33)	SGEE (AR(1))	SM	0.9997	0.5082		0.4974		
		SSE	0.2339	0.3073		0.0580		
	SGQL	SM	0.9993	0.5072		0.4988	0.2490	0.1278
		SSE	0.2340	0.3071		0.0504	0.0728	0.0973
	SGEE (UNS)	SM	0.9999	0.5076				
		SSE	0.2366	0.3105				
	SGEE(I)	SM	0.9999	0.5094				
		SSE	0.2343	0.3714				
	SGEE (MA(1))	SM	0.9996	0.5086	0.4693			
		SSE	0.2349	0.3099	0.0251			
	SGEE (EQC)	SM	0.9998	0.5087	0.3530			
		SSE	0.2333	0.3112	0.0549			
0.8 (2.78)	SGEE (AR(1))	SM	1.0005	0.5066		0.7998		
		SSE	0.2425	0.3149		0.0298		
	SGQL	SM	1.0003	0.5057		0.8002	0.6401	0.5140
		SSE	0.2425	0.3155		0.0316	0.0504	0.0730

Table 1(b) (*Continued*)

ϕ (σ^2)	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
	SGEE (UNS)	SM	1.0013	0.5047				
		SSE	0.2491	0.3253				
	SGEE(I)	SM	1.0022	0.5181				
		SSE	0.2513	0.6259				
	SGEE (MA(1))	SM	1.0018	0.5111	0.4800			
		SSE	0.2490	0.3911	0.0000			
	SGEE (EQC)	SM	1.0011	0.5083	0.6988			
		SSE	0.2424	0.3250	0.0400			

Table 2 *Estimates under the true* MA(1) *model. Simulated means* (*SMs*), *simulated standard errors* (*SSEs*) *and estimated standard errors* (*ESEs*) *of the estimates of regression parameters* $\beta_1 = 1$ *and* $\beta_2 = 0.5$, *under* MA(1) *correlation model for selected values of the model parameters* θ *and* σ^2; *with* $K = 100$; $n = 4$; *and* 1000 *simulations*

θ (σ^2)	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
0.1	SGEE (MA(1))	SM	1.0014	0.4982		0.0971		
		SSE	0.2005	0.2641		0.0586		
		ESE	0.2038	0.2685				
	SGQL	SM	1.0011	0.4986		0.0971	0.0000	−0.0018
		SSE	0.2009	0.2643		0.0586	0.0684	0.0982
		ESE	0.2035	0.2677				
	SGEE (UNS)	SM	1.0001	0.4976				
		SSE	0.2026	0.2654				
		ESE	0.2040	0.2688				
	SGEE(I)	SM	1.0013	0.4992				
		SSE	0.2005	0.2650				
		ESE	0.1985	0.2638				
	SGEE (AR(1))	SM	1.0013	0.4983	0.0865			
		SSE	0.2006	0.2651	0.0810			
		ESE	0.2018	0.2679				
	SGEE (EQC)	SM	1.0012	0.4985	0.0481			
		SSE	0.2005	0.2642	0.0449			
		ESE	0.1939	0.2714				
	SHGEE(I)	SM	1.0118	0.4996				
		SSE	0.2007	0.2656				
		ESE	0.1994	0.2651				
	SHGEE	SM	1.0015	0.4992		0.0972	−0.0000	−0.0017
		SSE	0.2014	0.2658		0.0583	0.0685	0.0990
		ESE	0.2044	0.2690				

Table 2 (*Continued*)

θ (σ^2)	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
0.4	SGEE (MA(1))	SM	1.0008	0.4948		0.3435		
		SSE	0.2275	0.2781		0.0528		
		ESE	0.2314	0.2838				
	SGQL	SM	1.0004	0.4954		0.3435	−0.0007	−0.0033
		SSE	0.2280	0.2784		0.0528	0.0726	0.0973
		ESE	0.2313	0.2830				
	SGEE (UNS)	SM	0.9991	0.4949				
		SSE	0.2298	0.2802				
		ESE	0.2318	0.2835				
	SGEE(I)	SM	1.0003	0.4970				
		SSE	0.2284	0.3010				
		ESE	0.2126	0.2826				
	SGEE (AR(1))	SM	1.0004	0.4959	0.2778			
		SSE	0.2278	0.2803	0.0731			
		ESE	0.2190	0.2894				
	SGEE (EQC)	SM	1.0002	0.4962	0.1702			
		SSE	0.2281	0.2825	0.0523			
		ESE	0.1945	0.3058				
	SHGEE(I)	SM	1.0011	0.4975				
		SSE	0.2281	0.3009				
		ESE	0.2137	0.2841				
	SHGEE	SM	1.0008	0.4969		0.3430	−0.0002	−0.0033
		SSE	0.2283	0.2784		0.0511	0.0728	0.0983
		ESE	0.2323	0.2846				

