Consistent Model Selection and Data-driven Smooth Tests for Longitudinal Data in the Estimating Equations Approach

Lan Wang [1] and Annie Qu [2]

Abstract

Model selection for marginal regression analysis of longitudinal data is challenging due to the presence of correlation and the difficulty of specifying the full likelihood, particularly for correlated categorical data. This paper introduces a novel BIC-type model selection criterion based on the quadratic inference function (Qu, Lindsay and Li, 2000), which does not require the full likelihood or quasilikelihood. With probability approaching one, the criterion selects the most parsimonious correct model. Although a working correlation matrix is assumed, there is no need to estimate the nuisance parameters in the working correlation matrix; moreover, the model selection procedure is robust against the misspecification of the working correlation matrix. The BIC-type criterion can also be used to construct a data-driven Neyman smooth test for checking the goodness-of-fit of a postulated model. This test is especially useful and often yields much higher power in situations where the classical directional test behaves poorly. The finite sample performance of the model selection and model checking procedures is demonstrated through Monte Carlo studies and analysis of a clinical trial data set.

Key Words: BIC, correlated data, generalized estimating equations, longitudinal data, marginal model, model checking, model selection, Neyman smooth test, quadratic inference function.

[1]385 Ford Hall, School of Statistics, University of Minnesota, 224 Church Street SE, Minneapolis, MN 55455. Email: lan@stat.umn.edu

[2]Kidder Hall, Room 82, Statistics Department, Oregon State University, Corvallis, OR 97331-4606. Email: qu@stat.orst.edu

1 Introduction

We consider the problem of model selection for the marginal regression analysis of longitudinal data. The intrinsic complexity of longitudinal data makes it challenging, if not impossible, to apply existing model selection procedures in the literature that were developed for independent data or based on the full likelihood, see Miller (1990), Shao (1997), McQuarrie and Tsai (1998), Burnham and Anderson (2002) and the references therein. First, the full likelihood for longitudinal data is often difficult to specify, particularly for correlated non-Gaussian data. This was actually the main motivation for Liang and Zeger (1986) to develop the generalized estimating equations (GEE) approach. Second, the presence of correlation makes it hard to establish the underlying asymptotic theory, which plays a crucial role in understanding the mechanism of model selection.

The approach developed in this article is based on the quadratic inference function (QIF) proposed by Qu, Lindsay and Li (2000), which generalizes the generalized method of moments (GMM, see Hansen, 1982) to the analysis of longitudinal data. We introduce a novel model selection criterion that involves a term accounting for goodness-of-fit based on the QIF and a penalty term corresponding to model complexity. We call this new criterion BIQIF, as it extends the basic idea of the Bayesian Information Criterion (BIC) from Schwarz (1978). We establish that BIQIF retains the desirable consistency property of Schwarz' BIC. That is, with probability approaching one, BIQIF selects the most parsimonious correct marginal regression model.

In addition to the consistency property, BIQIF possesses a few appealing practical advantages. The criterion does not require the specification of the full likelihood or quasilikelihood. Although a working correlation matrix is assumed, the implementation of BIQIF needs not to estimate the nuisance parameters. This is because the QIF is computed via a series of basis matrices which are formed based on the structure of the working correlation, and does not involve nuisance parameters. Furthermore, it is worth emphasizing that BIQIF inherits the built-in robustness of GEE with respect to the misspecification of the working correlation matrix.

There have been few existing model selection methods available for marginal analysis of longitudinal data. Pan (2001) was the first to formally tackle this problem. He proposed an AIC-type criterion by considering the quasi-likelihood. In addition to selecting covariates in regression, his procedure can also be used to select the covariance matrix. Cantoni, Flemming and Ronchetti (2005) developed a generalized version of Mallow's C_p which provides an estimate of a measure of a model's adequacy for prediction. Pan and Mackenzie (2003) introduced a model selection criterion for mean-covariance structure models. Although their reported numerical simulations suggest the effectiveness of these approaches, studies of theoretical asymptotic properties are still lacking.

Other related approaches in different settings include the AIC and BIC based on empirical likelihood (Kolaczyk, 1995, Small, 2002), and the moment selection criterion of Andrews (1999) for the GMM where the purpose is to select valid moment conditions when a vector of moment conditions is available and the number of correct moments is greater than the number of parameters. These, however, do not directly apply to longitudinal data analysis.

Another contribution of this paper is a data-driven goodness-of-fit test for model checking. This is motivated by the concern that classical tests whose power is focused on a fixed alternative may exhibit inferior power for alternatives in other directions, see also §5.4 in Hart (1997). Our approach chooses a data-driven alternative using BIQIF, which generalizes the data-driven tests by Ledwina (1994) and Kallenberg and Ledwina (1997) for independent data and relates to Neyman's (1937) original work on the smooth test. The flexibility of a data-driven alternative often leads to improved power performance, as suggested by simulations and the analysis of the clinical trial data for respiratory disorder in Section 5. Other work on goodness-of-fit tests for the GEE includes Barnhart and Williamson (1998) and Pan (2002), both of which specifically target correlated binary data.

Section 2 provides a brief review of GEE and QIF. Section 3 proposes the new model selection criterion. Section 4 introduces the data-driven goodness-of-fit test. Section 5 illustrates the application using a clinical trial data set. The final section contains some

concluding remarks.

2 The quadratic inference function

2.1 Marginal regression model

In a longitudinal study, an $n_i \times 1$ vector of responses $Y_i = (y_{i1}, \ldots, y_{in_i})^T$ is observed for each subject i. Let $X_i = (x_{i0}, x_{i1}, \ldots, x_{in_i})^T$ be the $n_i \times (q+1)$ matrix of covariate values associated with Y_i, where x_{i0} is a column of ones. The marginal regression model relates the covariates to the marginal mean through

$$g(E(y_{ij}|x_{ij})) = x_{ij}^T \beta, \quad \beta \in \mathcal{B} \tag{1}$$

where g is a known link function, and $\beta = (\beta_0, \beta_1, \ldots, \beta_q)^T$ is a $(q+1) \times 1$ vector of unknown parameters in the parameter space $\mathcal{B}$.

