International Statistical Review (1987), **55**, 3, pp. 231–244. Printed in Great Britain

Correspondent paper

Quasi-likelihood and Optimal Estimation

V.P. Godambe[1] and C.C. Heyde[2]

[1] Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1. [2] Department of Statistics, IAS, Australian National University, GPO Box 4, Canberra, ACT 2601, Australia.

Summary

Within the framework of estimating function theory, this paper provides a very general definition of quasi-likelihood estimating equations. Applications to stochastic processes are discussed. This work extends the previous results of Godambe (1985) and Hutton & Nelson (1986).

Key words: Asymptotic theory; Estimating functions; Generalized least-squares; Martingales; Maximum likelihood; Optimal estimators; Quasi-likelihood; Quasi-score function.

1 Historical introduction

Historically there are two distinct and sharply contrasting traditions in relation to estimation problems arising out of what are now known as classical linear models: (I) Gauss's relying exclusively on finite sample properties, and (II) Laplace's based on asymptotics.

Gauss in 1821, 1823 (Gauss, 1880) founded and substantially developed the methodology of what is usually called the Gauss–Markov theorem. Using only the 'first two moments' of the underlying distribution, the theorem, in terms of the modern concept of unbiasedness, asserts that in the class of linear unbiased estimates the variance is minimized for the 'least-squares estimate'. In this sense the least-squares estimates are *optimal.* This optimality is in relation to *finite samples.* Gauss did not investigate asymptotic properties of the least-squares estimates. However, he demonstrated (and emphasized) that in the location family of distributions *uniquely* for 'normal distributions', the 'maximum likelihood estimates' (using the present day terminology) *coincide* with the 'least-squares estimates', implying some kind of justification for the normal models. To clarify the point, we emphasize that the optimality of the least-squares estimate mentioned above depends on assumptions concerning the first two moments of the distribution but is otherwise *independent* of the distribution or distributional form. On the other hand, the maximum likelihood estimate generally depends on the entire form of the distribution. The observation of Gauss, that the methods of maximum likelihood and least-squares give identical estimates, if the distribution is normal, contain the germ of the modern theory of 'quasi-likelihood'. This will be clear from the subsequent discussion.

Gauss in his astronomical calculations apparently needed 'estimates' as numbers (point estimates) rather than 'intervals'. This is also suggested by the fact that he does *not* write of unbiased estimates but refers to estimates which equal 'true' values when the observations are without errors (Bertrand, 1889; Sprott, 1983).

Laplace, on the other hand, provided 'asymptotics'. He proved that in a certain class of

R. Maller et al. (eds.), *Selected Works of C.C. Heyde*, Selected Works in Probability and Statistics,
DOI 10.1007/978-1-4419-5823-5_49,

symmetric confidence intervals (to use the terminology common now), the one based on the least-squares estimate is asymptotically shortest. For an account of these historical contributions see Heyde & Seneta (1977, Ch. 4).

In the following we, rather informally, state the Gauss–Markov theorem in the framework of estimating function theory. This will help establish the connection of the above-mentioned historical results with what, in this paper, we define as the *quasi-score function* and the *quasi-likelihood.*

Let $y_1, \ldots, y_n$ be independent real random variables jointly distributed with cumulative distribution function F where $F \in \mathcal{F}$. On $\mathcal{F}$ are defined real parameters θ and σ^2 such that for $i = 1, \ldots, n$

$$E_F(y_i - \theta(F)) = 0, \quad E_F(y_i - \theta(F))^2 = \sigma^2(F) \quad (F \in \mathcal{F}),$$

where E_F denotes expectation under F. Now, for any specified numbers α_i $(i = 1, \ldots, n)$, we call

$$g = \sum_{i=1}^{n} (y_i - \theta)\alpha_i \tag{1}$$

a *linear unbiased estimating function*; 'linear' for g is 'linear in $(y_i - \theta)$' $(i = 1, \ldots, n)$, 'unbiased' for $E_F(g) = E_F\{\sum (y_i - \theta(F))\alpha_i\} = 0$, where the sum is over $i = 1, \ldots, n$, $F \in \mathcal{F}$, and 'estimating function' for the equation $g = 0$, when solved for θ, provides an estimate $\hat{\theta}$ for θ. Let $\mathcal{G}$ be the class of all linear unbiased estimating functions subject to the condition $\sum \alpha_i = c$, a constant, where the sum is over $i = 1, \ldots, n$. Now, in this context, an estimating function g^* is said to be *optimal* in $\mathcal{G}$ if $g^* \in \mathcal{G}$ and if, for all $g \in \mathcal{G}$,

$$E_F\{(g^*)^2\} \leq E_F(g^2) \quad (F \in \mathcal{F}). \tag{2}$$

Here this means that the unbiased estimator obtained from g^* has minimal variance among the estimators obtained from estimating functions in $\mathcal{G}$. It is easy to see that *up to a constant multiple* the estimating function g^* in $\mathcal{G}$ satisfying (2) is given by

$$g^* = \bar{y} - \theta, \tag{3}$$

where $\bar{y} = \sum y_i/n$, with the sum over $i = 1, \ldots, n$.

Now suppose that the class of underlying distributions $\mathcal{F} = \{F\}$ is such that it can be written as

$$\mathcal{F} = \{f\} \times \{\theta\}, \tag{4}$$

where 'f' denotes the 'form' of the distribution, and 'θ' the 'parameter'. That is, $\mathcal{F}$ is a 'union' of 'families' of parametric distributions, each family being indexed by the 'same' parameter. In other words, in (4) for each 'f' the 'parameter range' is the same, namely $\{\theta\}$. This covers examples like models with a nuisance parameter or semiparametric models. More conventionally, we may write

$$dF = f_\theta \, d\mu, \tag{5}$$

where f_θ is the density with respect to the measure μ on $\mathbb{R}^n$.

Now we assume, in relation to (5), the existence of the *score function*, namely $\partial \log f_\theta / \partial \theta$, for all θ and f in (4). It can be shown that the 'optimality' of the estimating function g^*, that is g^* in $\mathcal{G}$ satisfying the property (2), is *equivalent* to each of:

(i) $E_F(g^* - \partial \log f_\theta/\partial\theta)^2 \leq E_F(g - \partial \log f_\theta/\partial\theta)^2$,
(ii) $\operatorname{corr}(g^*, \partial \log f_\theta/\partial\theta) \geq \operatorname{corr}(g, \partial \log f_\theta/\partial\theta)$,

for all $g \in \mathcal{G}$ and $F \in \mathcal{F}$, where corr denotes correlation (Godambe, 1985; Godambe &

Thompson, 1985). Here $\partial \log f_\theta / \partial \theta$ is not known, in the sense that it is not known which element of $\{f\}$ is the true one. Further, as an implication of a result proved in this paper, under suitable conditions,

(iii) for large samples the confidence interval for θ associated with g^* is smaller or equal to that associated with g for all $g \in \mathcal{G}$ and $F \in \mathcal{F}$.

Now, because of the properties (i), (ii) and (iii) above of the estimating function g^*, and noting that (iii) is a characteristic property of the score function in a parametric family of distributions (Wilks, 1938), we *define* initially,

$$\text{'}\textit{the quasi-score function}\text{'} \equiv \text{'}g^*\text{'}, \tag{6}$$

(noting that g^* is defined only up to a constant multiple). The *quasi-likelihood* is obtained by the appropriate integration of (6).