Figures 5, 6 and 7 under true AR(1), MA(1) and EQC models, respectively. We have also computed the estimates for $\gamma(t) = \sin(2t)$ under all these three true correlation models, but displayed the EQC case only in Figure 8 to save space. It is clear from the these four figures that this nonparametric function is estimated very well by the semi-parametric QL approach.

Further note that as a reviewer suggested, by generating the data under the AR(1) correlation structure, we have estimated the $\gamma(t) = 3 + 2(t - \frac{n+1}{2}) + (t - \frac{n+1}{2})^2$ solving the semi-parametric QL estimating equation (1.5) but by using the well known Epanechnikov kernel (Pagan and Ullah (1999, p. 28), Fan and Gijbels (1996), Chen and Jin (2005))

$$p_{hu}(\psi) = \begin{cases} \frac{1}{4}[1 - \psi^2], & \text{for } |\psi| \le 1, \text{with } \psi = \dfrac{t_0 - t_{hu}}{b}, \\ 0, & \text{otherwise.} \end{cases}$$

Table 3 *Estimates under the true EQC model. Simulated means* (*SMs*), *simulated standard errors* (*SSEs*), *and estimated standard errors* (*ESEs*) *of the estimates of regression parameters* $\beta_1 = 1$ *and* $\beta_2 = 0.5$, *under Equi correlation model for selected values of the model parameters* ζ *and* σ^2; *with* $K = 100$; $n = 4$; *and* 1000 *simulations*

ζ (σ^2)	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
0.5	SGEE (EQC)	SM	0.9967	0.5211		0.4994		
		SSE	0.2111	0.4088		0.0504		
		ESE	0.2005	0.3933				
	SGQL	SM	0.9968	0.5194		0.5003	0.4986	0.4985
		SSE	0.2115	0.4088		0.0564	0.0577	0.0870
		ESE	0.2003	0.3933				
	SGEE (UNS)	SM	0.9979	0.5205				
		SSE	0.2125	0.4118				
		ESE	0.2000	0.3933				
	SGEE(I)	SM	0.9968	0.5215				
		SSE	0.2124	0.5019				
		ESE	0.2797	0.3717				
	SGEE (AR(1))	SM	0.9967	0.5204	0.6388			
		SSE	0.2131	0.4180	0.0450			
		ESE	0.2494	0.3231				
	SGEE (MA(1))	SM	0.9969	0.5201	0.4668			
		SSE	0.2140	0.4165	0.0282			
		ESE	0.3123	0.3520				
	SHGEE(I)	SM	0.9967	0.5214				
		SSE	0.2131	0.5036				
		ESE	0.2813	0.3738				
	SHGEE	SM	0.9971	0.5195		0.5011	0.4999	0.5011
		SSE	0.2121	0.4107		0.0551	0.0579	0.0768
		ESE	0.2011	0.3962				
0.8	SGEE (EQC)	SM	0.9968	0.5216		0.7992		
		SSE	0.2135	0.4725		0.0274		
		ESE	0.2023	0.4604				
	SGQL	SM	0.9968	0.5192		0.7998	0.7989	0.7986
		SSE	0.2138	0.4725		0.0317	0.0296	0.0532
		ESE	0.2012	0.4575				
	SGEE (UNS)	SM	0.9983	0.5212				
		SSE	0.2170	0.4777				
		ESE	0.1989	0.4528				
	SGEE(I)	SM	0.9981	0.5325				
		SSE	0.2279	0.8811				
		ESE	0.4420	0.5874				
	SGEE (AR(1))	SM	0.9964	0.5198	0.8715			
		SSE	0.2154	0.4860	0.0188			
		ESE	0.2614	0.3305				

