Institute of Mathematical Statistics

LECTURE NOTES — MONOGRAPH SERIES

ON THE CONSISTENCY OF GENERALIZED ESTIMATING EQUATIONS

By Bing Li
Pennsylvania State University

ABSTRACT

We study the consistency of generalized estimating equations. Our consistency result differs from the known results in two respects. First, it identifies a specific sequence of consistent solutions to be the minimax point of a deviance function; this is stronger than the known consistency results, which assert only the asymptotic existence of a consistent sequence. Second, the minimax procedure applies and gives consistent estimate even when the generalized estimating equation itself is not defined, as would be the case if the mean function is not differentiable, or if the support of the random observations depend on the parameters. We also provide two practical criteria based on which we can decide whether a solution is consistent by fairly simple computations.

Key words and phrases: Quasi likelihood estimation; generalized estimating equations; deviance; projected likelihood ratio; Doob-Wald approach to consistency.

1. INTRODUCTION

Many data sets that arise from scientific research consist of repeated, or clustered, measurements, which results in dependence among the observations. It is therefore necessary to incorporate the dependence into the estimation of the parameters and the assessment of errors. The generalized estimating equations and related techniques are very effective for such purposes. See Liang & Zeger (1986), and Zeger & Liang (1986). In the present paper we will establish the consistency of generalized estimating equations, and provide verifiable conditions under which consistency holds.

Applying Crowder's general theory (Crowder 1986; Small & McLeish 1994, page 96), it is not difficult to demonstrate that, with probability tending to one, a generalized estimating equation has a consistent solution. However, here we aim at a more specific statement of consistency; that is, a specific sequence of solutions, which can be identified in practice, is consistent. This latter statement is important because, in many applications, a generalized estimating equation may either have multiple solutions or have none at all; See McCullagh (1990), Hanfelt & Liang (1995), and Li (1996).

In the classical theory of maximum likelihood estimation, the more specific statement of consistency is known as the Doob-Wald statement, demonstrated by Doob (1934), Wald (1949), and Wolfowitz (1949). It is established via a property of likelihood functions, which states that the maximizer of expected log likelihood is the true parameter value. This implies that, under certain regularity assumptions, the maximizer of the averaged log likelihood converges in probability to the true parameter value.

For estimating equations, however, there is often no such likelihood functions to which the Doob-Wald argument can be directly applied. This is because, unlike a likelihood score function which is by definition the gradient of the log likelihood, an estimating equation need not be the gradient of any potential function, and hence we have no function to maximize. See McCullagh (1990), McCullagh & Nelder (1989), Firth & Harris (1991), and Li & McCullagh (1994). However, Li (1996) pointed out that (i) the likelihood property used in Doob-Wald argument can be restated as: the minimax of the expectation of the likelihood ratio is the true parameter value, (ii) for many estimating equations it is possible to construct a function which behaves similarly as the log likelihood ratio, so that the minimax of the expectation of this function is always the true parameter value. Based on these observations Li (1996) demonstrated that every minimax point of this function is necessarily a consistent solution. This leaves no ambiguity when the estimating equation has multiple solutions or has no solution at all. Moreover, the consistency of the minimax holds even when the estimating equation itself is not defined, as would be the case for quasi likelihood equation if the mean function is not differentiable or if the support of the random observations depend on the parameter.

In this note we will extend the minimax approach of Li (1996) to demonstrate the consistency of generalized estimating equations. In addition, we will provide two practical criteria by which one can judge, via fairly simple computation, whether a solution is consistent.

In §3 we demonstrate the consistency in the special cases in which the estimating equations do integrate to potential functions. In §4, we describe the minimax approach to consistency, and introduce a deviance function that will facilitate this approach. §5 contains all the technical preparations towards the proof of consistency: assumptions, lemmas, and examples. The consistency in general cases will then be demonstrated in §6. In §7, we provide two practical criteria for consistency. Finally, in §8, a numerical example is carried out to illustrate the use and effectiveness of these criteria.

2. GENERALIZED ESTIMATING EQUATIONS AND E-ANCILLARITY

The data sets with which we shall be concerned are of the form $\{Y_{it} : t = 1, ..., n_i, i = 1, ..., K\}$. The observations are independent for different i, but dependent for different t within the same i. For example, $\{Y_{it} : t = 1, ..., n_i\}$ may be the measurements from the same subject at different times. Associated with each Y_{it} are a p-dimensional explanatory variable X_{it} and a p-dimensional regression parameter β, which together determine the expectation of Y_{it}. We will denote the whole vector $\{Y_{it}\}$ by Y, and the sample space of Y by $\mathcal{Y}$.

Let $\mu_{it}(\beta)$ and $\phi^{-1}V_{it}(\beta)$ be the mean and variance of the observation Y_{it}. We shall assume that $\mu_{it}(\beta) = \mu(X_{it}^T\beta)$ and $V_{it}(\beta) = V(X_{it}^T\beta)$ for some known functions $\mu(\cdot)$ and $V(\cdot)$. The dispersion parameter ϕ is always taken to be positive. The inverse of μ is often called the link function, and $X_{it}^T\beta$ the linear predictor. Typically, we take $\mu^{-1}(\cdot)$ to be the natural link function from a linear exponential family and $V(\cdot)$ to depend on μ according to that family, but in principal they can be other functions as well. See McCullagh & Nelder (1989). The dependence within each i is modeled by correlation matrices $R(\alpha) = \{R_{tt'}(\alpha) : t, t' = 1, ..., n_i\}$, where α is an s dimensional parameter and $R_{tt'}(\cdot)$ are known functions. Across i, $R_{tt'}(\alpha)$ is assumed to remain constant, in so far as both observations Y_{it} and $Y_{it'}$ are present in the cluster i. In other words, $R(\alpha)$ is the same for all i except for its dimension. In practice, the matrices $R(\alpha)$ need not be correctly assumed, and is called the working correlation matrices. Whether or not $R(\alpha)$ is correctly assumed, we can formally calculate the covariance matrices of $Y_i = \{Y_{it} : t = 1, ..., n_i\}$ based on $R(\alpha)$, as

$$W_i(\beta, \phi, \alpha) \equiv \phi^{-1}\{V_i(\beta)\}^{\frac{1}{2}} R(\alpha) \{V_i(\beta)\}^{\frac{1}{2}}.$$

These are called the working covariance matrices.

If the working correlation assumption were correct, then for each fixed ϕ and α, the optimal linear combination of $\{Y_{it} - \mu_{it}(\beta)\}$ which yields the highest information about β, in the sense of Godambe (1960), is

$$q(\beta, \phi, \alpha) = \phi^{-1} \sum_{i=1}^{K} \{\dot{\mu}_i(\beta)\}^T \{W_i(\beta, \phi, \alpha)\}^{-1} \{Y_i - \mu_i(\beta)\} = 0, \quad (1)$$

where $\mu_i(\beta)$ is the $n_i \times 1$ vector $\{\mu_{it}(\beta) : t = 1, ...n_i\}$, $\dot{\mu}_i(\beta)$ is the $n_i \times p$ dimensional gradient matrix of $\mu_i(\beta)$. Equation (1), considered as the estimating equation for β given ϕ and α, combined with any $\sqrt{K}$-consistent estimate $\hat{\phi}(\beta)$ of ϕ, and $\sqrt{K}$-consistent estimate $\hat{\alpha}\{\beta, \hat{\phi}(\beta)\}$ of α, is the generalized estimating equation for β. In other words, we estimate β by solving the equation

$$g(\beta) \equiv q\{\beta, \hat{\phi}(\beta), \hat{\alpha}(\beta, \hat{\phi}(\beta))\} = 0. \quad (2)$$

If $R(\alpha)$ is the true correlation matrix, and if α is known, equation (2) reduces to the quasi-likelihood equation as defined by Wedderburn (1974) and McCullagh (1983). In this case, it is known that any solution to (2), if it is consistent, is asymptotically normally distributed, and is efficient among the solutions to linear estimating equations. See also Jarrett (1984), McLeish (1984), Godambe & Heyde (1987).