Gauss's justification for the 'normal model', discussed above, would seem to be fairly in line with our definition of 'quasi-likelihood' if we put, as in (3), $g^* = \bar{y} - \theta$ in (6).

In the following sections, the above concepts of quasi-score function and quasi-likelihood are cast in the general setting of stochastic processes. Within this general framework, it is shown that the confidence intervals obtained by the inversion of the appropriately standardized optimal estimating function (i.e. quasi-score function) are asymptotically smallest, generalizing the corresponding result of Laplace, in relation to the least-squares estimator, mentioned above.

The present-day usage of the term 'quasi-likelihood' seems to have been initiated by Wedderburn (1974). Deeply intuitively, he observed that the expression $\bar{y} - \theta$ in (3), or a similar expression in a more general setting, possesses mathematical properties of a 'score function'. The statistical significance of this mathematical fact is indicated by the term 'quasi-likelihood'. The underlying idea comes from the scalar linear regression model with independent error term, whose distribution is of exponential family type. Here, as noted by Bradley (1973) and Wedderburn (1974), the true score function depends on the parameters only through the means and variances. These authors further noted that this score function can also be written as a weighted least-squares estimating function. To estimate the parameters, Wedderburn (1974) *suggested* using the exponential family score function, even when the 'error' distribution is *unspecified*. Clearly then, the estimation depends on the first two moments only, as in the Gauss–Markov theorem. But this suggestion raises the *question* of what happens when the true underlying distribution is *not* from the exponential family? Both the suggestion and the question can be better formulated, extended and answered by relating the 'quasi-likelihood' to the 'optimal estimating function' as in (6) in view of its properties (i), (ii) and (iii). This is the subject of the remainder of the paper.

2 The general framework

Let $\{\mathbf{X}_t, 0 \leq t \leq T\}$ be a sample in discrete or continuous time drawn from some process taking values in r-dimensional Euclidean space whose distribution depends on a parameter $\boldsymbol{\theta}$ taking values in an open subset Θ of p-dimensional Euclidean space. Suppose that the possible probability measures for $\{\mathbf{X}_t\}$ are $\mathcal{P} = \{\mathcal{P}_{\boldsymbol{\theta}}\}$, a *union* of parametric families each family being indexed by the same parameter $\boldsymbol{\theta}$, and that each $(\Omega, \mathcal{B}, P_{\boldsymbol{\theta}})$ is a complete probability space. Let $\{\mathcal{B}_t, t \geq 0\}$ denote a standard filtration, $\mathcal{B}_s \subseteq \mathcal{B}_t \subseteq \mathcal{B}$ for $s \leq t$; $\mathcal{B}_0$ is augmented by sets of measure zero of $\mathcal{B}$ and $\mathcal{B}_t = \mathcal{B}_{t+}$, where $\mathcal{B}_{t+} = \bigcap_{s>t} \mathcal{B}_s$.

We shall focus attention on the class $\mathcal{G}$ of zero mean, square integrable estimating

functions $\mathbf{G}_T = \mathbf{G}_T(\{\mathbf{X}_t, 0 \leq t \leq T\}, \boldsymbol{\theta})$ for which $E\mathbf{G}_T(\boldsymbol{\theta}) = \mathbf{0}$ for each $P_{\boldsymbol{\theta}} \in \mathcal{P}$ with index $\boldsymbol{\theta}$ and on the class $\mathcal{M} \subset \mathcal{G}$ of martingale estimating functions $\{\mathbf{G}_T, \mathcal{B}_T\}$ which are martingales for each $P_{\boldsymbol{\theta}} \in \mathcal{P}$ with index $\boldsymbol{\theta}$. Here $\mathbf{G}_T$ is a vector of dimension p. Estimators $\boldsymbol{\theta}^*$ are found by solving the estimating equation $\mathbf{G}_T(\boldsymbol{\theta}^*) = \mathbf{0}$.

We should emphasize that this framework includes the standard methods of estimation: maximum likelihood, least-squares, conditional least-squares, minimum chi-squared, etc., under minor regularity conditions. Indeed, it is essentially the class of martingale M-estimators, albeit in a more general setting than the usual one; see for example, Basawa (1985).

The estimating function approach focuses on the function rather than the estimator derived from it and optimality properties are for the function. There is often loss of information about the parameter in moving from estimating function to estimator. For example, it is well known that, under minor regularity conditions, the score function provides a minimal sufficient partitioning of the sample space. However, there is often no single sufficient statistic.

The properties of estimators derived from optimal estimating functions can be investigated in detail for large samples but little explicit information can be given in general about estimators from small samples. This is a manifestation in a broader setting of the fact that, outside the field of sufficient statistics, the optimal properties of maximum likelihood (ML) estimators are asymptotic ones.

For a matrix $\mathbf{A}$ we shall write $\mathbf{A}'$ for the transpose and, if $\mathbf{A}$ is square, $|\mathbf{A}|$ for the determinant. If $\mathbf{A}$ is symmetric and nonnegative-definite its unique nonnegative-definite square root will be written as $\mathbf{A}^{\frac{1}{2}}$, while $\mathbf{A}^+$ denotes its Moore–Penrose pseudo inverse, namely the unique matrix possessing the properties $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$, $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$, $\mathbf{A}^+\mathbf{A} = \mathbf{A}\mathbf{A}^+$.

For $n \times 1$ vector valued martingales $\mathbf{M}_T$ and $\mathbf{N}_T$, the $n \times n$ process $\langle \mathbf{M}, \mathbf{N}' \rangle_T$ is the *mutual quadratic characteristic*, a predictable increasing process such that $\mathbf{M}_T\mathbf{N}_T' - \langle \mathbf{M}, \mathbf{N}' \rangle_T$ is an $n \times n$ martingale. We shall write $\langle \mathbf{M} \rangle_T$ for $\langle \mathbf{M}, \mathbf{M}' \rangle_T$, the *quadratic characteristic* of $\mathbf{M}_T$. A convenient sketch of these concepts is given by Shiryaev (1981).

3 Fixed sample criteria

Here we shall confine attention to the subclass $\mathcal{G}_1$ of $\mathcal{G}$ of estimating functions $\mathbf{G}_T$ which are almost surely differentiable with respect to the components of $\boldsymbol{\theta}$, and for which $E\dot{\mathbf{G}}_T(\boldsymbol{\theta}) = (E\,\partial G_{T,i}(\boldsymbol{\theta})/\partial\theta_j)$ and $E(\mathbf{G}_T(\boldsymbol{\theta})\mathbf{G}_T'(\boldsymbol{\theta}))$ are nonsingular. The expectations are always with respect to $P_{\boldsymbol{\theta}}$.

We shall usually suppose that $P_{\boldsymbol{\theta}}$ is absolutely continuous with respect to some σ-finite measure λ_t giving density $p_t(\boldsymbol{\theta})$. Then, we write $\mathbf{U}_t(\boldsymbol{\theta}) = p_t^{-1}(\boldsymbol{\theta})\dot{\mathbf{p}}_t(\boldsymbol{\theta})$ for the score function which we further suppose to be almost surely differentiable with respect to the components of $\boldsymbol{\theta}$. In addition we suppose that differentiation and integration can be interchanged in $E(\mathbf{G}_T\mathbf{U}_T')$ and $E(\mathbf{U}_T\mathbf{G}_T')$, for $\mathbf{G}_T \in \mathcal{G}_1$.