Table 3 (*Continued*)

ζ (σ^2)	Method	Quantity	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$
	SGEE (MA(1))	SM	0.9980	0.5252	0.4800			
		SSE	0.2255	0.5723	0.0000			
		ESE	0.4947	0.5515				
	SHGEE(I)	SM	0.9981	0.5325				
		SSE	0.2297	0.8828				
		ESE	0.4454	0.5917				
	SHGEE	SM	0.9975	0.5201		0.8007	0.8002	0.8010
		SSE	0.2167	0.4783		0.0288	0.0296	0.0364
		ESE	0.2029	0.4619				

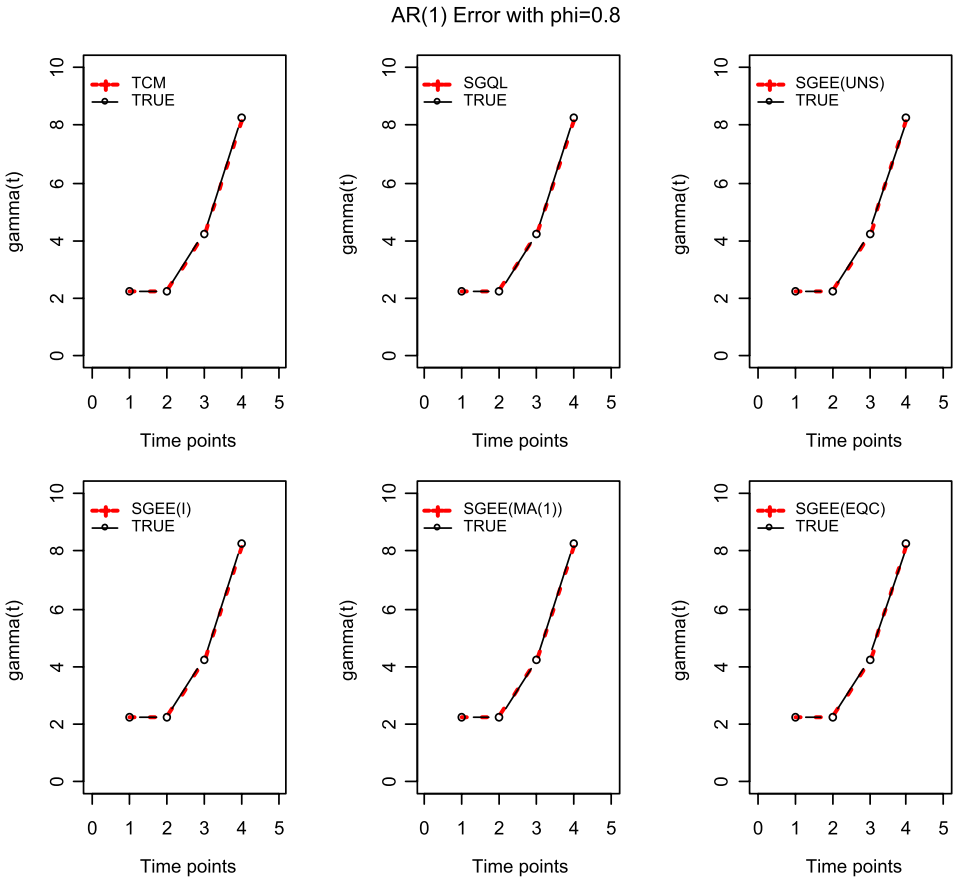

Figure 5 *Selected correlation structure based fully standardized semi-parametric GQL estimation for the unspecified time dependent function* $(\gamma(t))$ *in a linear model with* AR(1) *correlated errors.*