In the sequel, we will denote the nuisance parameter $(\phi, \alpha^T)^T$ by γ, and the combined parameter $(\beta, \gamma^T)^T$ by θ. The parameter space for β will be written as B. Notice that if the working assumption is incorrect, then the nuisance parameter γ may have nothing to do with the underlying distribution P. Therefore we will use E_β to represent the expectation under the distribution P for which $\beta = \beta(P)$. We denote the true value of the regression parameter as β_0, and will often abbreviate E_{β_0} by E.

What is remarkable is the way in which the nuisance parameters γ enter into the equation (2). Specifically,

$$E_\beta\{q(\beta,\gamma)\} = 0 \quad \text{for all } \gamma. \tag{3}$$

Property (3) is called the E-ancillarity of the estimating equation q relative to the parameter γ; See Small & McLeish (1988, 1989). This is a fundamental ingredient of a generalized estimating equation, out of which arise most of the desirable properties of the method. In particular, as implicit in Liang & Zeger (1986), if $R(\cdot)$ is correct, substituting $\sqrt{K}$ consistent estimates of ϕ and α into (1) does not impair the efficiency of the estimate of β; even if $R(\cdot)$ is incorrect, in which case $\hat{\alpha}$ may not estimate anything, substituting $\hat{\phi}$ and $\hat{\alpha}$ into (1) does not impair the consistency and asymptotic normality of $\hat{\beta}$, as long as $\sqrt{K}(\hat{\alpha}-\alpha)$ is bounded in probability for some α. Property (3) also plays a vital role in our demonstration of consistency.

3. DOOB-WALD CONSISTENCY IN SPECIAL CASES

If an estimating equation does integrate to a potential function, then, under certain conditions, the Doob-Wald argument can be extended to verify that the global maximum of the potential function is consistent. In this section we sketch the proof of the consistency of quasi likelihood estimation in these simple cases. The argument can be extended to generalized estimating equation (2), provided that, with respect to the regression parameter β, it integrates to a potential function for each fixed γ. This extension is fairly straightforward and will be omitted. For further studies of the potential functions of estimating equations, see Li & McCullagh (1994).

Let $Y = \{Y_i : i = 1, ..., n\}$ be independent observations with mean $\mu_i(\beta)$ and variance $V_i(\beta)$ where, similarly as in §2, $\mu_i(\beta) = \mu(X_i\beta)$ and $V_i(\beta) = V(X_i\beta)$ for some known functions μ and V. Let $s(\beta, Y)$ be the quasi score function

$$s(\beta, Y) = \sum_{i=1}^{n} \dot{\mu}_i(\beta)V_i^{-1}(\beta)(Y_i - \mu_i(\beta)). \tag{4}$$

PROPOSITION 1. *Suppose that the quasi score integrates to some potential function $l(\beta, Y)$; in other words $(\partial l/\partial\beta)(\beta, Y) = s(\beta, Y)$, and that there is a measure ν with respect to which the integral of $e^{l(\beta,y)}$ does not depend on β; that is*

$$\int e^{l(\beta,y)}d\nu(y) = c. \tag{5}$$

Then, under certain regularity conditions, the global maximum of $l(\beta, Y)$ *is a consistent estimate of* β_0.

PROOF. The argument is similar in spirit to that used in Gourieroux, Monfort, and Trognon (1984). For each β, define a probability measure $dQ_\beta(y) = e^{l(\beta,y)}d\nu(y)/c$. Differentiating equation (5) with respect β, we find that $\mu(\beta) = \int y dQ_\beta(y)$. Since $l(\beta, y)$ is a linear function of y, it follows that $El(\beta, Y) = \int l(\beta, y)dQ_{\beta_0}(y)$ for all β. Hence

$$\begin{aligned} E\{l(\beta, Y) - l(\beta_0, Y)\} &= \int \{l(\beta, y) - l(\beta_0, y)\}dQ_{\beta_0}(y) \\ &< \log \int e^{l(\beta,y)-l(\beta_0,y)}dQ_{\beta_0}(y) \\ &= \log \int \{dQ_\beta(y)/dQ_{\beta_0}(y)\}dQ_{\beta_0}(y) = 0. \end{aligned}$$

That is, the expectation $El(\beta, Y)$ is maximized at β_0. The rest of the argument follows Wald (1949), with appropriate regularity conditions imposed to ensure that $n^{-1}l(\beta, Y)$ converges in probability to $n^{-1}El(\beta, Y)$ uniformly in β. □

4. A DEVIANCE FUNCTION AND A MINIMAX APPROACH

The argument of the last section depends on the existence of the potential function $l(\beta, Y)$, and is inapplicable if no such function exists, as is the case for many applications. Thus the questions arise: What is the crucial element in Doob-Wald argument? Must one have something like a likelihood function to maximize in order to apply this argument?

At first sight, the existence of a likelihood function seems essential: if there is a function $f(\beta, Y)$, whose expectation is uniquely maximized at β_0, and if $f(\beta, Y)$ converges in probability to $Ef(\beta, Y)$ uniformly in β, then the maximizer of $f(\beta, Y)$ will be a consistent estimator of the maximizer of $Ef(\beta, Y)$. But a more careful look at the Doob-Wald argument reveals that the maximization of a likelihood is not indispensable: if we can uniquely identify the true parameter value β_0 by examining the function $Ef(\beta, Y)$, may it be the maximum, the minimum, the turning point, or the minimax, then the $\hat{\beta}$ that can be identified in the same way by empirical version of $Ef(\beta, Y)$, namely $f(\beta, Y)$, should be a consistent estimator of β_0.

To illustrate this idea, let $\{Y_1, ..., Y_n\}$ be independent and identically distributed observations with a common density $p_\beta(y_i)$. Since β_0 is the maximizer of $E_{\beta_0} \log p_\beta(Y_1)$, it is also the minimax point of the function

$E_{\beta_0} \log\{p_{\beta'}(Y_1)/\ p_\beta(Y_1)\}$. In symbols,

$$\sup_{\beta' \in B} E_{\beta_0} \log\{p_{\beta'}(Y_1)/p_{\beta_0}(Y_1)\} = \inf_{\beta \in B} \sup_{\beta' \in B} E_{\beta_0} \log\{p_{\beta'}(Y_1)/p_\beta(Y_1)\}.$$

This suggests that the minimax of $n^{-1} \sum \log\{p_{\beta'}(Y_i)/p_\beta(Y_i)\}$, namely the $\hat{\beta}$ defined by the relation

$$\sup_{\beta \in B} \sum \log\{p_\beta(Y_i)/p_{\hat{\beta}}(Y_i)\} = \inf_{\beta \in B} \sup_{\beta' \in B} \sum \log\{p_{\beta'}(Y_i)/p_\beta(Y_i)\},$$

should be a consistent estimator of β_0. This is indeed the case: it can be easily verified that $\hat{\beta}$ defined above is identical to the maximum likelihood estimate in this case. See Li (1996).

The passage from the maximum of a likelihood to the minimax of a likelihood ratio is important because, unlike the likelihood, the likelihood ratio can be generalized to many estimating equations, so that the minimax argument applies very generally. Li (1993a) introduced such a generalization to quasi likelihood equations. This is then further extended in Li (1993b) to generalized estimating equations, which is now recorded below.