With assumptions on the first two marginal moments, the GEE (Liang and Zeger, 1986) estimator of β is defined as the solution of

$$\sum_{i=1}^{N} \dot{\mu_i}^T V_i^{-1} (Y_i - \mu_i) = 0, \tag{2}$$

where $\mu_i = (\mu_{i1}, \ldots, \mu_{in_i})^T$ with $\mu_{ij} = E(y_{ij}|x_{ij})$, $\dot{\mu_i} = \partial \mu_i / \partial \beta$ is a $n_i \times q$ matrix, and $V_i = A_i^{1/2} R(\alpha) A_i^{1/2}$ with A_i being the diagonal matrix of marginal variances $Var(Y_i)$ and $R(\alpha)$ being the working correlation matrix. The main advantage of the GEE approach is that it yields a consistent estimator even if the working correlation matrix is misspecified.

There have been many important developments of GEE since the work of Liang and Zeger, see Diggle et al. (2002) and Hardin and Hilbe (2003). For instance, Prentice and Zhao (1991) proposed estimating equations for jointly modeling the mean and covariance parameters; Qu, Lindsay and Li (2000) introduced the quadratic inference function to improve the efficiency of GEE; Balan and Schiopu-Kratina (2005) derived a two-step estimation procedure for the marginal model based on the pseudo-likelihood and Chiou

and Müller (2005) developed a new marginal approach based on semiparametric quasi-likelihood regression. See also the recent work of Hall and Severini (1998), Shults and Chaganty (1998), Stoner and Leroux (2002) and Wang and Carey (2003).

2.2 Quadratic inference function

The QIF approach of Qu et al. (2000) is motivated by the observation that the inverse of the commonly used working correlation structures can be exactly represented or approximated by a linear combination of basis matrices, i.e.,

$$R^{-1} \approx \sum_{i=1}^{k} a_i T_i, \tag{3}$$

where T_1 is the identity matrix, $T_2 \ldots, T_k$ are basis matrices which are determined by the structure of $R(\alpha)$, and $a_1, \ldots, a_k$ are constant coefficients.

With this representation, the GEE defined in (2) becomes a linear combination of the elements of the following extended score vector

$$\overline{g}_N(\beta) = \frac{1}{N}\sum_{i=1}^{N} g_i(\beta) \approx \frac{1}{N}\begin{pmatrix} \sum_{i=1}^{N} \dot{\mu_i}^T A_i^{-1/2} T_1 A_i^{-1/2}(Y_i - \mu_i) \\ \vdots \\ \sum_{i=1}^{N} \dot{\mu_i}^T A_i^{-1/2} T_k A_i^{-1/2}(Y_i - \mu_i) \end{pmatrix}. \tag{4}$$

As the dimension of the above extended score is greater than the number of unknown parameters, not all of the components of the extended score vector can be made zero simultaneously. Qu et al. (2000) applied the method of GMM to obtain an estimator of β by minimizing the following quadratic inference function

$$Q_N(\beta) = N\overline{g}_N(\beta)^T C_N^{-1}(\beta)\overline{g}_N(\beta), \tag{5}$$

where $C_N(\beta) = N^{-1}\sum_{i=1}^{N} g_i(\beta) g_i^T(\beta)$ is the sample covariance matrix of g_i.

It is worth noting that estimation of the linear coefficients a_i's are not needed since the quadratic function in (5) does not involve nuisance parameters, see Section 3.3 for specific

examples. In addition, the QIF yields an optimal estimator in the class of estimating functions which are linear combinations of the elements of the extended score vector in (4). Since GEE is a special case in this class, the QIF has the potential to improve the efficiency of GEE when the working correlation is misspecified.

3 BIC based on the quadratic inference function

3.1 The BIQIF criterion

As in the typical model selection setting, we consider choosing an appropriate marginal regression model from a class of candidate models that corresponding to selecting different subsets of covariates. We assume that the intercept is always in the model, which of course can be relaxed if needed. Therefore there are at most 2^q candidate models. Let $\mathcal{M}$ be the class of candidate models. Each member of $\mathcal{M}$ can be identified with a unique set m, where m is a subset of $\{1, \ldots, q\}$ and contains the indices of the covariates that are included in that candidate model.

Schwarz' BIC (1978), originally motivated from a Bayesian point of view, selects the model that maximizes $\widehat{l}_p - \frac{p}{2}(\log n)$, where $\widehat{l}_p$ is the log-likelihood of the model with dimension p and n denotes the sample size. We modify Schwarz' BIC by replacing the negative two times loglikelihood with the QIF for the purpose of model selection. This is motivated by the observation that the role played by QIF in this semiparametric setting is similar to the one played by negative two times loglikelihood in the parametric setting. In fact, several key properties of parametric likelihood have been shown to carry over to the quadratic inference function. For instance, it allows for a likelihood ratio type test: $Q_N(\beta_0) - Q_N(\widehat{\beta})$ has an asymptotical chi-squared null distribution for testing the hypothesis $\beta = \beta_0$, where $\widehat{\beta}$ denotes the unrestricted estimator (see Lindsay and Qu, 2003, for a general review of the QIF from the perspective of the inference function).

For a candidate model indexed by m, let $\beta(m)$ denote the $(q+1) \times 1$ vector which sets the corresponding components of β to zero if they are not selected by this model. To

illustrate the notation, consider fitting a marginal regression model with three possible covariates x_1, x_2, x_3, then the candidate model that contains only x_1 and x_3 will be indexed by $m = \{1, 3\}$, and $\beta(m) = (\beta_0, \beta_1, 0, \beta_3)^T$. Furthermore, let $\mathcal{B}(m)$ be the corresponding parameter space, i.e., $\mathcal{B}(m) = \{\beta \in \mathcal{B} : \beta = \beta(m)\}$, a subset of the parameter space $\mathcal{B}$. The QIF-based BIC selects the model in $\mathcal{M}$ which minimizes:

$$\text{BIQIF}(m) = Q_N(\widehat{\beta}(m)) + |\beta(m)| \log(N), \tag{6}$$

where $Q_N(\widehat{\beta}(m)) = \inf_{\beta \in \mathcal{B}(m)} Q_N(\beta)$; i.e., $\widehat{\beta}(m)$ is the estimated marginal regression parameter when $\beta \in \mathcal{B}(m)$, and $|\beta(m)|$ denotes the number of nonzero elements in $\beta(m)$. In the linear model case, $|\beta(m)|$ represents the dimension of the model.

Note that the interpretation of the model selection criterion (6) is similar to that of AIC and BIC: the first term on the right-hand side penalizes lack of fit of the model and the second term penalizes the complexity of the model, and m is chosen to minimize the sum of these two penalties.