Here we note that corresponding to each parametric family in $\mathcal{P}$ is defined a score function. When the true underlying parametric family and hence the score function is unknown it may be argued that the best strategy is to choose an estimating function $\mathbf{G}_T$ which has minimum distance, in an appropriate sense, from $\mathbf{U}_T$ or maximum vector correlation with $\mathbf{U}_T$. These ideas are formalized in the following equivalent properties of an optimal estimating function $\mathbf{G}_T^*$.

Criterion 1. We have that

$$(E\dot{\mathbf{G}}_T)^{-1}(E\mathbf{G}_T\mathbf{G}_T')((E\dot{\mathbf{G}}_T)^{-1})' - (E\dot{\mathbf{G}}_T^*)^{-1}(E\mathbf{G}_T^*\mathbf{G}_T^{*\prime})((E\dot{\mathbf{G}}_T^*)^{-1})'$$

is nonnegative-definite for all $\mathbf{G}_T \in \mathcal{G}_1$, $\boldsymbol{\theta} \in \Theta$ and $P_{\boldsymbol{\theta}} \in \mathcal{P}$.

Criterion 2. For some fixed matrix function $\boldsymbol{\alpha}$ depending on $\boldsymbol{\theta}$ and T,

$$E((\boldsymbol{\alpha}_T\mathbf{U}_T - \mathbf{G}_T^*)\mathbf{G}_T') = E(\mathbf{G}_T(\boldsymbol{\alpha}_T\mathbf{U}_T - \mathbf{G}_T^*)') = \mathbf{0}$$

for every $\mathbf{G}_T \in \mathcal{G}_1$ for all $\boldsymbol{\theta} \in \Theta$ and $P_{\boldsymbol{\theta}} \in \mathcal{P}$.

Criterion 3. For some fixed matrix function $\boldsymbol{\alpha}$ depending on $\boldsymbol{\theta}$ and T,

$$E(\boldsymbol{\alpha}_T\mathbf{U}_T - \mathbf{G}_T)(\boldsymbol{\alpha}_T\mathbf{U}_T - \mathbf{G}_T)' - E(\boldsymbol{\alpha}_T\mathbf{U}_T - \mathbf{G}_T^*)(\boldsymbol{\alpha}_T\mathbf{U}_T - \mathbf{G}_T^*)'$$

is nonnegative-definite for all $\mathbf{G}_T \in \mathcal{G}_1$, $\boldsymbol{\theta} \in \Theta$ and $P_{\boldsymbol{\theta}} \in \mathcal{P}$.

The proof of the equivalence of these criteria, which extends the results of Godambe (1985), Godambe & Thompson (1985) and Thavaneswaran & Thompson (1986) to the vector case, is given in the Appendix. Note from Criterion 1 that an optimal estimating function is defined only up to a constant (matrix) multiplier, while Criterion 3 can be interpreted as $\mathbf{G}_T^*$ having minimum dispersion distance from $\boldsymbol{\alpha}_T\mathbf{U}_T$. Criterion 1 can also be inverted to give an equivalent form as follows.

Criterion 1*. We have that

$$(E\dot{\mathbf{G}}_T^*)'(E\mathbf{G}_T^*\mathbf{G}_T^{*\prime})^{-1}(E\dot{\mathbf{G}}_T^*) - (E\dot{\mathbf{G}}_T)'(E\mathbf{G}_T\mathbf{G}_T')^{-1}(E\dot{\mathbf{G}}_T)$$

is nonnegative-definite for all $\mathbf{G}_T \in \mathcal{G}_1$, $\boldsymbol{\theta} \in \Theta$ and $P_{\boldsymbol{\theta}} \in \mathcal{P}$.

We shall subsequently use this form. It should be noted that Criterion 1* (or 1) does not require the existence of the score function and this is taken as the general *defining* relationship for an optimal estimating function for fixed samples, denoted by O_F.

In particular, if

$$(E\dot{\mathbf{G}}_T^*)'(E\mathbf{G}_T^*\mathbf{G}_T^{*\prime})^{-1}(E\dot{\mathbf{G}}_T^*) - (E\dot{\mathbf{G}}_T)'(E\mathbf{G}_T\mathbf{G}_T')^{-1}(E\dot{\mathbf{G}}_T)$$

is nonnegative-definite for all $\mathbf{G}_T \in \mathcal{G}_2 \subset \mathcal{G}_1$, $\boldsymbol{\theta} \in \Theta$ and $P_{\boldsymbol{\theta}} \in \mathcal{P}$, we will say that $\mathbf{G}_T^*$ is *O_F-optimal within $\mathcal{G}_2$*. We emphasize that Criteria 1, 2 and 3 are not necessarily equivalent if consideration is restricted to a subset $\mathcal{G}_2 \subset \mathcal{G}_1$.

Now the vector correlation which measures the association between $\mathbf{G}_T = (G_{T,1}, \ldots, G_{T,p})'$ and $\mathbf{U}_T = (U_{T,1}, \ldots, U_{T,p})'$, defined, for example, by Hotelling (1936), is

$$\rho^2 = \frac{|E(\mathbf{G}_T\mathbf{U}_T')|^2}{|E(\mathbf{G}_T\mathbf{G}_T')|\,|E(\mathbf{U}_T\mathbf{U}_T')|}.$$

However, under the regularity conditions that have been imposed, $E\dot{\mathbf{G}}_T = -E(\mathbf{G}_T\mathbf{U}_T')$, so a maximal correlation requirement is to maximize

$$|E\dot{\mathbf{G}}_T|^2/|E\mathbf{G}_T\mathbf{G}_T'|, \tag{7}$$

which can be achieved by maximizing

$$(E\dot{\mathbf{G}}_T)'(E\mathbf{G}_T\mathbf{G}_T')^{-1}(E\dot{\mathbf{G}}_T) \tag{8}$$

in the (partial) order of nonnegative-definite matrices. This corresponds to Criterion 1*.

The single-parameter optimality criterion, obtained from (7), is originally due to Godambe (1960). Multiparametric optimality criteria corresponding to (7) and (8) above were previously proposed by Durbin (1960), Bhapkar (1972), Morton (1981), Ferreira (1982), Chandrasekar & Kale (1984) and Godambe & Thompson (1986).

Following discussion in § 1, we call the optimal estimating function $\mathbf{G}_T^*$ the *quasi-score*

(QS) *function* and the corresponding equation $\mathbf{G}_T^* = \mathbf{0}$ the *quasi-likelihood* (QL) *equation*. A solution of the QL equation would be called a *maximum quasi-likelihood estimator* (MQLE).

Optimal estimating functions in the sense of Criterion 1* do not necessarily exist for the full class $\mathcal{G}_1$; see, for example, Godambe & Thompson (1985). However, by restricting consideration to a statistically meaningful subclass of $\mathcal{G}_1$, such as in (13) of § 5 below, an optimum may be obtained.