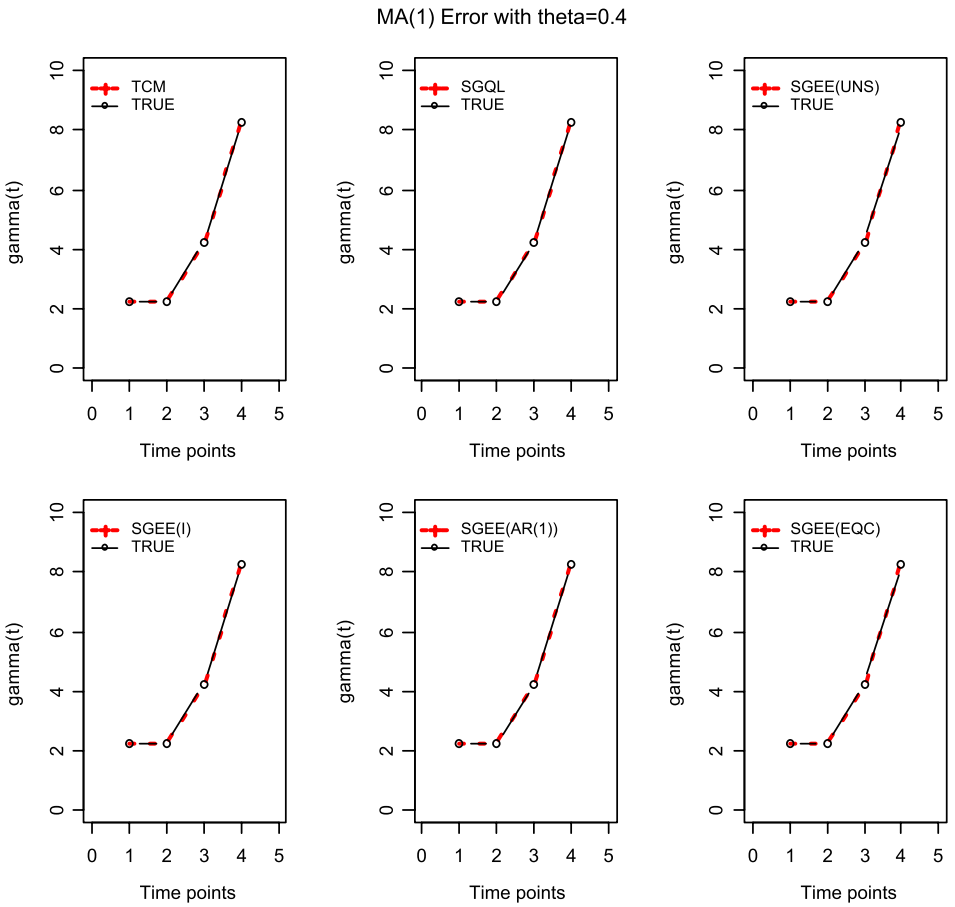

Figure 6 *Selected correlation structure based fully standardized semi-parametric GQL estimation for the unspecified time dependent function* $(\gamma(t))$ *in a linear model with* MA(1) *correlated errors.*

in place of the Gaussian kernel

$$p_{hu}\left(\frac{t_0 - t_{hu}}{b}\right) = \frac{1}{\sqrt{2\pi}b}\exp\left(\frac{-1}{2}\left(\frac{t_0 - t_{hu}}{b}\right)^2\right).$$

The estimates of $\gamma(t)$ were found to be almost the same as those produced in Figure 5, and hence are not reproduced. The corresponding regression estimates under this alternative kernel are given in Table 1(b). When compared to Table 1(a), the estimates and standard errors are also found to be almost the same. Thus the Gaussian and the Epanechnikov kernel appear to perform almost the same in estimating $\gamma(t)$ and hence β, in the present simulation setup.