3.2 The consistency property

Schwarz' BIC, which was originally motivated to select a model to maximize the posterior model probability, has the well-known consistency property: it selects the true model, if it is among the candidate models, with probability approaching one as the sample size goes to infinity. This was rigorously established by Nishii (1984) in the setting of parametric normal regression, see also Haughton (1988). We will demonstrate that the semiparametric criterion BIQIF proposed in Section 3.1 retains this important property for selecting the marginal regression model.

To formally state this property, we first introduce some notation. Let $\mathcal{M}^c$ be the subset of $\mathcal{M}$ that contains the correct models, i.e.,

$$\mathcal{M}^c = \{m \in \mathcal{M} : g(E(y_{ij}|x_{ij})) = x_{ij}^T \beta(m), \quad \text{for some } \beta(m) \in \mathcal{B}(m)\}.$$

There may exist more than one correct model, for example, if a linear model is correct then a quadratic model is also correct. We therefore need to consider the class of the most parsimonious correct models, defined as

$$\mathcal{PM}^c = \{m \in \mathcal{M}^c : |\beta(m)| \leq |\beta(m^*)|, \ \forall \ m^* \in \mathcal{M}^c\}.$$

In many situations, $\mathcal{PM}^c$ has a unique element, for instance, when $\mathcal{M}$ consists of a sequence of nested models which includes the true model.

Denote the model selected by BIQIF from $\mathcal{M}$ by $\widetilde{m}$. The following theorem describes the consistency property of BIQIF.

Theorem 3.1 *Under assumptions 1-4 in Appendix A,*

$$P(\widetilde{m} \in \mathcal{PM}^c) \to 1$$

when $N \to \infty$.

Theorem 1 immediately implies that, with probability approaching one, the BIQIF procedure selects the most parsimonious correct model. Proof of this theorem is given in the appendix. The proof also shows that in case the class of candidate models does not contain any correct model, the first term of the BIQIF asymptotically dominates the second term and as a result BIQIF selects the model within that class with the smallest quadratic inference function value. To better understand this, a geometrical interpretation of the quadratic inference function from the projection point of view can be found in Lindsay and Qu (2003).

3.3 Monte Carlo simulations

We compare, via numerical simulations, the BIQIF with several alternative methods: the AIC and BIC procedures based on the full likelihood; the Z-test procedure; and the QIC procedure of Pan (2001). To reduce the computational burden, we use the true correlation matrix for computing AIC, BIC, and the Z-test procedure which fits the full model and

throws out all variables with p-values greater than 0.05. The correlated binary data in Example 2 below are simulated from Bahadur's representation, see Fitzmaurice (1995). The AIC and BIC procedures based on the full likelihood serve as benchmarks.

In the examples below, the extended score vector for calculating BIQIF is

$$\frac{1}{N}\begin{pmatrix} \sum_{i=1}^{N} \dot{\mu_i}^T A_i^{-1/2} T_1 A_i^{-1/2}(Y_i - \mu_i) \\ \sum_{i=1}^{N} \dot{\mu_i}^T A_i^{-1/2} T_2 A_i^{-1/2}(Y_i - \mu_i) \end{pmatrix},$$

where T_1 is the identity matrix. For the exchangeable (or CS) correlation, T_2 has zeros on the diagonal and all off-diagonal elements equal to one; and for the AR-1 correlation, T_2 has one on the two main off-diagonals and zero elsewhere, see Qu et al. (2000).

Example 1: continuous responses. The response variable is generated by

$$y_{ij} = \beta_0 + \beta_1 x_{1,ij} + \beta_2 x_{2,ij} + \beta_3 x_{3,ij} + \epsilon_{ij}, \quad i = 1, \ldots, N \text{ and } j = 1, \ldots, 4,$$

where $(x_{1,i1}, \ldots, x_{1,i4})^T$, $(x_{2,i1}, \ldots, x_{2,i4})^T$ and $(x_{3,i1}, \ldots, x_{3,i4})^T$ are independently generated from a multivariate normal distribution with mean $(0.1, 0.2, 0.3, 0.4)^T$ and an identity covariance matrix, and the $(\epsilon_{i1}, \ldots, \epsilon_{i4})^T$ are generated independently of the covariates from a four-dimensional normal distribution with mean 0, marginal variance 1 and an AR(1) correlation structure with auto-correlation coefficient α. The true model has $\beta_0 = 0.2$, $\beta_1 = \beta_2 = 1$ and $\beta_3 = 0$. We consider all possible eight candidate models that correspond to different subsets of the three covariates with the intercept included.

Table 1 summarizes the proportion of times that the true model is selected out of 500 simulation runs for three different sample sizes: $N = 40, 80, 120$, and three different values of α: 0.3, 0.5 and 0.7.

Put Table 1 about here

Among these methods, the Z-test procedure consistently yields the poorest performance. The performance of the BIQIF using either the AR-1 working correlation matrix or the CS working correlation matrix is quite satisfactory. For sample size $N = 80$ and $\alpha = 0.7$,

the estimated probability for the BIQIF to select the true model is above 95%, which is close to the ideal BIC performance (97.2%) when the full likelihood is completely known. The QIC performs similarly as the BIQIF for $N = 40$ but becomes less competitive when the sample size increases. We note that the AIC does not yield a high probability of selecting the true model. This is due to the intrinsic difference between the AIC and BIC type criteria, see Yang (2005) for related discussion.

Example 2: discrete responses. The simulation setting is the same as that of Pan (2001). The response variable Y_{ij} is binary and its marginal mean μ_{ij} is

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 x_{1,ij} + \beta_2(j-1) + \beta_3 x_{3,ij} + \beta_4 x_{4,ij}, \quad i = 1, \ldots, N \text{ and } j = 1, 2, 3,$$

where $x_{1,ij}$ are iid Bernoulli with probability 0.5, $x_{3,ij}$ and $x_{4,ij}$ have iid uniform distribution on the interval (-1,1) and are independent of $x_{1,ij}$. The correlation matrix is exchangeable with correlation $\rho = 0.5$. The true model has $\beta_0 = 0.25$, $\beta_1 = \beta_2 = -0.25$ and $\beta_3 = \beta_4 = 0$. As in Pan (2001), five non-nested candidate models are considered.

The performance of different methods for two different sample sizes, $N = 100$ and 200, is summarized in Table 2.