4 An asymptotic criterion

Let $\mathcal{M}_1$ denote the subset of $\mathcal{G}_1$ which are square integrable martingales. For $\{\mathbf{G}_T, \mathcal{B}_T\} \in \mathcal{M}_1$ there is, under quite broad conditions, a multivariate central limit result

$$\langle \mathbf{G} \rangle_T^{-\frac{1}{2}} \mathbf{G}_T \to MVN(\mathbf{0}, \mathbf{I}_p) \tag{9}$$

in distribution, as $T \to \infty$; see, for example, Feigin (1985), Hall & Heyde (1980, Ch. 3), with the Cramér–Wold device in mind to obtain multivariate versions, and Hutton & Nelson (1984). Let $\mathcal{M}_2 \subset \mathcal{M}_1$ be the subclass for which (9) obtains.

Next, with $\mathbf{G}_T \in \mathcal{M}_2$ let $\boldsymbol{\theta}^*$ be a solution of $\mathbf{G}_T(\boldsymbol{\theta}) = \mathbf{0}$ and use Taylor's expansion to obtain

$$\mathbf{0} = \mathbf{G}_T(\boldsymbol{\theta}^*) = \mathbf{G}_T(\boldsymbol{\theta}) + \dot{\mathbf{G}}_T(\boldsymbol{\theta}\dagger)(\boldsymbol{\theta}^* - \boldsymbol{\theta}), \tag{10}$$

where $\|\boldsymbol{\theta} - \boldsymbol{\theta}\dagger\| < \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|$, the norm denoting sum of squares of elements. Then, if $\dot{\mathbf{G}}_T(\boldsymbol{\theta})$ is nonsingular for $\boldsymbol{\theta}$ in a suitable neighbourhood and $(\dot{\mathbf{G}}_T(\boldsymbol{\theta}\dagger))^{-1}\dot{\mathbf{G}}_T(\boldsymbol{\theta}) \to \mathbf{I}_p$ in probability as $T \to \infty$, expressions (9) and (10) lead to

$$\langle \mathbf{G}(\boldsymbol{\theta}) \rangle_T^{-\frac{1}{2}} \dot{\mathbf{G}}_T(\boldsymbol{\theta})(\boldsymbol{\theta}^* - \boldsymbol{\theta}) \to MVN(\mathbf{0}, \mathbf{I}_p)$$

in distribution.

Now define the predictable process

$$\bar{\mathbf{G}}_t(\boldsymbol{\theta}) = \int_0^t E(d\dot{\mathbf{G}}_s(\boldsymbol{\theta}) \mid \mathcal{B}_{s-}),$$

$\mathcal{B}_{s-}$ being the σ-field generated by $\bigcup_{r<s} \mathcal{B}_r$, and assume that $\dot{\mathbf{G}}_T(\boldsymbol{\theta})$ admits a Doob–Meyer type decomposition

$$\dot{\mathbf{G}}_T(\boldsymbol{\theta}) = \mathbf{M}_{\mathbf{G},T}(\boldsymbol{\theta}) + \bar{\mathbf{G}}_T(\boldsymbol{\theta}),$$

$\mathbf{M}_{\mathbf{G},T}(\boldsymbol{\theta})$ being a martingale. Then, under modest conditions, for example if $\dot{\mathbf{G}}_T(\boldsymbol{\theta}) \to -\infty$ almost surely as $T \to \infty$,

$$|\mathbf{M}_{\mathbf{G},T}(\boldsymbol{\theta})| = o_p(|\bar{\mathbf{G}}_T(\boldsymbol{\theta})|)$$

as $T \to \infty$, o_p denoting small order in probability.

Thus, considerations which can be formalized under appropriate regularity conditions indicate that, for $\mathbf{G}_T$ belonging to some $\mathcal{M}_3 \subset \mathcal{M}_2$,

$$\langle \mathbf{G}(\boldsymbol{\theta}) \rangle_T^{-\frac{1}{2}} \bar{\mathbf{G}}_T(\boldsymbol{\theta})(\boldsymbol{\theta}^* - \boldsymbol{\theta}) \to MVN(\mathbf{0}, \mathbf{I}_p)$$

in distribution, and hence

$$(\boldsymbol{\theta}^* - \boldsymbol{\theta})' \bar{\mathbf{G}}_T'(\boldsymbol{\theta}) \langle \mathbf{G}(\boldsymbol{\theta}) \rangle_T^{-1} \bar{\mathbf{G}}_T(\boldsymbol{\theta})(\boldsymbol{\theta}^* - \boldsymbol{\theta}) \to \chi_p^2 \tag{11}$$

in distribution. Best (asymptotic) estimation within a class $\mathcal{M}_4 \subseteq \mathcal{M}_3$ of estimating functions is then achieved by choosing $\mathbf{G}_T^* \in \mathcal{M}_4$ so that Criterion 4 is satisfied.

Criterion 4. We have that

$$\bar{\mathbf{G}}_T^{*\prime}(\boldsymbol{\theta})\langle \mathbf{G}^*(\boldsymbol{\theta})\rangle_T^{-1}\bar{\mathbf{G}}_T^*(\boldsymbol{\theta}) - \bar{\mathbf{G}}_T'(\boldsymbol{\theta})\langle \mathbf{G}(\boldsymbol{\theta})\rangle_T^{-1}\bar{\mathbf{G}}_T(\boldsymbol{\theta})$$

is nonnegative-definite for all $\mathbf{G}_T \in \mathcal{M}_4$, $\boldsymbol{\theta} \in \Theta$, $P_{\boldsymbol{\theta}} \in \mathcal{P}$ and $T > 0$.

This follows because maximizing

$$\mathbf{C}_T(\boldsymbol{\theta}) = \bar{\mathbf{G}}_T'(\boldsymbol{\theta})\langle \mathbf{G}(\boldsymbol{\theta})\rangle_T^{-1}\bar{\mathbf{G}}_T(\boldsymbol{\theta})$$

leads to (asymptotic) confidence regions centred on $\boldsymbol{\theta}^*$ of minimum size, for example Rao (1965, § 4b.2). Note that $\mathbf{C}_T(\boldsymbol{\theta})$ can be replaced by $\mathbf{C}_T(\boldsymbol{\theta}^*)$ in (11). If $\mathbf{G}_T^*$ satisfies Criterion 4 we will say that it is *O_A-optimal within $\mathcal{M}_4$*, O_A meaning optimal in the asymptotic sense.

The relation between O_F and O_A optimality, both restricted to the same class of estimating functions, is very close and it should be noted that

$$E\bar{\mathbf{G}}_T(\boldsymbol{\theta}) = E\dot{\mathbf{G}}_T(\boldsymbol{\theta}), \quad E\langle \mathbf{G}(\boldsymbol{\theta})\rangle_T = E\mathbf{G}_T(\boldsymbol{\theta})\mathbf{G}_T'(\boldsymbol{\theta}).$$

Both criteria are satisfied by the same (quasi-score) estimating function under broad conditions as we shall indicate below.