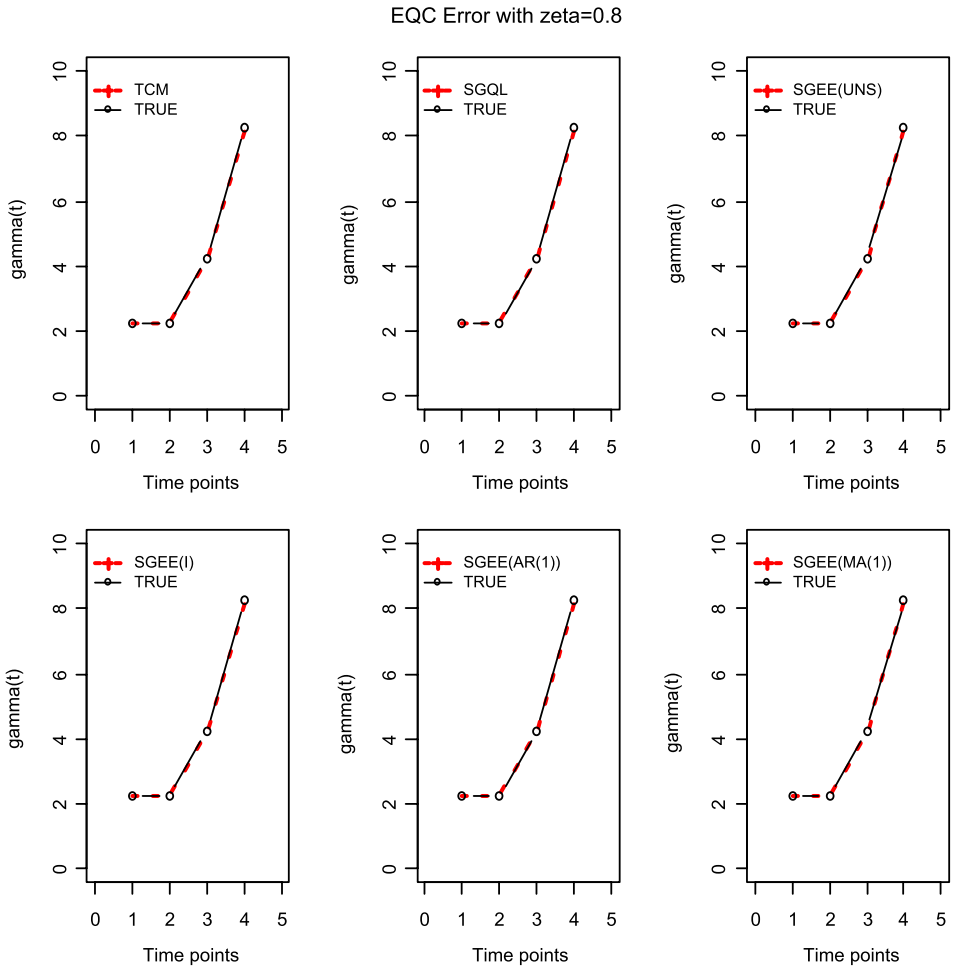

Figure 7 *Selected correlation structure based fully standardized semi-parametric GQL estimation for the unspecified time dependent function* $(\gamma(t))$ *in a linear model with Equi correlated errors.*

4 Concluding remarks

In a semi-parametric correlation model, the estimation of the finite dimensional regression parameters is affected by both the estimation of infinite dimensional nonparametric function and the estimation of correlation structure of the repeated responses. Under the assumption that the responses follow a class of Gaussian type auto-correlations, this paper has demonstrated that the proposed FSSGQL approach is highly efficient as compared to the existing various PSSGEE approaches for the estimation of regression parameters. Note that among all PSSGEE approaches, as expected, the independence assumption based semi-parametric GEE approach was found to be the worst in almost all cases.