Put Table 2 about here

Similarly as in Example 1, the performance of the model selection procedure based on the Z-test is poor. For sample size 100, the performance of BIQIF and that of QIC are similar. The estimated probability that the full likelihood-based BIC selects the true model is less than that of the full likelihood-based AIC. This indicates that the sample size 100 is still not large enough for the asymptotic consistency to kick in. Indeed, it takes a larger sample size to observe the consistency in a discrete case than in a continuous case. For sample size 200, the advantage of the BIQIF becomes evident, and its performance is close to that of the full likelihood-based BIC. The table also suggests that although the probability of the BIC-type procedures choosing the true model is higher than that of AIC, the BIC-type procedures also tend to select simpler models in finite samples.

4 Data-driven test for model checking

The null hypothesis for checking the goodness-of-fit of a postulated model states that the conjectured model provides an adequate fit to the data, i.e.,

$$H_0 : \ g(E(y_{ij}|x_{ij})) = x_{ij}^T\beta, \quad \text{for some } \beta \in \mathcal{B}. \tag{7}$$

In general, the alternative is assumed to be smooth and embed the null hypothesis,

$$H_a : \ g(E(y_{ij}|x_{ij})) = x_{ij}^T\beta + \sum_{k=1}^{d} \theta_k b_k(x_{ij}), \tag{8}$$

where $b_1, \ldots, b_d$ are basis functions. For example, for testing whether a certain covariate has an effect or not, the basis functions are often chosen to be orthogonal Legendre polynomials (see, for example, Szegö, 1975) of that covariate.

4.1 Data-driven smooth test

We consider a two-step data-driven smooth test, which generalizes the ideas of Ledwina (1994) and Kallenberg and Ledwina (1997) in an independent data setting. Compared with existing goodness-of-fit tests for GEE, the new test has the advantage of not restricting the power of the test to a specific direction, since a specific alternative might be difficult to validate and can cause serious loss of power if it is misspecified.

First, a data-driven alternative that best describes the complexity of the data is selected from l candidate alternatives (usually corresponding to letting $d = 1, \ldots, l$ in (8)). For example, to test whether a covariate has any effect, the candidate alternatives might include the linear model, the quadratic model and up to d-th order polynomial model of that covariate. The candidate alternatives are compared using the BIQIF criterion and the one leading to the smallest BIQIF value is chosen. The selected alternative dimension is denoted by $\widehat{d}$.

Next, with the chosen alternative, we apply a QIF-based score-type test T_d in (12) (see

Appendix B). That is, set the d in the alternative (8) to be $\widehat{d}$ and calculate $T_{\widehat{d}}$. The test proceeds as if the alternative is fixed. Other tests, such as the QIF-based likelihood ratio test, can also be applied. The score-type test, however, possesses the computational advantage that only the parameter under the null hypothesis needs to be estimated.

Under the null hypothesis (7), all candidate alternatives contain the true model and therefore are all correct. With probability approaching one, the BIQIF criterion selects the most parsimonious correct alternative, i.e., the one corresponding to $d = 1$. This results in $T_{\widehat{d}}$ converging in distribution to a chi-squared distribution with one degree of freedom, see the following theorem.

Theorem 4.1 *Assume conditions 1-3 in the Appendix A, under* H_0*,*

$$T_{\widehat{d}} \to \chi^2_1$$

in distribution when $N \to \infty$*.*

In applications we suggest applying a finite sample approximation of the null distribution of $T_{\widehat{d}}$ (Kallenberg and Ledwina, 1999) based on the second order expansion. More specifically, the probability $P(T_{\widehat{d}} \leq x)$ under H_0 can be approximated by

$$\begin{cases} \{2\Phi(\sqrt{x}) - 1\}\{2\Phi(\sqrt{\log N}) - 1\} & \text{if } x \leq \log N \\ \{2\Phi(\sqrt{x}) - 1\}\{2\Phi(\sqrt{\log N}) - 1\} + 2\Phi(-\sqrt{\log N}) & \text{if } x \geq 2\log N \\ \text{interpolate} & \text{if } \log N < x < 2\log N \end{cases}$$

where $\Phi(\cdot)$ stands for the cumulative distribution function of a standard normal random variable, and "interpolate" means that for any $\log N < x < 2\log N$, the approximated probability is obtained as a linear interpolation of the two values evaluated at $\log N$ and $2\log N$.

4.2 A Simulated Example

We consider data generated from $y_{ij} = 0.2 + x_{1,ij} + h(x_{2,ij}) + \epsilon_{ij}$, $i = 1, \ldots, 200$, $j = 1, \ldots, 8$, where $(x_{1,i1}, \ldots, x_{1,i8})^T$ has a multivariate normal distribution with mean zero and an

identity covariance matrix; $x_{2,ij}$, $j = 1, \ldots, 8$ are independent uniform random variables on (-1,1), and the random errors $(\epsilon_{i1}, \ldots, \epsilon_{i8})^T$ are independent of the covariates and have a multivariate normal distribution with mean 0, marginal variance 1 and an AR(1) correlation structure with auto-correlation coefficient 0.5.

We test the null hypothesis that $h(\cdot)$ is constant, i.e, the covariate x_2 does not have any effect on the response in the marginal regression. Four different alternatives for $h(\cdot)$ are considered: (1) linear alternative $h(x) = \theta x$, $\theta \in [0, 0.2]$; (2) quadratic alternative $h(x) = \theta x^2$, $\theta \in [0, 0.4]$; (3)cosine alternative $h(x) = \theta cos(0.5\pi x)$, $\theta \in [0, 0.4]$; (4) exponential alternative $h(x) = \theta exp(x^2)$, $\theta \in [0, 0.2]$. Note that $\theta = 0$ corresponds to H_0.

Figure 1 illustrates the estimated power curves (based on 500 runs) corresponding to the data-driven smooth test and the GEE Z-test. For the data-driven smooth test, we use $l = 6$. We have tried different values for l and observed very similar results. This conforms with the finding of Kallenberg and Ledwina (1997, 1999) in independent data settings that the data-driven test has stable power behavior with respect to l when a Schwarz-type criterion is applied. For computational simplicity and to save space, we use and report results using a working independence correlation matrix for the score-type test and the true correlation structure for the Z-test.