5 A particular QL solution

We shall illustrate the use of the criteria for O_F and O_A optimality by supposing that the observed process $\{\mathbf{X}_t, t \geqslant 0\}$ is (a semimartingale) representable for all T in the form

$$\mathbf{X}_T = \int_0^T \mathbf{f}_t(\boldsymbol{\theta})\, d\lambda_t + \mathbf{m}_T(\boldsymbol{\theta}). \tag{12}$$

Here $\{\lambda_t\}$ is a real, monotone increasing, right continuous process with $\lambda_0 = 0$, while $\{\mathbf{m}_T(\boldsymbol{\theta}), \mathcal{B}_T\}$ is a cadlag, square integrable $r \times 1$ vector martingale with characteristic $\langle \mathbf{m}(\boldsymbol{\theta})\rangle_T$ given for all T by

$$\langle \mathbf{m}(\boldsymbol{\theta})\rangle_T = \int_0^T \mathbf{a}_t(\boldsymbol{\theta})\, d\lambda_t,$$

the processes $\{\lambda_t\}$, $\{\mathbf{a}_t(\boldsymbol{\theta})\}$ and $\{\mathbf{f}_t(\boldsymbol{\theta})\}$ being predictable and the elements of $\mathbf{f}_t(\boldsymbol{\theta})$ being almost surely continuously differentiable with respect to the elements of $\boldsymbol{\theta}$. This is the setting used by Hutton & Nelson (1986) and it is very widely applicable. It covers many continuous time stochastic models of the diffusion, counting, queueing or population process type as well as most discrete time models via the use of $\lambda_t = [t]$, the greatest integer less than or equal to t.

Now in the context of (12), utilizing the motivation provided by Godambe (1985, § 4), we shall confine attention to estimating functions of the form

$$\mathbf{G}_T(\boldsymbol{\theta}) = \int_0^T \boldsymbol{\alpha}_s(\boldsymbol{\theta})\, d\mathbf{m}_s(\boldsymbol{\theta}), \tag{13}$$

where $\{\boldsymbol{\alpha}_s(\boldsymbol{\theta})\}$ is a $p \times r$ predictable process whose elements are almost surely continuously differentiable with respect to the components of $\boldsymbol{\theta}$. We shall proceed to find an estimator which is optimal in both the O_F and O_A senses within this class. It should be remarked, however, that this discussion is illustrative and cases in which both O_F and O_A optimality are satisfied are certainly not peculiar to the semimartingale setting. This point is discussed further in § 6.

We have, under appropriate regularity conditions, that

$$\dot{\mathbf{G}}_T(\boldsymbol{\theta}) = \int_0^T \dot{\boldsymbol{\alpha}}_s(\boldsymbol{\theta})\, d\mathbf{m}_s(\boldsymbol{\theta}) + \int_0^T \boldsymbol{\alpha}_s(\boldsymbol{\theta})\, d\dot{\mathbf{m}}_s(\boldsymbol{\theta})$$

and, since $\dot{\mathbf{m}}_s(\theta)$ is $\mathcal{B}_{s-}$-measurable from (12),

$$\bar{\mathbf{G}}_T(\boldsymbol{\theta}) = \int_0^T \boldsymbol{\alpha}_s(\boldsymbol{\theta})\, d\dot{\mathbf{m}}_s(\boldsymbol{\theta}) = -\int_0^T \boldsymbol{\alpha}_s(\boldsymbol{\theta})\dot{\mathbf{f}}_s(\boldsymbol{\theta})\, d\lambda_s.$$

Furthermore,

$$\langle \mathbf{G}(\boldsymbol{\theta})\rangle_T = \int_0^T \boldsymbol{\alpha}_s(\boldsymbol{\theta}) \frac{d\langle \mathbf{m}(\boldsymbol{\theta})\rangle_s}{d\lambda_s} \boldsymbol{\alpha}_s'(\boldsymbol{\theta})\, d\lambda_s,$$

and, more generally, if $\mathbf{H}_T(\boldsymbol{\theta}) = \int \boldsymbol{\beta}_s(\boldsymbol{\theta})\, d\mathbf{m}_s(\boldsymbol{\theta})$, where the integral is over $(0, T)$, and where $\{\boldsymbol{\beta}_s(\boldsymbol{\theta})\}$ is predictable, then

$$\langle \mathbf{G}(\boldsymbol{\theta}), \mathbf{H}'(\boldsymbol{\theta})\rangle_T = \int_0^T \boldsymbol{\alpha}_s(\boldsymbol{\theta}) \frac{d\langle \mathbf{m}(\boldsymbol{\theta})\rangle_s}{d\lambda_s} \boldsymbol{\beta}_s'(\boldsymbol{\theta})\, d\lambda_s.$$

Of course

$$E\langle \mathbf{G}(\boldsymbol{\theta}), \mathbf{H}'(\boldsymbol{\theta})\rangle_T = E\mathbf{G}_T(\boldsymbol{\theta})\mathbf{H}_T'(\boldsymbol{\theta}), \quad E\bar{\mathbf{G}}_T(\boldsymbol{\theta}) = E\dot{\mathbf{G}}_T(\boldsymbol{\theta}).$$

We shall suppose that $\langle \mathbf{G}(\boldsymbol{\theta})\rangle_T$ and $\bar{\mathbf{G}}_T(\boldsymbol{\theta})$ are almost surely nonsingular and that their expected values are nonsingular.

Now we introduce the estimating function $\mathbf{Q}_T$ defined by

$$\mathbf{Q}_T(\boldsymbol{\theta}) = \int_0^T \dot{\mathbf{f}}_s'(\boldsymbol{\theta})\mathbf{a}_s^+(\boldsymbol{\theta})\, d\mathbf{m}_s(\boldsymbol{\theta}). \tag{14}$$

We shall show that this estimating function is optimal in the sense of the criteria for both O_F and O_A optimality within the class of estimating functions of the form (13). This justifies the use of the terminology *quasi-score* function for $\mathbf{Q}_T$. It is useful, where possible, to interpret the quasi-score function as the derivative of an underlying log quasi-likelihood whose maximum provides the quasi-likelihood estimator. In fact, it is frequently, but not always, possible to obtain (14) as the true score function for members of a certain exponential family. This important topic will be taken up elsewhere.

Note that, if $\mathbf{a}_s(\boldsymbol{\theta})$ is positive-definite,

$$\langle \mathbf{G}(\boldsymbol{\theta}), (\mathbf{Q}(\boldsymbol{\theta}))'\rangle_T = -\bar{\mathbf{G}}_T(\boldsymbol{\theta}),$$

while

$$\langle \mathbf{Q}(\boldsymbol{\theta})\rangle_T = \int_0^T \dot{\mathbf{f}}_s'(\boldsymbol{\theta})\mathbf{a}_s^+(\boldsymbol{\theta})\dot{\mathbf{f}}_s(\boldsymbol{\theta})\, d\lambda_s = -\bar{\mathbf{Q}}_T(\boldsymbol{\theta}).$$

Next, write $\mathbf{G}_T = (G_{1,T}, \ldots, G_{p,T})'$, $\mathbf{Q}_T = (Q_{1,T}, \ldots, Q_{p,T})'$. The $2p \times 2p$ dispersion and mutual quadratic characteristic matrices of the set of variables

$$(G_{1,T}, \ldots, G_{p,T}, Q_{1,T}, \ldots, Q_{p,T})$$

may be written in partitioned matrix form as

$$\mathbf{D} = \begin{pmatrix} E\mathbf{G}_T\mathbf{G}_T' & E\mathbf{G}_T\mathbf{Q}_T' \\ (E\mathbf{G}_T\mathbf{Q}_T')' & E\mathbf{Q}_T\mathbf{Q}_T' \end{pmatrix} = \begin{pmatrix} E\mathbf{G}_T\mathbf{G}_T' & -E\dot{\mathbf{G}}_T \\ (-E\dot{\mathbf{G}}_T)' & E\mathbf{Q}_T\mathbf{Q}_T' \end{pmatrix},$$

$$\mathbf{C} = \begin{pmatrix} \langle \mathbf{G}\rangle_T & \langle \mathbf{G}, \mathbf{Q}'\rangle_T \\ \langle \mathbf{G}, \mathbf{Q}'\rangle_T' & \langle \mathbf{Q}\rangle_T \end{pmatrix} = \begin{pmatrix} \langle \mathbf{G}\rangle_T & -\bar{\mathbf{G}}_T \\ -\bar{\mathbf{G}}_T' & \langle \mathbf{Q}\rangle_T \end{pmatrix},$$

respectively, the former matrix being the expected value of the latter.