DEFINITION 1. *Let, in the notation of §2,* $\theta_1 = (\beta_1, \phi_1, \alpha_1^T)^T$ *and* $\theta_2 = (\beta_2, \phi_2, \alpha_2^T)^T$ *be two points in the parameter space* Θ. *Let* $D : \Theta \times \Theta \times \mathcal{Y} \mapsto R^1$ *be the function*

$$D(\theta_1, \theta_2) = \sum_{i=1}^{K} \{\mu_i(\beta_2) - \mu_i(\beta_1)\}^T \Big[\frac{\phi_1}{2} W_i^{-1}(\theta_1)\{Y_i - \mu_i(\beta_1)\} + \frac{\phi_2}{2} W_i^{-1}(\theta_2)\{Y_i - \mu_i(\beta_2)\}\Big].$$

The deviance function of the generalized estimating equation (2) is a mapping $R : B \times B \times \mathcal{Y} \mapsto R^1$ *defined by*

$$R(\beta_1, \beta_2) = D\{\beta_1, \hat{\phi}(\beta_1), \hat{\alpha}(\beta_1, \hat{\phi}(\beta_1)); \beta_2, \hat{\phi}(\beta_2), \hat{\alpha}(\beta_2, \hat{\phi}(\beta_2))\}.$$

The centering function of $R(\beta_1, \beta_2)$ *is the mapping* $J : B \times B \mapsto R^1$ *defined by*

$$J(\beta_1, \beta_2) = ED\{\beta_1, E\hat{\phi}(\beta_1), E\hat{\alpha}(\beta_1, E\hat{\phi}(\beta_1)); \beta_2, E\hat{\phi}(\beta_2), E\hat{\alpha}(\beta_2, E\hat{\phi}(\beta_2))\}.$$

Notice that the dependence of D and R on the random observations is suppressed from the notation. The next proposition provides some elementary properties of J, which can be verified along the lines of Li (1993a).

PROPOSITION 2. *The centering function $J(\beta_1, \beta_2)$ has the following properties:*

(i) $J(\beta_1, \beta_2) = -J(\beta_2, \beta_1)$ for all β_1, β_2 in B ;

(ii) $J(\beta_0, \beta)$ is the negative quadratic form

$$-\sum_i \{\mu_i(\beta) - \mu_i(\beta_0)\}^T W_i^{-1}\{\beta, E\hat{\phi}(\beta), E\hat{\alpha}(\beta, E\hat{\phi}(\beta))\}\{\mu_i(\beta) - \mu_i(\beta_0)\}.$$

If the matrices W's in (ii) are all positive definite, and if β is identifiable by the assumption of the mean; that is, different βs correspond to different sets of means $\{\mu_{it}(\beta)\}$, then $J(\beta_0, \beta) < 0$ for all β in B. This, combined with (i), implies that

$$\sup_{\beta \in B} J(\beta_0, \beta) = \inf_{\beta \in B} \sup_{\beta' \in B} J(\beta, \beta').$$

Thus, intuitively, if $R(\beta, \beta')$ converges to $J(\beta, \beta')$, then any $\hat{\beta}$ that satisfies the minimax relation

$$\sup_{\beta \in B} R(\hat{\beta}, \beta) = \inf_{\beta \in B} \sup_{\beta' \in B} R(\beta, \beta')$$

should be a consistent estimator of β_0. This will be proved rigorously in the next two sections.

5. ASSUMPTIONS AND LEMMAS

We shall frequently use the condition of stochastic equicontinuity, which can be found in Pollard (1984, page 139). The following definition is a combination of the condition C1 and Lemma 2.1 of Crowder (1986); it is to assume that a sequence of random functions and the corresponding sequence of centered random functions are both stochastically equicontinuous. Let T be a compact set in a Euclidean space, let $\{\hat{f}_n(t; X_1, ..., X_n) : n = 1, 2, ...\}$, abbreviated as $\{\hat{f}_n(t, X)\}$, be a sequence of random functions defined on T, and let $\{f_n(t) : n = 1, 2, ...\}$ be a sequence of (deterministic) functions of t, considered as the "centering" functions of $\{\hat{f}_n\}$.

DEFINITION 2. *Let T be a compact set in a Euclidean space. The sequence of random functions, $\{\hat{f}_n(t, X)\}$, and the associated centering sequence, $\{f_n(t)\}$, are said to obey condition C1 if the following are satisfied:*

(i) $\{f_n(t, X)\}$ is stochastically equicontinuous on T; that is, for each $\epsilon > 0, \eta > 0$ there is a positive number $\delta > 0$ such that,

$$\limsup_{n\to\infty} P\Big\{\sup_{\|s-t\|<\delta} |\hat{f}_n(t, X) - \hat{f}_n(s, X)| > \epsilon\Big\} < \eta;$$

(ii) $\{f_n(t)\}$ is equicontinuous on T;

(iii) There is $t_0 \in T$ such that $\{f_n(t_0) : n = 1, 2, ...\}$ is bounded.

By the Arzelà-Ascoli theorem (Conway 1990, page 175) and the compactness of T, (ii) and (iii) imply that $\{f_n(t)\}$ is totally bounded. Therefore, if $\{\hat{\phi}(\beta)\}$ and $\{E\hat{\phi}(\beta)\}$ obey C1, then the set $F_0 \equiv \{E\hat{\phi}(\beta) : \beta \in B, n = 1, 2, ...\}$ is contained in a compact set, F say. And, without loss of generality we assume that F is such that $\inf\{|\phi - \phi'| : \phi \in F_0, \phi' \in F^c\} > \epsilon_0$ for some $\epsilon_0 > 0$. Suppose that, in addition to C1, we have the convergence

$$\hat{\phi}(\beta) - E\hat{\phi}(\beta) \xrightarrow{p} 0. \tag{6}$$

By Lemma 3.2 of Crowder (1986), the pointwise convergence (6), together with assumptions (i) and (ii), implies the the uniform convergence $\sup_{\beta\in B} |\hat{\phi}(\beta) - E\hat{\phi}(\beta)| \xrightarrow{p} 0$. It follows that, with probability tending to one, the set $\{\hat{\phi}(\beta) : \beta \in B\}$ lies in F, because

$$\begin{aligned} P\{\hat{\phi}(\beta) \in F \text{ for all } \beta\} &\geq P\{|\hat{\phi}(\beta) - E\hat{\phi}(\beta)| < \epsilon_0 \text{ for all } \beta\} \\ &= P\{\sup |\hat{\phi}(\beta) - E\hat{\phi}(\beta)| < \epsilon_0\} \to 1. \end{aligned}$$

This proves the following result.

LEMMA 1. *Suppose $\{\hat{\phi}(\cdot), E\hat{\phi}(\cdot) : n = 1, 2, ...\}$ obeys C1, and suppose (6) holds. Then, with probability tending to one, $\{\hat{\phi}(\beta) : \beta \in B\}$ is contained in a compact set F.*

With this construction, the condition C1 is easily passed from simple functions to composite functions, a passage of importance since we are to study the limit behaviour of the substituted estimates such as $\hat{\alpha}(\beta, \hat{\phi}(\beta))$ as a random function of β.