Put Figure 1 about here

Both tests display accurate levels under H_0 but their power performance differs dramatically. The Z-test is more powerful than the data-driven smooth test with respect to the linear hypothesis because it is designed to have higher power when the alternative direction is specified correctly; however its power breaks down for the other three alternatives where the data-driven smooth test exhibits high power. Very similar patterns are observed if other working correlation structures are used for the data-driven test, and the power performance is even better if the true correlation structure is applied.

5 Application to a clinical trial set

We now apply the proposed procedures to data from a clinical trial comparing two treatments for a respiratory disorder (Stokes, Davis, and Koch, 1995). There are 111 patients, who are randomly assigned to receive either the active treatment or the placebo. The response variable is the respiratory status (coded here as 0=poor, 1=good), which is measured for each patient on four scheduled visits. Five explanatory variables are recorded: gender (binary), treatment (binary), clinic center (binary), age at the beginning of the study (continuous) and baseline respiratory status (binary).

We consider 32 candidate models that include all possible subsets of the five covariates. The best model chosen by BIQIF (using AR-1 working correlation) includes intercept, treatment and baseline respiratory status, which yields BIQIF = 27.025; while the second-best model contains intercept, treatment, center and baseline respiratory status, which yields BIQIF = 28.053. Table 3 gives the top four models chosen by the BIQIF using AR-1 working correlation (these top four choices remain the same if the CS working correlation is used). The table also reports the estimated coefficients and p-values for each of these four models using both the QIF and the GEE (for which the p-values are reported via a robust variance estimation). We note that the QIC (using either the AR-1 or CS working correlation) selects a different best model that includes intercept, treatment, center, age and baseline respiratory status. This model is ranked number 4 by the BIQIF.

Put Table 3 about here

We next test whether age has any effect on the response if the intercept, treatment and baseline respiratory status are included for the β part in (8). As there are only four measurements for each patient, the most complex polynomial model for age effect we could use is a cubic one. We thus perform the data-driven smooth test with $l = 3$ and obtain a p-value 0.001, which suggests that age effect is significant. In contrast, the GEE Z-test (with linear alternative) gives a p-value of 0.21. A closer look at the output of

the data-driven smooth test shows that dimension two ($l = 2$) is selected; thus age has a quadratic effect. In fact, if we repeat the Z-test with both age and age^2 included, both of them are significant. This show that the Z-test is not powerful when a linear alternative is used, while the actual age effect is likely to be nonlinear. Finally, based on the new information obtained from model checking, we repeat the above model selection process with age^2 added, and the gender effect is left out as it was not significant in the previous analysis. The best model obtained now includes intercept, treatment, baseline, age and age^2, based on our model selection and model checking procedures.

6 Discussion

We have proposed a quadratic inference function based BIC-type criterion for selecting the marginal regression model for longitudinal data analysis and a data-driven score-type test for checking the validity of a proposed marginal model. They are useful for analyzing the relationship between the marginal expectation and the covariates. If primary interest is focused on subject-specific effects, then random or mixed effects models are probably more appropriate; but this is not addressed in the current paper.

Although a working correlation matrix needs to be assumed, the quadratic inference function does not include any nuisance parameters. This is different from GEE, where to estimate β, one also needs to estimate the nuisance parameters in the working correlation matrix. The BIQIF only depends on β and the model dimension. Moreover, even if the working correlation is misspecified, BIQIF remains consistent.

Parallel to the construction in this paper, one may propose a criterion similar to AIC by modifying the penalty term. As is well known, AIC and BIC have different strengths: BIC is consistent in selecting the true model and AIC aims at minimizing the Kullback-Leibler divergence between the true model and the estimated model. The AIC criterion usually yields minimax-rate optimal estimators for the regression function under a squared-error loss (see Yang and Barron, 1999; and Yang, 2005). Another important future research topic is to consider covariance matrix selection, as in Pan (2001), although Pan's method

is based on the quasilikelihood approach.

Acknowledgements

We would like to thank the referee, the Associate Editor and the Joint Editor for their helpful comments that significantly improved the quality of the paper. We also thank Wei Pan and Yuhong Yang for their insightful suggestions for an earlier version of the paper. Wang's research is supported by NSF grant DMS-0706842 and Qu's research is supported by NSF grant DMS-0348764.

References

Andrews, D. W. K. (1999) Consistent moment selection procedures for generalized method of moment selection. *Econometrica*, **67**, 543-564.

Balan, R. M. and Schiopu-Kratina. (2005) Asymptotic results with generalized estimating equations for longitudinal data. *Ann. Statist.*, **32**, 522-541.

Barnhart, H. X. and Williamson, J. M. (1998) Goodness-of-fit tests for GEE modeling with binary data. *Biometrics*, **54**, 720-729.

Billingsley, P. (1995). *Probability and Measure*, 3rd ed. New York: John Wiley & Sons.

Burnham, K. P. and Anderson, D. R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed, New York: Springer-Verlag.

Cantoni, E., Flemming, J. M. and Ronchetti, E. (2005) Variable selection for marginal longitudinal generalized linear models. *Biometrics*, **61**, 507-514.

Chiou, J-M. and Müller, H-G. (2005) Estimated estimating equations: semiparametric inference for clustered and longitudinal data. *J. R. Statist. Soc.* B, **67**, 531-553.

Diggle, P., Heagerty, P., Liang, K.-Y. and Zeger, S. (2002) *Analysis of Longitudinal Data*, 2nd edition. New York: Oxford University Press.

Fitzmaurice, G. M. (1995) A caveat concerning independence estimating equations with multivariate binary data. *Biometrics*, **51**, 309-317.

Hall, D. B. and Severini, T. A. (1998) Extended generalized estimating equations for clustered data. *J. Am. Statist. Ass.*, **93**, 1365-1375.

Hansen, L. (1982) Large sample properties of generalized method of moments estimators. *Econometrica*, **50**, 1029-54.

Hardin, J. W. and Hilbe, J. M. (2003) *Generalized Estimating Equations.* Chapman & Hall/CRC.

Hart, J. (1997) *Nonparametric Smoothing and Lack-of-fit Test.* New York: Springer-Verlag.

Haughton, D. (1988) On the choice of a model to fit data from an exponential family. *Ann. Statist.*, **16**, 342-355.

Kallenberg, W. C. M. and Ledwina, T. (1997) Data-driven smooth tests when the hypothesis is composite. *J. Am. Statist. Ass.*, **92**, 1094-1104.