Now $\mathbf{C}$ and $\mathbf{D}$ are nonnegative-definite since, if

$$\mathbf{Z}' = (G_{1,T}, \ldots, G_{p,T}, Q_{1,T}, \ldots, Q_{p,T})$$

and $\mathbf{u}$ is an arbitrary $2p \times 1$ vector,

$$\mathbf{u}'\mathbf{D}\mathbf{u} = \mathbf{u}'E(\mathbf{Z}\mathbf{Z}')\mathbf{u} = E(\mathbf{u}'\mathbf{Z}\mathbf{Z}'\mathbf{u}) = E(\mathbf{Z}'\mathbf{u})^2 \geq 0,$$

while, almost surely,

$$\mathbf{u}'\mathbf{C}\mathbf{u} = \mathbf{u}'\langle \mathbf{Z}, \mathbf{Z}' \rangle_T \mathbf{u} = \langle \mathbf{u}'\mathbf{Z}, \mathbf{Z}'\mathbf{u} \rangle_T = \langle \mathbf{u}'\mathbf{Z} \rangle_T \geq 0,$$

and the method of Rao (1965, p. 266) gives the nonnegative-definiteness of

$$E\mathbf{G}_T\mathbf{G}_T' - (E\dot{\mathbf{G}}_T)(E\mathbf{Q}_T\mathbf{Q}_T')^{-1}(E\dot{\mathbf{G}}_T)',$$

and, almost surely, $\langle \mathbf{G} \rangle_T - \bar{\mathbf{G}}_T \langle \boldsymbol{Q} \rangle_T^{-1} \bar{\mathbf{G}}_T'$. Furthermore, since all the matrices involved in these expressions are almost surely nonsingular, we obtain by inversion that

$$E\mathbf{Q}_T\mathbf{Q}_T' - (E\dot{\mathbf{G}}_T)'(E\mathbf{G}_T\mathbf{G}_T')^{-1}(E\dot{\mathbf{G}}_T),$$

and, almost surely, $\langle \mathbf{Q} \rangle_T - \bar{\mathbf{G}}_T' \langle \mathbf{G} \rangle_T^{-1} \bar{\mathbf{G}}_T$ are nonnegative-definite.

Then, since

$$E\mathbf{Q}_T\mathbf{Q}_T' = -E\dot{\mathbf{Q}}_T, \quad \langle \mathbf{Q} \rangle_T = -\bar{\mathbf{Q}}_T,$$

both $(E\dot{\mathbf{G}}_T)'(E\mathbf{G}_T\mathbf{G}_T')^{-1}(E\dot{\mathbf{G}}_T)$ and $\bar{\mathbf{G}}_T' \langle \mathbf{G} \rangle_T^{-1} \bar{\mathbf{G}}_T$ are maximized in the partial order of nonnegative-definite matrices by choosing $\mathbf{G}_T = \mathbf{Q}_T$. Thus, the estimating function defined by (14) is optimal in both the senses of O_F and O_A optimality as required.

The maximum quasi-likelihood estimator (MQLE) is a solution of the QL estimating equation and general properties of the MQLE can be established along similar lines to those for the MLE which they closely resemble. Sufficient conditions for strong consistency and asymptotic normality are given in Theorems 3.1 and 4.1, respectively, of Hutton & Nelson (1986).

The framework discussed above specializes to what has been described as quasi-likelihood in the context of the general linear model, for example, McCullagh (1983) and McCullagh & Nelder (1983, Ch. 8) provided the $\mathbf{X}_t$ are sums of independent random vectors $\mathbf{Y}_s$, for $1 \leq s \leq t$. Then

$$\mathbf{m}_T(\boldsymbol{\theta}) = \sum_{t=1}^{T} (\mathbf{Y}_t - \boldsymbol{\mu}_t(\boldsymbol{\theta})),$$

where $\boldsymbol{\mu}_t(\boldsymbol{\theta}) = E\mathbf{Y}_t$, and, if

$$E(\mathbf{Y}_t - \boldsymbol{\mu}_t(\boldsymbol{\theta}))(\mathbf{Y}_t - \boldsymbol{\mu}_t(\boldsymbol{\theta}))' = \mathbf{v}_t(\boldsymbol{\theta}) \quad (t = 1, 2, \ldots, T),$$

then the QL estimating function given by (14) is

$$\mathbf{Q}_T(\boldsymbol{\theta}) = \sum_{t=1}^{T} \dot{\boldsymbol{\mu}}_t'(\boldsymbol{\theta})\mathbf{v}_t^+(\boldsymbol{\theta})(\mathbf{Y}_t - \boldsymbol{\mu}_t(\boldsymbol{\theta})) = \dot{\boldsymbol{\mu}}'\mathbf{V}^+(\mathbf{Y} - \boldsymbol{\mu}),$$

where

$$\boldsymbol{\mu}' = (\boldsymbol{\mu}_1'(\boldsymbol{\theta}) \vdots \ldots \vdots \boldsymbol{\mu}_T'(\boldsymbol{\theta})), \quad \mathbf{Y}' = (\mathbf{Y}_1' \vdots \ldots \vdots \mathbf{Y}_T'), \quad \mathbf{V}^+ = \operatorname{diag}(\mathbf{v}_1^+(\boldsymbol{\theta}) \vdots \ldots \vdots \mathbf{v}_T^+(\boldsymbol{\theta})).$$

This agrees with the usual definition but the extensions that have been used for the generalized linear model cannot be used directly to deal with most cases of dependent variables.

A simple example of dependent variables for which the QL formalism discussed herein is essential is given by the estimation of the mean of the offspring distribution in a single

type (Bienaymé–) Galton–Watson branching process $\{Z_0 = 1, Z_1, \ldots, Z_T\}$. Here we have $EZ_1 = \theta$ and we assume var $Z_1 = \sigma^2 < \infty$.