LEMMA 2. *Suppose $\{\hat{\phi}(\cdot), E\hat{\phi}(\cdot)\}$ obeys C1 on B, and $\{\hat{\alpha}(\cdot, \cdot), E\hat{\alpha}(\cdot, \cdot)\}$ obeys C1 on $B \times F$. Then $\{\hat{\alpha}(\cdot, \hat{\phi}(\cdot)), E\hat{\alpha}(\cdot, E\hat{\phi}(\cdot))\}$ obeys C1 on B.*

PROOF. By Billingsley (1968, page 221), conditions (ii) and (iii), as applied to $\{\hat{\phi}(\cdot), E\hat{\phi}(\cdot)\}$ and $\{\hat{\alpha}(\cdot,\cdot), E\hat{\alpha}(\cdot,\cdot)\}$, imply that the sets $\{E\hat{\phi}(\beta) : \beta \in B, n = 1, 2, ...\}$ and $\{E\hat{\alpha}(\beta, \phi) : (\beta, \phi) \in B \times F, n = 1, 2, ...\}$ are bounded. So for any $\beta \in B$,

$$\sup_n |E\hat{\alpha}(\beta, E\hat{\phi}(\beta))| \le \sup_n \sup_{\phi \in F} |E\hat{\alpha}(\beta, \phi)| \le \sup_n \sup_{\beta \in B} \sup_{\phi \in F} |E\hat{\alpha}(\beta, \phi)| < \infty.$$

Hence (iii) holds for $\{E\hat{\alpha}(\cdot, E\hat{\phi}(\cdot))\}$. To prove (ii), let $\epsilon > 0$, and let $\delta_0 > 0$ be such that

$$\limsup_{n\to\infty} \sup\{|E\hat{\alpha}(\beta, \phi) - E\hat{\alpha}(\beta', \phi')| : \|\beta - \beta'\| < \delta_0, \|\phi - \phi'\| < \delta_0\} < \epsilon.$$

This is possible by stochastic equicontinuity of $\{\hat{\alpha}(\cdot,\cdot)\}$ and compactness of $B \times F$. Let $\delta_1 > 0$ be such that $\limsup_{n\to\infty} \sup\{|E\hat{\phi}(\beta) - E\hat{\phi}(\beta')| : \|\beta - \beta'\| < \delta_1\} < \delta_0$. It follows that

$$\limsup_{n\to\infty} \sup\{|E\hat{\alpha}(\beta, E\hat{\phi}(\beta)) - E\hat{\alpha}(\beta', E\hat{\phi}(\beta'))| : \|\beta - \beta'\| < \min\{\delta_0, \delta_1\}\} < \epsilon.$$

To prove (iii), let $\epsilon > 0$, $\eta > 0$. Since

$$\lim_{n\to\infty} P\{\hat{\phi}(\beta) \in F, \text{ for all } \beta \in B\} = 1,$$

we can, and do, assume $\hat{\phi}(\beta) \in F$ without altering our limit argument. Let $\delta_0 > 0$ be such that

$$\limsup_{n\to\infty} P\{\sup |\hat{\alpha}(\beta, \phi) - \hat{\alpha}(\beta', \phi')| > \epsilon\} < \eta/2,$$

supremum being over the set $\{(\beta, \phi, \beta', \phi') : \|\beta - \beta'\| < \delta_0, \|\phi - \phi'\| < \delta_0\}$. Let $\delta_1 > 0$ be such that

$$\limsup_{n\to\infty} P\{\sup |\hat{\phi}(\beta) - \hat{\phi}(\beta')| > \delta_0\} < \eta/2,$$

supremum being over the set $\{(\beta, \beta') : \|\beta - \beta'\| < \delta_1\}$. Put $\delta = \min\{\delta_0, \delta_1\}$. It follows that

$$\limsup_{n\to\infty} P\{ \sup_{\|\beta-\beta'\|<\delta} \left|\hat{\alpha}(\beta, \hat{\phi}(\beta)) - \hat{\alpha}(\beta', \hat{\phi}(\beta'))\right| > \epsilon\}$$

$$\le \limsup_{n\to\infty} P\{ \sup_{\|\beta-\beta'\|<\delta} \left|\hat{\alpha}(\beta, \hat{\phi}(\beta)) - \hat{\alpha}(\beta', \hat{\phi}(\beta'))\right| > \epsilon,$$

$$\sup_{\|\beta-\beta'\|<\delta}\left|\hat{\phi}(\beta)-\hat{\phi}(\beta')\right|<\delta_0\}+\frac{\eta}{2}$$

$$\leq \limsup_{n\to\infty} P\left\{\sup\left|\hat{\alpha}(\beta,\phi)-\hat{\alpha}(\beta',\phi')\right|>\epsilon\right\}+\frac{\eta}{2}\leq \eta,$$

where the supremum on the last line is over the set $\{\|\beta-\beta'\|<\delta_0, \|\phi-\phi'\|<\delta_0\}$. □

LEMMA 3. *Suppose (a) for each $\beta\in B$ and $\phi\in F$, $\hat{\phi}(\beta)-E\hat{\phi}(\beta)\xrightarrow{p}0$ and $\hat{\alpha}(\beta,\phi)-E\hat{\alpha}(\beta,\phi)\xrightarrow{p}0$, and (b) $\{\hat{\phi}(\beta),E\hat{\phi}(\beta)\}$ obeys C1 on B, and $\{\hat{\alpha}(\beta,\phi),E\hat{\alpha}(\beta,\phi)\}$ obeys C1 on $B\times F$. Then the sequence of random functions*

$$\{\hat{\alpha}(\cdot,\hat{\phi}(\cdot))-E\hat{\alpha}(\cdot,E\hat{\phi}(\cdot)):n=1,2,...\} \tag{7}$$

converges weakly in $C(B)$ to a random function degenerated at constant 0, where $C(B)$ is the class of continuous functions defined on the compact set B.

PROOF. First, we show that for each $\beta\in B$, $\hat{\alpha}(\beta,\hat{\phi}(\beta))-E\hat{\alpha}(\beta,E\hat{\phi}(\beta))$ converges in probability to 0. In other words, the finite dimensional distributions of the sequence of random functions (7) converge to those of the random function degenerated at constant 0. Let $\epsilon>0$, and let $\delta>0$ be such that

$$\limsup_{n\to\infty} P\Big\{\sup_{\|\phi-\phi'\|<\delta}|\hat{\alpha}(\beta,\phi)-\hat{\alpha}(\beta,\phi')|>\epsilon/2\Big\}=0. \tag{8}$$

From the discussion preceding Lemma 1, assumptions (a) and (b) imply that the sequence $\{\hat{\alpha}(\cdot,\cdot):n=1,2,...\}$ converges weakly to constant 0 in $C(B\times F)$. It follows that

$$\begin{aligned}\limsup_{n\to\infty} & P\left\{|\hat{\alpha}(\beta,E\hat{\phi}(\beta))-E\hat{\alpha}(\beta,E\hat{\phi}(\beta))|>\epsilon/2\right\}\\ &\leq \limsup_{n\to\infty} P\{\sup_{\beta\in B,\phi\in F}|\hat{\alpha}(\beta,\phi)-E\hat{\alpha}(\beta,\phi)|>\epsilon/2\}=0. \end{aligned}\tag{9}$$

Therefore,

$$\limsup_{n\to\infty} P\{|\hat{\alpha}(\beta,\hat{\phi}(\beta))-E\hat{\alpha}(\beta,E\hat{\phi}(\beta))|>\epsilon\}$$

$$=\limsup_{n\to\infty} P\{|\hat{\alpha}(\beta,\hat{\phi}(\beta))-E\hat{\alpha}(\beta,E\hat{\phi}(\beta))|>\epsilon, |\hat{\phi}(\beta)-E\hat{\phi}(\beta)|<\delta\}$$

[by assumption (a)]

$$\le \limsup_{n\to\infty} P\{|\hat\alpha(\beta,\hat\phi(\beta)) - E\hat\alpha(\beta,E\hat\phi(\beta))| > \epsilon, |\hat\phi(\beta) - E\hat\phi(\beta)| < \delta,$$

$$|\hat\alpha(\beta,E\hat\phi(\beta)) - E\hat\alpha(\beta,E\hat\phi(\beta))| < \epsilon/2\} \text{ [by (9)]}$$

$$\le \limsup_{n\to\infty} P\{|\hat\alpha(\beta,\hat\phi(\beta)) - \hat\alpha(\beta,E\hat\phi(\beta))| > \epsilon/2, \quad |\hat\phi(\beta) - E\hat\phi(\beta)| < \delta\} = 0, \tag{10}$$

where the right most limit equals 0 because it is no more than the left hand side of (8). By Lemma 2 we know that the sequence (7) obeys C1. This, together with (10), implies that the sequence (7) is tight in $C(B)$. This proves the asserted result. □

At this stage it is helpful to see through an example just what assumption C1 means, and how it can be verified for a specific moment structure. We do so by look into the moment estimate of ϕ suggested by Liang & Zeger (1986). The assumptions about the estimates of α suggested may be investigated following a similar procedure.