Kallenberg, W. C. M. and Ledwina, T. (1999) Data-driven rank tests for independence. *J. Am. Statist. Ass.*, **94**, 285-301.

Kolaczyk, E. D. (1995) An information criterion for empirical likelihood with general estimating equations. *Technical Report.* Department of Statistics, University of Chicago, Chicago.

Ledwina, T. (1994) Data-driven version of Neyman's smooth test of fit. *J. Am. Statist. Ass.*, **89**, 1000-1005.

Liang, K. Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalised linear models. *Biometrika* **73**, 12-22.

Lindsay, B. G. and Qu, A. (2003) Inference functions and quadratic score tests. *Statist. Sci.*, 394-410.

McQuarrie, A. D. R. and Tsai, C.-L. (1998) *Regression and Time Series Model Selection.* River Edge, NJ: World Scientific Publishing Company.

Miller, A. J.(1990) *Subset Selection in Regression.* London: Chapman and Hall.

Neyman, J. (1937) *Skand. Aktuar. Tidskr.*, **20**, 149-199.

Nishii, R. (1984) Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, **12**, 758-765.

Pan, W. (2001) Akaike's information criterion in generalized estimating equations. *Biometrics*, **57**, 120-125.

Pan, W. (2002) Goodness-of-fit tests for GEE with correlated binary data. *Scand. J. Statist.*, **29**, 101-110.

Pan, J. and Mackenzie, G. (2003) On modelling mean-covariance structures in longitudinal studies. *Biometrika*, **90**, 239-244.

Prentice, R. L. and Zhao, L. P. (1991) Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, **47**, 825-839.

Qu, A., Lindsay, B. G. and Li, B. (2000) Improving generalised estimating equations using quadratic inference functions. *Biometrika*, **87**, 823-836.

Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461-464.

Shao, J. (1997) An asymptotic theory for linear model selection. *Statist. Sin.*, **7**, 221-264.

Shults, J. and Chaganty, N. R. (1998) Analysis of serially correlated data using quasi-least squares. *Biometrics*, **54**, 1622-1630.

Small, D. (2002) Inference and model selection for instrumental variables regression. *PhD Thesis*. Stanford University, Stanford.

Stokes, M. E., Davis, C. S. and Koch, G. G.(1995) *Categorical Data Analysis Using the SAS System*. Cary, NC: SAS Institute Inc.

Stoner, J. A. and Leroux, B. G. (2002) Analysis of clustered data: A combined estimating equations approach. *Biometrika*, **89**, 567-578.

Szegö, G. (1975) *Orthogonal Polynomials*, 4th ed. Providence, RI: Amer. Math. Soc.

Wang, Y.-G. and Carey, V. (2003) Working correlation structure misspecification, estimation and covariate design: Implications for generalised estimating equations performance.

Biometrika, **90**, 29-41.

Yang, Y. (2005) Can the strengths of AIC and BIC be shared? *Biometrika*, **92**, 937-950.

Yang,Y. and Barron, A. R. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, **27**, 1564-1599.

Appendix A: Proof of Theorems

For transparency of exposition, we follow the notation of Liang and Zeger (1986) and Balan and Schiopu-Kratina (2005) and assume that there are N subjects and $n_i = n$ repeated measurements on each subject. We first state below a set of regularity conditions that consist of the following four assumptions.

Assumption 1. (X_i, Y_i), $i = 1, \ldots, N$, are independent and identically distributed from an $n \times (q+2)$-dimensional distribution. The marginal regression parameter in (1) is identifiable, i.e., $E(g(E(y_{ij}|x_{ij})) - x_{ij}^T\beta) = 0$, $j = 1, \ldots, n$, if and only if $\beta = \beta_0$.

Assumption 2. The domain $\mathcal{B}$ is a compact subset of R^q and the true parameter value β_0 is in its interior.

Assumption 3. Let $\mu_i = (\mu_{i1}, \ldots, \mu_{in})^T$ where $\mu_{ij} = E(y_{ij}|x_{ij})$, then μ_i is almost surely continuously differentiable in β. Denote this $n \times (q+1)$ derivative matrix by $\dot{\mu}_i$, then $\dot{\mu}_i$ has full column rank $q+1$ almost surely.

Assumption 4. For $m \in \mathcal{M}$, $g_i(\beta(m))$ is almost surely (with respect to the data) continuously differentiable on $\mathcal{B}(m)$. There exists a function with finite expectation that dominates both $g_i(\beta(m))g_i(\beta(m))^T$ and the derivative matrix $\dot{g}_i(\beta(m))$. Conditional on the covariate, the weighting matrix $C_N(\beta(m)) = N^{-1}\sum_{i=1}^N g_i(\beta(m))g_i^T(\beta(m))$ converges in probability to a positive definite matrix $C(\beta(m))$.

These conditions are essentially the same as those required to establish the asymptotic normality of the GEE estimator and those given in Hansen (1982) for establishing the asymptotic normality of a generalized method of moments estimator.

Proof of Theorem 3.1. The proof consists of two steps, similar in spirit to Nishii (1984) and Andrews (1999).

Step 1. Consider any $m \in \mathcal{M}$ with $m \notin \mathcal{M}^c$. Thus $(g(E(y_{i1}|x_{i1})), \ldots, g(E(y_{in}|x_{in}))) \neq (x_{i1}^T\beta(m), \ldots, x_{in}^T\beta(m))$, for any $\beta(m) \in \mathcal{B}(m)$. Then conditional on the observed covariate matrix X, $E(\overline{g}_N(\beta(m))|X) \neq 0$ for any $\beta(m) \in \mathcal{B}(m)$ since it contains ${\dot{\mu}_i}^T A_i^{-1} E(y_i - \mu_i|X_i)$ and $\dot{\mu}_i$ has full row rank q. We thus have,

$$\begin{aligned} \frac{\text{BIQIF}(m)}{N} &= \inf_{\beta \in \mathcal{B}(m)} \overline{g}_N(\beta)^T C_N^{-1}(\beta) \overline{g}_N(\beta) + |\beta(m)| \frac{\log N}{N} \\ &\to \inf_{\beta \in \mathcal{B}(m)} E(\overline{g}_N(\beta(m))^T|X) C^{-1}(\beta(m)) E(\overline{g}_N(\beta(m))|X)' > 0 \end{aligned} \tag{9}$$

in probability by the Uniform Law of Large Numbers and continuous mapping theorem (Billingsley, 1995) and the fact that $E(\overline{g}_N(\beta(m))|X) \neq 0$ and $C(\beta(m))$ is positive definite.