We take $X_T = \sum Z_t$, $m_T(\theta) = \sum (Z_t - \theta Z_{t-1})$, and then $f_t(\theta) = \theta Z_{t-1}$ and $\langle m \rangle_T = \sigma^2 \sum Z_{t-1}$, where the sums are all over $t = 1, \ldots, T$, so that $a_t(\theta) = \sigma^2 Z_{t-1}$. Here the quasi-score function given by (14) is $\sigma^{-2} \sum (Z_t - \theta Z_{t-1})$, with the sum over $t = 1, \ldots, T$. The maximum quasi-likelihood estimating equation is then

$$\sum_{t=1}^{T} (Z_t - \hat{\theta} Z_{t-1}) = 0.$$

The estimator $\hat{\theta}$ so obtained is easily seen to be the maximum likelihood estimator (MLE) for the power series family of offspring distributions:

$$P(Z_1 = j) = A(j) \frac{(a(\theta))^j}{F(\theta)} \quad (j = 0, 1, 2, \ldots),$$

where $F(\theta) = \sum A(j)(a(\theta))^j$, with the sum over $j = 0, \ldots, \infty$, for example, Heyde (1975). These form the discrete linear exponential family for this context. The estimator $\hat{\theta}$ can also be obtained quite generally as a nonparametric MLE, (Feigin, 1977).

As another example, now in continuous time, we consider the multivariate counting process $\mathbf{X}_T = (X_{1,T}, \ldots, X_{p,T})'$ such that each $X_{i,T}$ is of the form

$$X_{i,T} = \theta_i \int_0^T J_i(s)\, ds + M_{i,T}$$

with multiplicative intensity $\Lambda_i(t) = \theta_i J_i(t)$, $J_i(t) > 0$ almost surely being a predictable process, and $M_{i,t}$ a square integrable martingale. This is a special case of the framework considered by Aalen (1978), see also Andersen et al. (1982), and it covers a variety of contexts for processes such as those of birth and death type. The case $p = 1$ has been discussed by Thavaneswaran & Thompson (1986).

Now for counting processes,

$$\langle M_i \rangle_T = A_{i,T} = \theta_i \int_0^T J_i(s)\, ds,$$

$A_{i,T}$ being a (nondecreasing) compensator and $\langle M_i, M_j \rangle_T = 0$ whenever $i \neq j$ (Aalen, 1978, Th. 3.2). Thus we can write

$$\mathbf{X}_T = \left(\int_0^T \mathbf{J}(s)\, ds \right) \boldsymbol{\theta} + \mathbf{M}_T,$$

where

$$\mathbf{J}(s) = \text{diag}\,(J_1(s), \ldots, J_p(s)), \quad \boldsymbol{\theta}' = (\theta_1, \ldots, \theta_p), \quad \mathbf{M}_T = (M_{1,T}, \ldots, M_{p,T})',$$

so that $\mathbf{f}_t(\boldsymbol{\theta}) = \mathbf{J}(t)\boldsymbol{\theta}$ and, since $\langle \mathbf{M} \rangle_T = \text{diag}\,(A_{1,T}, \ldots, A_{p,T})$,

$$a_t = \text{diag}\,(\theta_1 J_1(t), \ldots, \theta_p J_p(t)).$$

The MQLE is then given by $\mathbf{X}_T = \hat{\boldsymbol{\theta}} \int \mathbf{J}(s)\, ds$, where the integral is over $(0, T)$. That this $\hat{\boldsymbol{\theta}}$ is also the MLE follows from § 3.3 of Aalen (1978). The simplest particular case is where each $X_{i,t}$ is a Poisson process with parameter θ_i.

6 Extensions

The results of § 5 are focused on the case where the observed process $\{X_t, t \geq 0\}$ is a semimartingale of the form (12) but it is not essential to restrict consideration to such a

setting. All that is required is the choice of a basic martingale $\{\mathbf{m}_T(\boldsymbol{\theta}), \mathcal{B}_T\}$ and then a quasi-score estimating function can be chosen from amongst the competitors of the form $\int \boldsymbol{\alpha}_s(\boldsymbol{\theta})\, d\mathbf{m}_s(\boldsymbol{\theta})$, where the integral is over $(0, T)$, and where $\{\boldsymbol{\alpha}_s(\boldsymbol{\theta})\}$ is a predictable process. Here the QS estimating function is

$$\int_0^T (d\bar{\mathbf{m}}_s(\boldsymbol{\theta}))'(d\langle \mathbf{m}(\boldsymbol{\theta})\rangle_s)^+ \, d\mathbf{m}_s(\boldsymbol{\theta}),$$

where $d\bar{\mathbf{m}}_t(\boldsymbol{\theta}) = E(d\mathbf{m}_t(\boldsymbol{\theta}) \mid \mathcal{B}_{t-})$, and it is easily seen that this reduces to (14) in the particular case of the model (12).

Of course the martingale $\{\mathbf{m}_t(\boldsymbol{\theta}), \mathcal{B}_t\}$ can always be chosen in a variety of ways, for example with robustness requirements in mind as in the M-estimation context. A family of competing quasi-score estimating functions can then be envisaged based on the different possible choices of the basic martingale. The comparison of these and their combination, when this is advantageous, is treated by Heyde (1987). For example, referring to the (Bienaymé–) Galton–Watson branching process $\{Z_0 = 1, Z_1, \ldots, Z_T\}$ mentioned in § 5, the quasi-score estimating functions

$$G^*_{1,T} = \sum_{t=1}^{T} (Z_t - \theta Z_{t-1}), \quad G^*_{2,T} = \sum_{t=1}^{T} \{(Z_t - \theta Z_{t-1})^2 - \sigma^2 Z_{t-1})\}/(C + 2Z_{t-1}),$$

$C = \sigma^{-4}E(Z_1 - \theta)^4 - 3$ being the kurtosis of the distribution of Z_1, are *jointly* optimal for estimating the mean $EZ_1 = \theta$ and the variance $E(Z_1 - \theta)^2 = \sigma^2$ provided that $E(Z_1 - \theta)^3 = 0$. This result is also given by Godambe (1986). On the other hand, without this third moment condition, the *separate* optimality of $G^*_{1,T}$ and $G^*_{2,T}$, for estimating θ and σ^2 respectively, holds (Godambe, 1985). Godambe (1987) has initiated a theory of estimation which is aimed at spatial processes. This is based on a generalization of the martingale structure and the standard filtration of § 2.

Acknowledgment

Particular thanks are due to a referee for a very detailed set of comments.

Appendix: Equivalence of optimality criteria from § 3

The proof that Criteria 1, 2 and 3 are equivalent follows the essential lines of that of Godambe & Thompson (1985) for the scalar case. We shall drop the subscript T for convenience.

Let $\mathbf{G}_f = (E\dot{\mathbf{G}})^{-1}\mathbf{G}$. Then, note that

$$E\mathbf{G}_f\mathbf{U}' = (E\dot{\mathbf{G}})^{-1}E\mathbf{G}\mathbf{U}' = -\mathbf{I}_p, \quad E\mathbf{U}\mathbf{G}_f' = E\mathbf{U}\mathbf{G}'((E\dot{\mathbf{G}})^{-1})' = -\mathbf{I}_p.$$

We shall first show that Criteria 2 and 3 are equivalent and then that Criteria 1 and 2 are equivalent.