EXAMPLE 1. For simplicity, we assume that $n_i = n$ for all i. Let B be a compact set in R^p. Liang & Zeger (1986) suggested estimating ϕ by averaging over the residues based on the marginal moment assumptions, as follows

$$\hat\phi(\beta) = \frac{1}{Kn}\sum_{i=1}^{K}\sum_{t=1}^{n}\left\{\frac{Y_{it}-\mu_{it}(\beta)}{V_{it}^{\frac{1}{2}}(\beta)}\right\}^2.$$

Actually, Liang & Zeger used $(K-p)n$ as the denominator; we can ignore the finite number p without affecting the limit argument. Let β be an arbitrary but fixed parameter value, and β' a point in an open ball O_β centered at β, whose radius is yet to be determined. Let ϵ and η be two positive numbers. By direct calculation,

$$\begin{aligned}|\hat\phi(\beta') \; - \; \hat\phi(\beta)| &\\ \le \frac{1}{Kn}\sum_{i,t} V_{it}^{\frac{1}{2}}(\beta') \times \{2|Y_{it}| + |\mu_{it}(\beta') + \mu_{it}(\beta)|\} \times |\mu_{it}(\beta') - \mu_{it}(\beta)| &\\ + \frac{1}{Kn}\sum_{i,t}\frac{|Y_{it}-\mu_{it}(\beta)|^2}{V_{it}(\beta)V_{it}(\beta')} \times |V_{it}(\beta') - V_{it}(\beta)|. &\end{aligned} \tag{11}$$

We now verify that $\{\hat\phi(\beta)\}$ is stochastically equicontinuous if the following conditions are satisfied

(i) For each t, the sequences $\{\mu_{it}(\cdot) : i = 1, 2, ...\}$ and $\{V_{it}(\cdot) : i = 1, 2, ...\}$ are equicontinuous and the second sequence is bounded away from 0.

(ii) There are parameter values β_a and β_b for which $\{\mu_{it}(\beta_a) : i = 1, 2, ...\}$ and $\{V_{it}(\beta_b) : i = 1, 2, ...\}$ are bounded.

By (i) and (ii), for each t, $|\mu_{it}(\beta)|$ is bounded for $i = 1, 2, ...,$ and $\beta \in B$. Therefore, since n is finite, $|\mu_{it}(\beta)| < M_1$. Similarly, $1/V_{it}(\beta) < M_2$. Hence (11) implies that

$$\sup_{\beta' \in O_\beta} |\hat{\phi}(\beta') - \hat{\phi}(\beta)| \le \frac{2M_2^{\frac{1}{2}}}{Kn} \sum_{i,t} \{|Y_{it}| \times \sup_{\beta' \in O_\beta} |\mu_{it}(\beta') - \mu_{it}(\beta)|\}$$

$$+\frac{2M_1 M_2^{\frac{1}{2}}}{Kn} \sum_{i,t} \sup_{\beta' \in O_\beta} |\mu_{it}(\beta') - \mu_{it}(\beta)|$$

$$+\frac{M_2^2}{Kn} \sum_{i,t} \{|Y_{it} - \mu_{it}(\beta)|^2 \times \sup_{\beta' \in O_\beta} |V_{it}(\beta') - V_{it}(\beta)|\}.$$

Assumptions (i) and (ii) also imply that the sequence

$$\{(Kn)^{-1} \sum_{i,t} \{V_{it}(\beta_0) + \mu_{it}^2(\beta_0)\} : K = 1, 2, ...\}$$

is bounded, and so both $\{(Kn)^{-1} \sum_{i,t} |Y_{it}| : i = 1, 2, ...\}$ and $\{(Kn)^{-1} \sum_{i,t} |Y_{it} - \mu_{it}(\beta)|^2 : i = 1, 2, ...\}$ are bounded in probability. In other words, there is a positive number M_3 for which

$$\limsup_{K\to\infty} P\left\{\frac{2M_2^{\frac{1}{2}}}{Kn} \sum_{i,t} |Y_{it}| > M_3 \quad \text{or} \quad \frac{M_2^2}{Kn} \sum_{i,t} |Y_{it} - \mu_{it}(\beta)|^2 > M_3\right\} < \frac{\epsilon}{2}$$

By assumption (i) we can find $\delta > 0$ as the radius of O_β such that

$$\max\left\{\sup_{\beta' \in O_\beta} |\mu_{it}(\beta') - \mu_{it}(\beta)|, \sup_{\beta' \in O_\beta} |V_{it}(\beta') - V_{it}(\beta)|\right\} < \min\left\{\frac{\eta}{3M_3 M_2^{1/2}}, \frac{\eta}{6M_1}\right\}$$

for all $t = 1, 2, ..., n; i = 1, 2, ...$ Then, $\limsup_{n\to\infty} P\{|\hat{\phi}(\beta') - \hat{\phi}(\beta)| > \eta\} < \epsilon$ and, hence, $\{\hat{\phi}(\cdot)\}$ is stochastic equicontinuous. Evidently assumptions (i) and (ii) also imply that $\{E\hat{\phi}(\cdot)\}$ is a equicontinuous and bounded sequence. Therefore $\{\hat{\phi}(\beta), E\hat{\phi}(\beta)\}$ satisfies C1. □

From Lemma 3 and the argument preceding Lemma 1, it is easy to see that, with probability tending to one, the random set $\{\hat{\alpha}(\beta, \hat{\phi}(\beta)) : \beta \in B\}$ is contained in a compact set in R^s. We write the set as A, and write the compact set $B \times F \times A$ in R^{p+1+s} as Θ. It now takes only a small step further to obtain the limit behaviour of the deviance function introduced in the last section.

LEMMA 4. *Suppose that, in addition to the assumptions of Lemma 3, the sequence of random functions $\{n^{-1}D(\theta_1, \theta_2) - n^{-1}ED(\theta_1, \theta_2) : n = 1, 2, ...\}$ obeys C1 in Θ^2, and that for each $(\theta_1, \theta_2) \in \Theta^2$,*

$$n^{-1}D(\theta_1, \theta_2) - n^{-1}ED(\theta_1, \theta_2) \xrightarrow{p} 0.$$

Then the sequences $\{n^{-1}R(\cdot, \cdot), n^{-1}J(\cdot, \cdot) : n = 1, 2, ...\}$ obey C1 in B^2, and $n^{-1}\{R(\cdot, \cdot) - J(\cdot, \cdot)\}$ converges weakly in $C(B^2)$ to a constant function 0.

The proof is similar to that of lemma 3, and will be omitted.

6. CONSISTENCY OF GENERALIZED ESTIMATING EQUATIONS

We are now ready to prove that any minimax point of the function $R(\beta_1, \beta_2)$ is consistent. As we shall see in the next section, under mild conditions, any minimax points are solutions to generalized estimating equation (2). In other words, we can identify the consistent solutions of (2) by verifying that they are minimax points. Two numerically more convenient criteria will be presented in the next section.