Step 2. Consider any $m \in \mathcal{M}^c$. There exists a $\beta(m) \in \mathcal{B}(m)$ such that $E(\overline{g}_N(\beta(m))) = 0$. A similar argument as in Hansen (1982) establishes that $\overline{g}_N(\beta(m)) = O_p(N^{-1/2})$. Therefore,

$$\begin{aligned} \frac{\inf_{\beta \in \mathcal{B}(m)} Q_N(\beta)}{N} &= \inf_{\beta \in \mathcal{B}(m)} \overline{g}_N(\beta)^T C_N^{-1}(\beta) \overline{g}_N(\beta) \\ &\to \inf_{\beta \in \mathcal{B}(m)} E(\overline{g}_N(\beta(m))^T|X) C^{-1}(\beta(m)) E(\overline{g}_N(\beta(m))|X)' \\ &= O_p(N^{-1}). \end{aligned} \tag{10}$$

Therefore,

$$\frac{\text{BIQIF}(m)}{N} = O_p(N^{-1}) + O(\log N/N) = o_p(1). \tag{11}$$

Comparison of (9) and (11) leads to $P(\widetilde{m} \in \mathcal{M}^c) \to 1$ as $N \to \infty$. Now $\forall\ m_1 \in \mathcal{M}^c$, $m_2 \in \mathcal{M}^c$, $m_1 \notin \mathcal{P}\mathcal{M}^c$ and $m_2 \in \mathcal{P}\mathcal{M}^c$, we will show that BIQIF will choose m_2 with probability approaching one. This is immediate from

$$\begin{aligned} &\frac{\text{BIQIF}(m_1)}{N} - \frac{\text{BIQIF}(m_2)}{N} \\ = \ &|\beta(m_1)| \frac{\log N}{N} - |\beta(m_2)| \frac{\log N}{N} + O_p(N^{-1}) > 0 \end{aligned}$$

with probability approaching one as $N \to \infty$ because $|\beta(m_1)| > |\beta(m_2)|$. □

Proof of Theorem 4.1. It follows directly from Theorem 3.1 that $P(\widehat{d} = 1) \to 1$ as $n \to \infty$. The proof is completed by noting that $T_1 \to \chi^2_1$ under H_0. □

Appendix B: QIF-based score-type test with a known alternative

For a given d, the null hypothesis H_0 in (7) reduces to testing

$$H_0 : \theta_1 = \ldots = \theta_d = 0.$$

Write $\psi = (\psi_1, \ldots, \psi_q)^T$, $\theta = (\theta_1, \ldots, \theta_d)^T$ and $\beta = (\psi^T, \theta^T)^T$. Let $S_1(\beta) = \frac{1}{2}\frac{\partial}{\partial \psi^T} Q_N(\beta)$ and $S_2(\beta) = \frac{1}{2}\frac{\partial}{\partial \theta^T} Q_N(\beta)$, where $Q_N(\beta)$ is the quadratic inference function defined in (5) when the marginal mean has the form (8). Furthermore, let $\widehat{\psi}$ be the estimator of ψ under H_0, thus $S_1(\widehat{\psi}, 0) = 0$. The score-type test for H_0 is then based on $S_2(\widehat{\psi}, 0)$ by observing that $S_2(\widehat{\psi}, 0)$ should be close to 0 under the null, but far away from 0 if H_0 is not true, that is, when part of θ is non-zero.

Using Taylor expansion and techniques similar as those in Qu et al. (2000) for the likelihood ratio test, we can show that $N^{-1/2} S_2$ has an asymptotic multivariate normal distribution:

$$N^{-1/2} S_2(\widehat{\psi}, 0) \to N_d(0, \Omega),$$

for some positive definite $d \times d$ matrix Ω, which can be consistently estimated by $\widehat{\Sigma}_{\theta\theta} - \widehat{\Sigma}_{\theta\psi}\widehat{\Sigma}^{-1}_{\psi\psi}\widehat{\Sigma}_{\psi\theta}$, where

$$\begin{pmatrix} \widehat{\Sigma}_{\psi\psi} & \widehat{\Sigma}_{\psi\theta} \\ \widehat{\Sigma}_{\theta\psi} & \widehat{\Sigma}_{\theta\theta} \end{pmatrix} = \frac{1}{2}\frac{\partial^2}{\partial^2 \xi^T} Q_N((\widehat{\psi}, 0)).$$

The score-type test can therefore be defined as

$$T_d = N^{-1} S_2(\widehat{\psi}, 0)^T \widehat{\Omega}^{-1} S_2(\widehat{\psi}, 0). \tag{12}$$

Under the H_0, T_d has an asymptotic χ^2_d distribution.

Table 1: The proportion of times the true model is selected out of 500 simulation runs.

		BIQIF^1	BIQIF^2	AIC	BIC	Z test	QIC^1	QIC^2
$N = 40$	$\alpha = 0.3$	0.854	0.886	0.844	0.944	0.504	0.836	0.842
	$\alpha = 0.5$	0.874	0.884	0.852	0.948	0.350	0.886	0.856
	$\alpha = 0.7$	0.868	0.870	0.842	0.948	0.326	0.906	0.890
$N = 80$	$\alpha = 0.3$	0.924	0.944	0.804	0.936	0.732	0.826	0.838
	$\alpha = 0.5$	0.944	0.936	0.856	0.962	0.644	0.870	0.846
	$\alpha = 0.7$	0.958	0.966	0.844	0.972	0.574	0.912	0.896
$N = 120$	$\alpha = 0.3$	0.968	0.950	0.844	0.964	0.870	0.850	0.850
	$\alpha = 0.5$	0.976	0.958	0.870	0.978	0.816	0.908	0.894
	$\alpha = 0.7$	0.970	0.968	0.866	0.964	0.728	0.922	0.904

Note: N is the number of subjects and α is the autocorrelation coefficient. In Table 1-3, BIQIF^1 is the BIQIF using the AR-1 working correlation, BIQIF^2 is the BIQIF using the CS working correlation, AIC and BIC are based on the full likelihood with known covariance matrix, the Z test uses the true covariance structure, QIC^1 is the QIC procedure using the AR-1 working correlation matrix, and QIC^2 is the QIC procedure using the CS working correlation matrix.