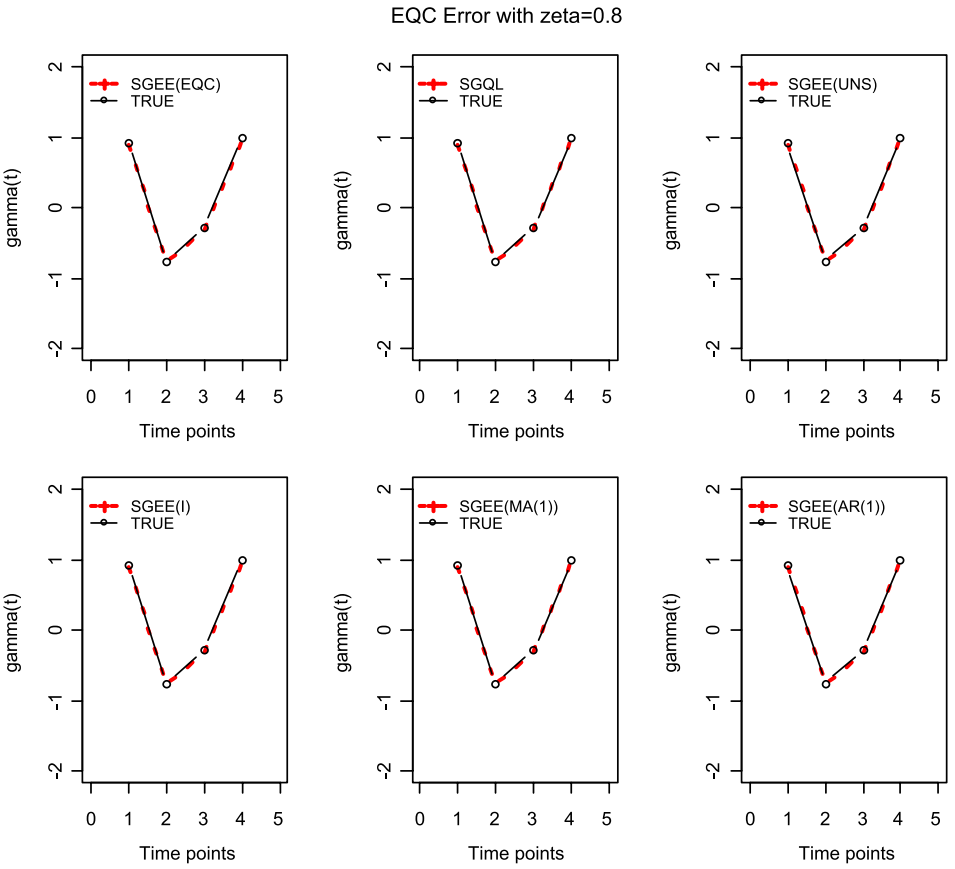

Figure 8 *Selected correlation structure based fully standardized semi-parametric GQL estimation for the unspecified time dependent function* $(\gamma(t) = \sin 2t)$ *in a linear model with Equi correlated errors.*

Acknowledgments

The authors would like to thank the editor and two referees for their valuable comments and suggestions leading to improvement of the paper.

References

Chen, K. and Jin, Z. (2005). Local polynomial regression analysis for clustered data. *Biometrika* **92**, 59–74. MR2158610

Crowder, M. (1995). On the use of a working correlation matrix in using generalized linear models for repeated measures. *Biometrika* **82**, 407–410.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* London: Chapman & Hall. MR1383587

Fan, J., Huang, T. and Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of the American Statistical Association* **102**, 632–641. MR2370857

Fan, J. and Wu, Y. (2008). Semiparametric estimation of covariance matrices for longitudinal data. *Journal of the American Statistical Association* **103**, 1520–1533. MR2504201

Hua, L. (2010). Spline-based sieve semiparametric generalized estimating equation for panel count data. Ph.D. thesis, Dept. Biostatistics, Univ. Iowa. MR2941442

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22. MR0836430

Lin, X. and Carroll, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association* **96**, 1045–1056. MR1947252

Pagan, A. and Ullah, A. (1999). *Nonparametric Econometrics.* Cambridge: Cambridge Univ. Press. MR1699703

Severini, T. A. and Staniswalis, J. G. (1994). Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association* **89**, 501–511. MR1294076

Sneddon, G. and Sutradhar, B. C. (2004). On semiparametric familial-longitudinal models. *Statistics and Probability Letters* **69**, 369–379. MR2089012

Sutradhar, B. C. (2003). An overview on regression models for discrete longitudinal responses. *Statistical Science* **18**, 377–393. MR2056579

Sutradhar, B. C. (2010). Inferences in generalized linear longitudinal mixed models. *Canadian Journal of Statistics* **38**, 174–196. MR2682757

Sutradhar, B. C. (2011). *Dynamic Mixed Models for Familial Longitudinal Data.* New York: Springer. MR2777359

Sutradhar, B. C. and Das, K. (1999). On the efficiency of regression estimators in generalized linear models for longitudinal data. *Biometrika* **86**, 459–465. MR1705378

You, J. and Chen, G. (2007). Semiparametric generalized least squares estimation in partially linear regression models with correlated errors. *Journal of Statistical Planning and Inference* **137**, 117–132. MR2292845

Zeger, S. L. and Diggle, P. J. (1994). Semi-parametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**, 689–699.

Mathematics and Statistics Department
Memorial University
St. John's, NL, A1C 5S7
Canada
E-mail: vwkv13@mun.ca
bsutradh@mun.ca