The function $J(\beta, \beta_0)$ plays the same role as the Kullback-Leibler information number in the Wald's proof of consistency of maximum likelihood estimate, and $R(\beta_1, \beta_2)$ plays the role of log likelihood ratio. However, in our case, neither $R(\beta_1, \beta_2)$ nor $J(\beta_1, \beta_2)$ has the form $f(\beta_2) - f(\beta_1)$. This is why a minimax deviance procedure must be used in place of the maximum likelihood procedure used in the Wald's proof.

THEOREM 1. *Suppose*

(a) The sequence $\{\hat{\phi}(\beta), E\hat{\phi}(\beta)\}$ obeys C1 in B, $\{\hat{\alpha}(\beta, \phi), E\hat{\alpha}(\beta, \phi)\}$ obeys C1 in $B \times F$, and $\{n^{-1}D(\theta_1, \theta_2), n^{-1}ED(\theta_1, \theta_2)\}$ obeys C1 in Θ^2;

(b) For each $\theta \in \Theta$, $\hat{\phi}(\beta) - E\hat{\phi}(\beta) \xrightarrow{p} 0$, $\hat{\alpha}(\beta, \phi) - E\hat{\alpha}(\beta, \phi) \xrightarrow{p} 0$, and $n^{-1}D(\theta_1, \theta_2) - n^{-1}ED(\theta_1, \theta_2) \xrightarrow{p} 0$;

(c) For each β in B, $\beta \neq \beta_0$, the sequence of quadratic forms $\{n^{-1}J(\beta, \beta_0)\}$ satisfies

$$\liminf_{n\to\infty} n^{-1}J(\beta, \beta_0) > 0.$$

Then any parameter value $\hat{\beta}$ that satisfy the relation

$$\sup_{\beta\in B} R(\hat{\beta}, \beta) = \inf_{\beta\in B} \sup_{\beta'\in B} R(\beta, \beta') \tag{12}$$

is a consistent estimate of β_0.

Notice that the estimates $\hat{\phi}$, $\hat{\alpha}$ need not be consistent; in fact, they need not converge in probability at all. Furthermore, the theorem asserts that all minimax points (if there are more than one) are consistent. The proof of the theorem is along the lines of Theorem 1 of Li (1996); here we only describe the idea briefly and highlight the difference.

PROOF. Let O_{β_0} be an arbitrary but fixed open ball centered at the true parameter value β_0. Let $\beta \neq \beta_0$ and O_β be an open ball whose closure does not contain β_0. By assumptions (a), (b), and Corollary 1, the sequence $n^{-1}\{R(\beta, \beta_0) - J(\beta, \beta_0)\}$ converges weakly in $C(B)$ to the random function degenerate at constant 0. This, together with assumption (c), implies that with probability tending to 1, $\inf_{\beta'\in O_\beta} n^{-1}R(\beta', \beta_0) > \delta$ for some positive δ that may be dependent on β. Now the class of such open balls $\{O_\beta : \beta \in B \setminus O_{\beta_0}\}$ form an open cover of $B \setminus O_{\beta_0}$. By compactness of $B \setminus O_{\beta_0}$ there is a finite subcover $\{O_i : i = 1, ..., k\}$. Now on each O_i, there is a $\delta_i > 0$ such that, with probability tending to one, $\inf_{\beta'\in O_i} n^{-1}R(\beta', \beta_0) > \delta_i$. It follows that, with probability tending to 1, $\inf_{\beta'\notin O_{\beta_0}} n^{-1}R(\beta', \beta_0) > \delta$ for some positive δ. Since $\inf_{\beta_1\notin O_{\beta_0}} \sup_{\beta_2\in B} n^{-1}R(\beta_1, \beta_2) \geq \inf_{\beta'\notin O_{\beta_0}} n^{-1}R(\beta', \beta_0)$, we see that

$$\lim_{n\to\infty} P\Big\{\inf_{\beta\notin O_{\beta_0}} \sup_{\beta'\in B} n^{-1}R(\beta, \beta') > \delta\Big\} = 1. \tag{13}$$

Meanwhile, by a similar argument, one can show that for every positive $\delta > 0$

$$\lim_{n\to\infty} P\Big\{\inf_{\beta\in B} \sup_{\beta'\in B} n^{-1}R(\beta, \beta') < \delta\Big\} = 1. \tag{14}$$

However, (13) and (14) imply that, with probability tending to one, any minimax point $\hat{\beta}$ of $R(\beta, \beta')$ is in O_{β_0}. In other words, $\hat{\beta}$ converges in probability to β_0. □

7. TWO PRACTICAL CRITERIA FOR CONSISTENCY

Theoretically, Theorem 1 solves the consistency problem. Practically, the search for the global minimax of $R(\beta, \beta')$ may be numerically difficult, and so we need simpler criteria for consistency. In this section we will introduce two such criteria. To achieve these, we need to assume that the function $R(\beta, \beta')$ do not behave too irregularly along the straight line $\beta_2 = \beta_1$, specifically, that, with probability tending to one, the following condition is satisfied

$$\inf_{\beta \in B} \sup_{\beta' \in B} R(\beta, \beta') = \sup_{\beta' \in B} \inf_{\beta \in B} R(\beta, \beta'). \tag{15}$$

Roughly, the condition requires that, as β moves pass the true parameter value β_0, the mode of the function $R(\beta, \cdot)$ moves continuously. For further discussion of this point, see Li (1996). In practice, such continuity is almost always satisfied; It is a challenge to find a counter example. Nevertheless, condition (15) is not implied by the antisymmetry of the function of R, nor assumption (iii) of Theorem 1. The latter two conditions only guarantee that

$$P\left\{n^{-1}\left|\inf_{\beta \in B} \sup_{\beta' \in B} R(\beta, \beta') - \sup_{\beta' \in B} \inf_{\beta \in B} R(\beta, \beta')\right| < \epsilon\right\} \to 1 \text{ for each } \epsilon > 0,$$

as can be seen from the proof of Theorem 1. Condition (15) is not crucial from a theoretical point of view because, as we have seen, the minimax $\hat{\beta}$ defined in (12) is consistent whether or not (15) holds. And, without requiring condition (15), we can show that $\hat{\beta}$ is efficient using a method similar to that used in the proof of Theorem 2 of Li (1996). However, the condition does simplify the computation and discussion, because it guarantees that the minimax point of R is necessarily a solution to the generalized estimating equation. Let $\hat{B}$ be the set of all solutions to the generalized estimating equation (2).

COROLLARY 1. *Suppose (a) the condition (15) is satisfied, (b) the minimax of R is in the interior of B, (c) $R(\cdot,\cdot)$ is a differentiable everywhere in B^2. Then, under the assumptions of Theorem 1, any solution $\tilde{\beta}$ of equation (2) that satisfies*

$$\sup_{\beta \in \hat{B}} R(\tilde{\beta}, \beta) = 0 \tag{16}$$

is consistent.

PROOF. Let $\hat{\beta}$ be a (any) minimax of R. Under the assumptions (a), (b), and (c), it is easy to show that $\hat{\beta}$ is in $\hat{B}$. The argument is similar to the proof of Theorem 3(b) of Li (1996), and the detail is omitted. Now let $\tilde{\beta}$ be a solution to (2) that satisfies (16). Since $\hat{\beta}$ is the minimax, and since $\hat{\beta} \in \hat{B}$, we have

$$0 \leq \sup_{\beta \in \hat{B}} R(\hat{\beta}, \beta) \leq \sup_{\beta \in B} R(\hat{\beta}, \beta) = 0; \quad \text{so} \quad \sup_{\beta \in \hat{B}} R(\hat{\beta}, \beta) = 0.$$

Since $\tilde{\beta} \in \hat{B}$, the above implies $R(\hat{\beta}, \tilde{\beta}) \leq 0$. But, since $\hat{\beta} \in \hat{B}$, (16) implies $R(\tilde{\beta}, \hat{\beta}) \leq 0$. Hence, by the antisymmetry of R, $R(\hat{\beta}, \tilde{\beta}) = 0$.