Table 2: The proportion of times each of five non-nested models is selected out of 500 simulation runs by different procedures. The true model has $\{x_1, x_2\}$.

		BIQIF^1	BIQIF^2	AIC	BIC	Z test	QIC^1	QIC^2
$N = 100$	x_1	0.314	0.318	0.144	0.414	0.006	0.176	0.190
	x_1, x_2	0.600	0.644	0.640	0.538	0.060	0.658	0.610
	x_1, x_3	0.034	0.014	0.036	0.014	0.000	0.026	0.090
	x_1, x_2, x_3	0.036	0.018	0.100	0.028	0.004	0.086	0.080
	x_1, x_2, x_3, x_4	0.016	0.006	0.080	0.006	0.000	0.054	0.030
$N = 200$	x_1	0.098	0.100	0.020	0.092	0.000	0.020	0.046
	x_1, x_2	0.868	0.876	0.768	0.888	0.208	0.818	0.812
	x_1, x_3	0.006	0.004	0.004	0.000	0.000	0.006	0.004
	x_1, x_2, x_3	0.024	0.016	0.138	0.020	0.016	0.100	0.090
	x_1, x_2, x_3, x_4	0.004	0.004	0.070	0.000	0.000	0.056	0.048

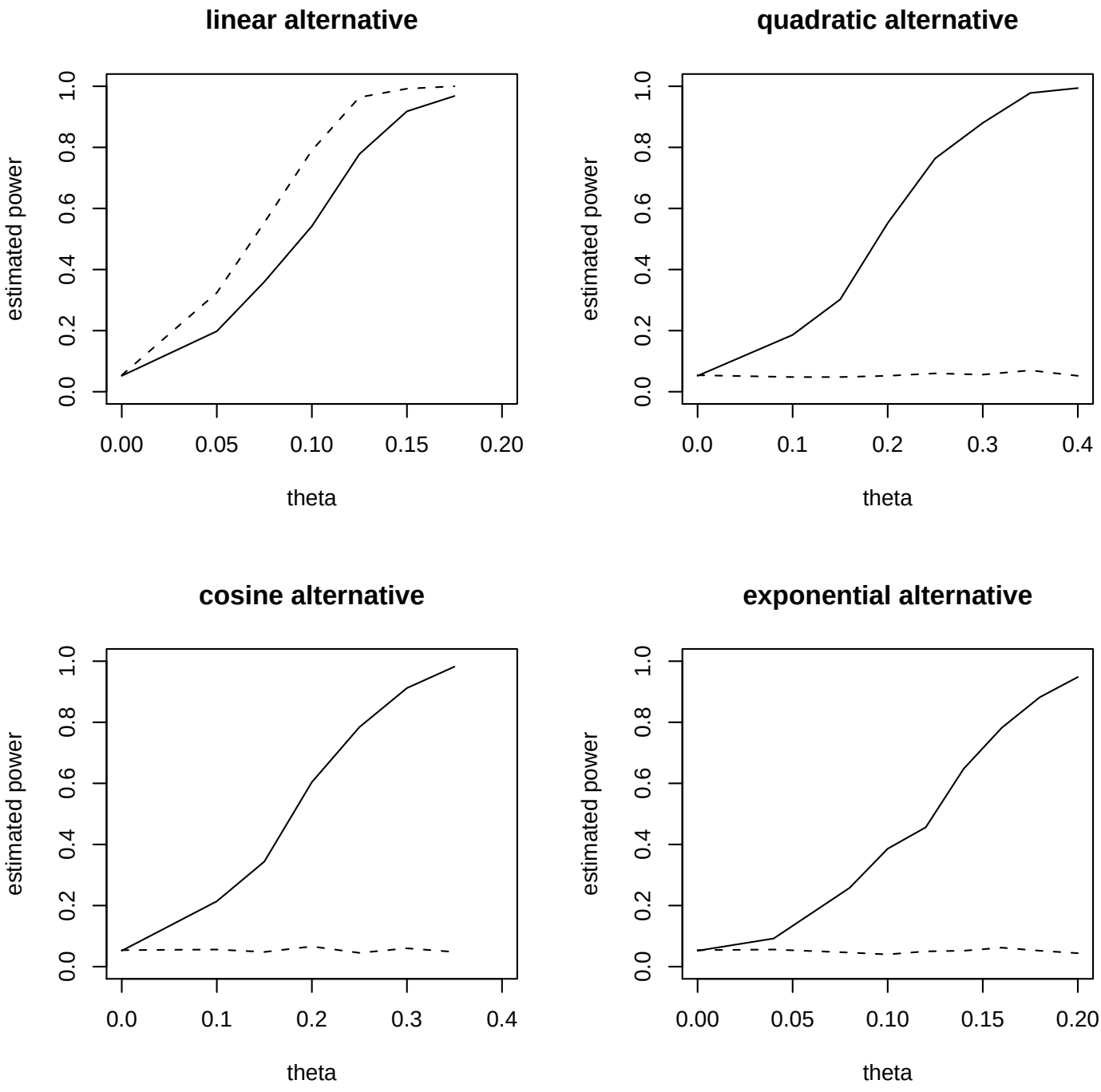

Figure 1: Simulated power curves of the data-driven smooth test (denoted by the solid curve) and the Z-test of GEE (denoted by the dashed curve) for four different alternatives.

Table 3: Top four models chosen by BIQIF for a clinical trial data set comparing two treatments for a respiratory illness. The entries are the estimated coefficients using QIF and GEE (with p-values in parenthesis). The last row gives the BIQIF values for the four models using an AR-1 working correlation matrix.

	Model 1		Model 2		Model 3		Model 4	
	QIF	GEE	QIF	GEE	QIF	GEE	QIF	GEE
trt	-1.215	-1.162	-1.225	-1.177	-1.238	-1.152	-1.236	-1.167
	(<0.001)	(0.001)	(0.001)	(<0.001)	(<0.001)	(<0.001)	0.002	(<0.001)
center	-	-	0.575	0.632	-	-	0.676	0.749
	-	-	(0.084)	(0.059)	-	-	(0.042)	(0.029)
age	-	-	-	-	-0.010	-0.009	-0.016	-0.016
	-	-	-	-	(0.519)	(0.440)	(0.342)	(0.168)
gender	-	-	-	-	-	-	-	-
baseline	2.030	2.030	1.991	1.896	2.064	2.022	1.966	1.859
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
BIQIF	27.025		28.053		30.513		30.614	