Now let β_0 be the true value of the parameter β, let $\epsilon > 0$ and $\rho > 0$ be arbitrary but fixed, and let O_{β_0} be the open ball of radius ρ centered at β_0. By an argument similar to that used in the proof of Theorem 1, there is a positive number η for which

$$P\Big\{ \sup_{\beta \notin O_{\beta_0}} n^{-1} R(\beta_0, \beta) < -\eta \Big\} \to 1, \quad \text{as } n \to \infty. \tag{17}$$

Now by Lemma 4, the sequence $\{n^{-1}R(\cdot,\cdot) : n = 1, 2, ...\}$ is stochastically uniformly equicontinuous, therefore there is a $\delta > 0$ for which

$$\limsup_{n \to \infty} P\{\sup |n^{-1}R(\beta_1, \beta_2) - n^{-1}R(\beta_1', \beta_2)| > \eta/2\} < \epsilon, \tag{18}$$

where the supremum is taken over the set $\{(\beta_1, \beta_1', \beta_2) \in B^3 : \|\beta_1 - \beta_1'\| < \delta\}$. By Theorem 1,

$$\lim_{n \to \infty} P(\|\hat{\beta} - \beta_0\| < \delta) = 1. \tag{19}$$

Combining (18) and (19), it follows that

$$\limsup_{n \to \infty} P\Big\{ n^{-1} \sup_{\beta \in B} |R(\hat{\beta}, \beta) - R(\beta_0, \beta)| > \eta/2 \Big\} < \epsilon. \tag{20}$$

Hence, by (17) and (20),

$$\begin{aligned}
&\limsup_{n \to \infty} P(\tilde{\beta} \notin O_{\beta_0}) \\
&\leq \limsup_{n \to \infty} P\{\tilde{\beta} \notin O_{\beta_0},\ n^{-1} \sup_{\beta \notin O_{\beta_0}} R(\beta_0, \beta) < -\eta, \\
&\qquad n^{-1} \sup_{\beta \in B} |R(\hat{\beta}, \beta) - R(\beta_0, \beta)| < \eta/2\} + \epsilon \\
&\leq \limsup_{n \to \infty} P\{n^{-1}R(\beta_0, \tilde{\beta}) < -\eta,\ n^{-1}|R(\hat{\beta}, \tilde{\beta}) - R(\beta_0, \tilde{\beta})| < \eta/2\} + \epsilon \\
&\leq \limsup_{n \to \infty} P\{n^{-1}R(\hat{\beta}, \tilde{\beta}) < -\eta/2\} + \epsilon = \epsilon.
\end{aligned}$$

Since ϵ is arbitrary we conclude that $P(\tilde{\beta} \notin O_{\beta_0}) \to 0$. □

If we do not want to find all the solutions of (2), we can use the following immediate consequence of Corollary 1.

COROLLARY 2. *Under the conditions of Corollary 1, any solution $\tilde{\beta}$ of (2) that satisfies*

$$\sup_{\beta \in B} R(\tilde{\beta}, \beta) = 0 \tag{21}$$

is consistent.

Thus, in order to determine whether a solution $\tilde{\beta}$ belongs to a consistent sequence, it suffices to check whether $R(\tilde{\beta}, \beta) \leq 0$ for all β. We now apply the minimax-deviance approach to a numerical example.

8. A NUMERICAL EXAMPLE

EXAMPLE 2. Let $\{Y_{it} : t = 1, 2; i = 1, ..., 30\}$ be thirty bivariate observations. For each i, (Y_{i1}, Y_{i2}) follows a bivariate normal distribution with expectation located at an unknown point of a cardioid and correlation matrix completely unspecified. That is

$$\begin{pmatrix} \mu_\beta^1 \\ \mu_\beta^2 \end{pmatrix} = \begin{pmatrix} 2\cos\beta - \cos 2\beta \\ 2\sin\beta - \sin 2\beta \end{pmatrix}, \qquad \mathrm{corr}_\theta(Y_{i1}, Y_{i2}) = \phi \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix},$$

where $\theta = (\beta, \phi, \alpha)$ are unknown and β is the parameter of interest. Thirty observations are generated with $\beta = \pi/4$, $\phi = 0.3$ and $\alpha = 0.3$, and are presented in Figure 1. There are four solutions to the generalized estimating equation. The solution $\beta = 0$ will be ignored since it has nothing to do with data. The data and the solutions are presented in Figure 1.

The three non-trivial solutions are $\hat{\beta}_0 = 0.28\pi$, $\hat{\beta}_1 = 0.91\pi$, and $\hat{\beta}_2 = 1.92\pi$. In this simple example, the likelihood function is available, and it takes a global maximum at $\hat{\beta}_0$, a local maximum at $\hat{\beta}_2$, and a global minimum at $\hat{\beta}_1$. Plotted in Figure 2 is the curve $l(\beta) = \sup_{\beta' \in B} R(\beta, \beta')$; so the minimum point of $l(\beta)$ is the minimax of $R(\beta, \beta')$. Thus by Theorem 1, $\hat{\beta}_0$ is a consistent solution. Figure 2 also indicates that $l(\hat{\beta}_0) = \inf_{\beta \in B} l(\beta) = 0$; in other words $\sup_{\beta \in B} R(\hat{\beta}_0, \beta) = 0$. So Corollary 2 also tells us that $\hat{\beta}_0$ is consistent.

ACKNOWLEDGEMENT

I wish to thank a referee for his (or her) very useful comments. The research is supported by the National Science Foundation Grant DMS-9306738.

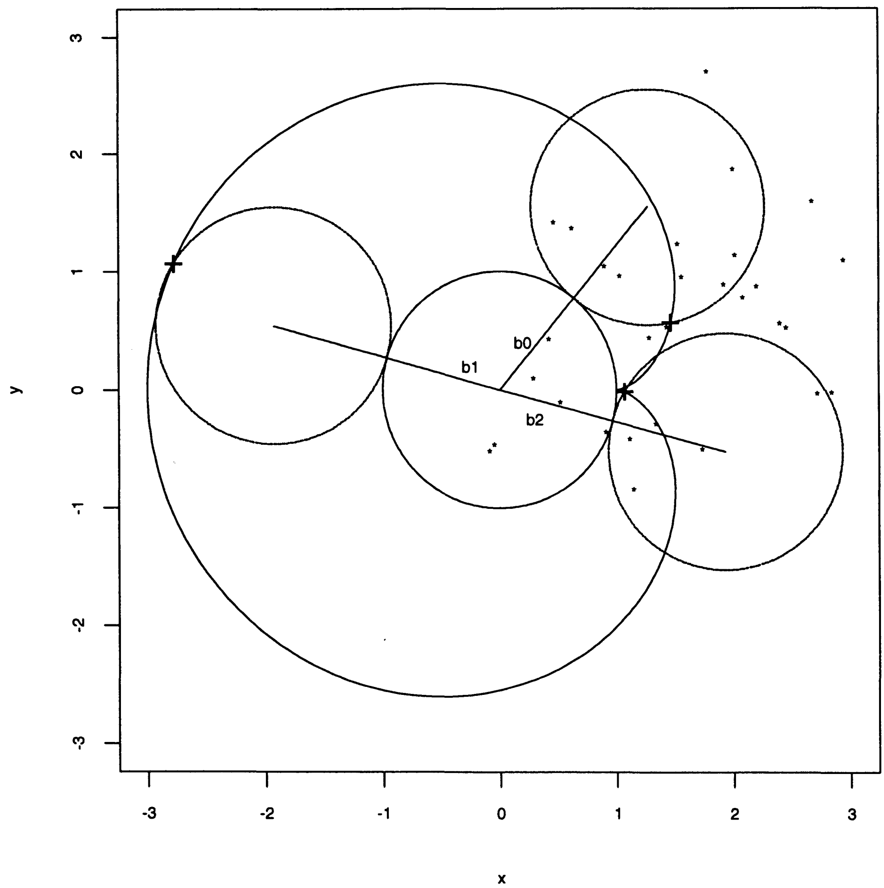

Figure 1: Multiple solutions for the cardioid model. b0, b1, b2 are $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$. The three +'s mark the positions that correspond to the three solutions.

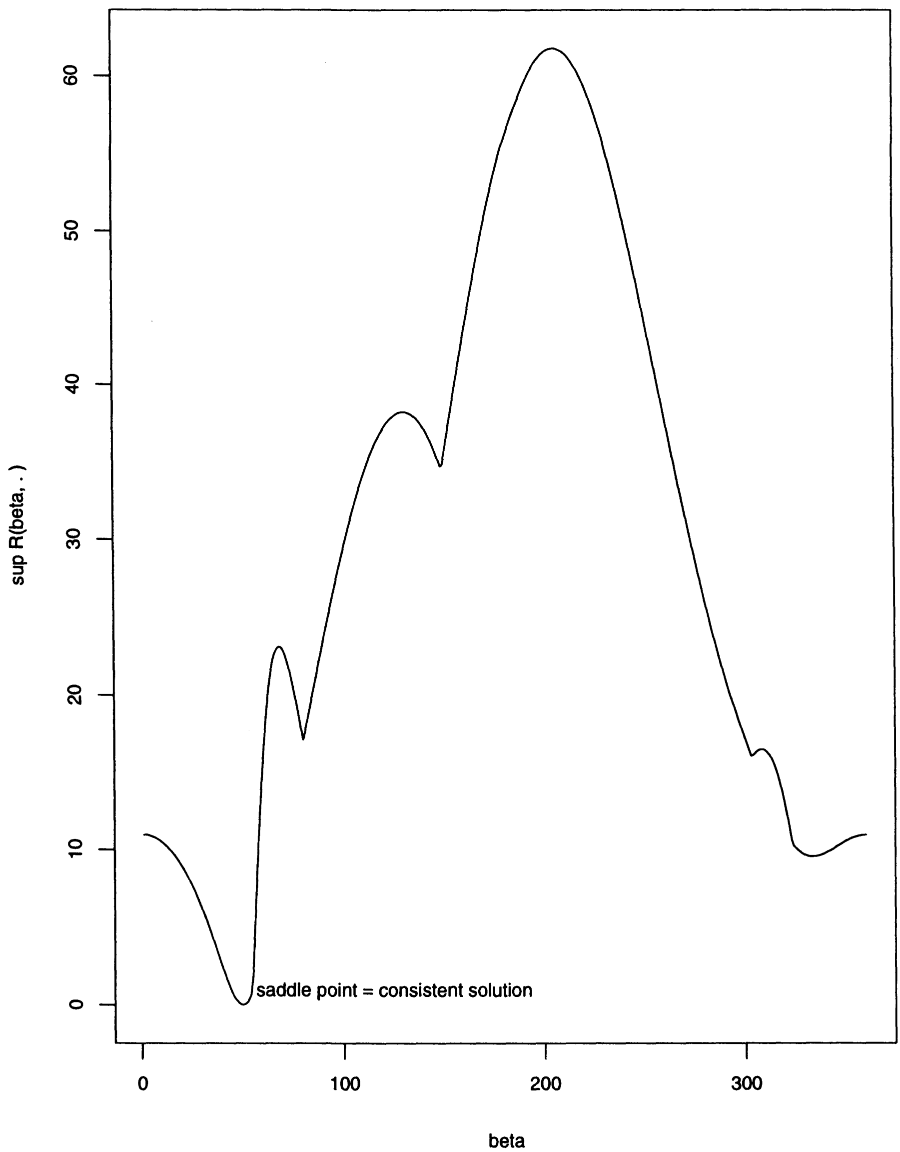

Figure 2: Saddle point of the deviance function. The curve is $\sup_{\beta' \in B} R(\beta, \beta')$ as a function of β. The minimum point corresponds to the saddle point of R.

REFERENCES

[1] Billingsley, P. (1968). *Convergence of Probability Measures.* John Wiley & Sons, Inc.

[2] Conway, J.B. (1990). *A Course in Functional Analysis.* 2nd edition. New York: Springer.

[3] Doob, J.S. (1934). Probability and Statistics. *Trans. Amer. Math. Soc.* **36**, 759-775.

[4] Crowder, M. (1986). On consistency and inconsistency of estimating equations. *Econometric Theory* **3**, 305-30.

[5] Firth, D. & Harris, I.R. (1991). Quasi-likelihood for multiplicative random effects. *Biometrika* **78**, 545-555.

[6] Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* **31**, 1208-1211.

[7] Godambe, V.P. and Heyde, C.C. (1987) Quasi-likelihood and optimal estimation. *Int. Statist. Rev.* **55**, 231-244.

[8] Gourieroux, C., Monfort, A., & Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica* **52**, 681-700.

[9] Jarrett, R.G. (1984). Bounds and expansions for Fisher information when moments are known. *Biometrika* **74**, 233-245.

[10] Li, B. (1993a). A deviance function for the quasi likelihood method. *Biometrika* **80**, 741-753.

[11] Li, B. (1993b). Deviance functions for generalized estimating equations. unpublished manuscript.

[12] Li, B. & McCullagh, P. (1994). Potential functions and conservative estimating functions. *Ann. Statist.* **22**, 340-356.

[13] Li, B. (1996). A minimax approach to consistency and efficiency for estimating equations. To appear in *Ann. Statist.* **24**.

[14] Liang, K.Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

[15] McCullagh, P. (1983). Quasi-likelihood functions. *Ann. Statist.* **11**, 59-67.

[16] McCullagh, P. (1990). Quasi-likelihood and estimating functions. In *Statistical Theory and Modelling: in honour of Sir David Cox.* Ed by D.V.Hinkley, N.Reid, and E.J.Snell. London: Chapman & Hall.

[17] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models.* 2nd edition. London: Chapman & Hall.

[18] McLeish, D.L. (1984). Estimation for aggregate models: the aggregate Markov chain. *Can. J. Statist.* **12**, 265-282.

[19] Pollard, D. (1984). *Convergence of Stochastic Processes.* New York: Springer.

[20] Small, C.G. & McLeish, D.L. (1988). Generalization of ancillarity, completeness and sufficiency in an inference function space, *Ann. Statist.* **16**, 534-551.

[21] Small, C.G. & McLeish, D.L. (1989). Projection as a method for increasing sensitivity and eliminating nuisance parameters, *Biometrika* **76**, 693-703.

[22] Small, C.G. & McLeish, D.L. (1994). *Hilbert Space Methods in Probability and Statistical Inference.* New York: John Wiley.

[23] Wald, A. (1949). Note on the consistency of maximum likelihood estimate, *Ann. Math. Statist.* **20**, 595-601.

[24] Wedderburn, R.W.M. (1974). Quasi-likelihood, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439-447.

[25] Wolfowitz, J. (1949) On Wald's proof of the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20**, 601-602.