Suppose that Criterion 2 holds. Then

$$\begin{aligned} E(\boldsymbol{\alpha}\mathbf{U} - \mathbf{G})(\boldsymbol{\alpha}\mathbf{U} - \mathbf{G})' - E(\boldsymbol{\alpha}\mathbf{U} - \mathbf{G}^*)(\boldsymbol{\alpha}\mathbf{U} - \mathbf{G}^*)' \\ &= -\boldsymbol{\alpha}E(\mathbf{U}\mathbf{G}') - E(\mathbf{G}\mathbf{U}')\boldsymbol{\alpha}' + E(\mathbf{G}\mathbf{G}') \\ &\quad + \boldsymbol{\alpha}E(\mathbf{U}\mathbf{G}^{*\prime}) + E(\mathbf{G}^*\mathbf{U}')\boldsymbol{\alpha}' - E(\mathbf{G}^*\mathbf{G}^{*\prime}) \\ &= -E(\mathbf{G}^*\mathbf{G}') - E(\mathbf{G}\mathbf{G}^{*\prime}) + E(\mathbf{G}\mathbf{G}') + E(\mathbf{G}^*\mathbf{G}^{*\prime}) \\ &= E(\mathbf{G} - \mathbf{G}^*)(\mathbf{G} - \mathbf{G}^*)', \end{aligned}$$

which is nonnegative-definite, being a dispersion matrix. This gives Criterion 3.

Now suppose that Criterion 3 holds. Then for all scalar β and $\mathbf{G} \in \mathcal{G}_1$,

$$E(\boldsymbol{\alpha}\mathbf{U} - \mathbf{G}^* - \beta\mathbf{G})(\boldsymbol{\alpha}\mathbf{U} - \mathbf{G} - \beta\mathbf{G})' - E(\boldsymbol{\alpha}\mathbf{U} - \mathbf{G}^*)(\boldsymbol{\alpha}\mathbf{U} - \mathbf{G}^*)'$$

is nonnegative-definite which gives, after some algebra, that

$$\beta^2 E\mathbf{G}\mathbf{G}' - \beta E((\boldsymbol{\alpha}\mathbf{U} - \mathbf{G}^*)\mathbf{G}' + \mathbf{G}(\boldsymbol{\alpha}\mathbf{U} - \mathbf{G}^*)')$$

is nonnegative-definite. This is of the form $\beta^2\mathbf{A} - \beta\mathbf{B}$, where $\mathbf{A}$ and $\mathbf{B}$ are symmetric and $\mathbf{A}$ is positive-definite. Then, using the results of Rao (1965, 1c.3(ii), p. 37), there exists a nonsingular matrix $\mathbf{R}$ such that

$$\mathbf{A} = (\mathbf{R}^{-1})'\mathbf{R}^{-1}, \quad \mathbf{B} = (\mathbf{R}^{-1})'\boldsymbol{\Lambda}\mathbf{R}^{-1},$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix. Thus $(\mathbf{R}^{-1})'[\beta^2\mathbf{I}_p - \beta\boldsymbol{\Lambda}]\mathbf{R}^{-1}$ is nonnegative-definite, and hence $\mathbf{I}_p - \beta^{-1}\boldsymbol{\Lambda}$ is nonnegative-definite for all β which forces $\boldsymbol{\Lambda} = \mathbf{0}$ and hence $\mathbf{B} = \mathbf{0}$. But, this holds for each $\mathbf{G} \in \mathcal{G}_1$ so it is possible to replace $\mathbf{G}$ by $\mathbf{D}\mathbf{G}$ where $\mathbf{D} = \text{diag}\,(\lambda_1, \ldots, \lambda_p)$ is an arbitrary constant matrix. Then

$$E\{(\boldsymbol{\alpha}\mathbf{U} - \mathbf{G}^*)\mathbf{G}'\mathbf{D} + \mathbf{D}\mathbf{G}(\boldsymbol{\alpha}\mathbf{U} - \mathbf{G}^*)'\} = \mathbf{0},$$

which yields, in obvious notation,

$$\lambda_j E((\boldsymbol{\alpha}\mathbf{U} - \mathbf{G}^*)_i\mathbf{G}_j) + \lambda_i E((\boldsymbol{\alpha}\mathbf{U} - \mathbf{G}^*)_j\mathbf{G}_i) = 0 \quad (i, j = 1, 2, \ldots, p).$$

Then, first taking $\lambda_i = \lambda_j = 1$, we find

$$E((\boldsymbol{\alpha}\mathbf{U} - \mathbf{G}^*)_i\mathbf{G}_j) = -E((\boldsymbol{\alpha}\mathbf{U} - \mathbf{G}^*)_j\mathbf{G}_i)$$

and hence $(\lambda_i - \lambda_j)E((\boldsymbol{\alpha}\mathbf{U} - \mathbf{G}^*)_j\mathbf{G}_i) = 0$, which forces, for all i, j,

$$E((\boldsymbol{\alpha}\mathbf{U} - \mathbf{G}^*)_j\mathbf{G}_i) = 0,$$

and hence Criterion 2 holds. Thus Criteria 2 and 3 are equivalent.

Next suppose that Criterion 2 holds and note that this gives

$$E(\mathbf{G}_f^*(\mathbf{G}_f - \mathbf{G}_f^*)') = \mathbf{0} = E((\mathbf{G}_f - \mathbf{G}_f^*)\mathbf{G}_f^{*\prime}). \tag{A1}$$

Then,

$$\begin{aligned}(E\dot{\mathbf{G}})^{-1}E\mathbf{G}\mathbf{G}'((E\dot{\mathbf{G}})^{-1})' - (E\dot{\mathbf{G}}^*)^{-1}E\mathbf{G}^*\mathbf{G}^{*\prime}((E\dot{\mathbf{G}}^*)^{-1})' \\ &= E\mathbf{G}_f\mathbf{G}_f' - E\mathbf{G}_f^*\mathbf{G}_f^{*\prime} \\ &= E(\mathbf{G}_f - \mathbf{G}_f^*)(\mathbf{G}_f - \mathbf{G}_f^*)' + E(\mathbf{G}_f^*(\mathbf{G}_f - \mathbf{G}_f^*)') + E((\mathbf{G}_f - \mathbf{G}_f^*)\mathbf{G}_f^{*\prime}) \\ &= E(\mathbf{G}_f - \mathbf{G}_f^*)(\mathbf{G}_f - \mathbf{G}_f^*)'\end{aligned}$$

in view of (A1), and this dispersion matrix is nonnegative-definite giving Criterion 1.

Finally, suppose that Assumption 1 holds. Let $\tilde{\mathbf{G}} = \mathbf{G}_f^* + \mathbf{B}\mathbf{G}_f$, where $\mathbf{B}$ is a constant diagonal matrix. Then,

$$Ed\tilde{G}/d\theta = E\dot{\mathbf{G}}_f^* + \mathbf{B}E\dot{\mathbf{G}}_f = \mathbf{I}_p + \mathbf{B},$$

and Criterion 1 implies that

$$(\mathbf{I}_p + \mathbf{B})^{-1}E\tilde{\mathbf{G}}\tilde{\mathbf{G}}'(\mathbf{I}_p + \mathbf{B})^{-1} - E\mathbf{G}_f^*\mathbf{G}_f^{*\prime}$$

is nonnegative-definite. This gives, after some algebra, that

$$\mathbf{B}(E\mathbf{G}_f\mathbf{G}_f' - E\mathbf{G}_f^*\mathbf{G}_f^{*\prime})\mathbf{B} - \mathbf{B}(E(\mathbf{G}_f^* - \mathbf{G}_f)\mathbf{G}_f^{*\prime} - E\mathbf{G}_f^*(\mathbf{G}_f^* - \mathbf{G}_f)')\mathbf{B}$$

is nonnegative-definite for all $\mathbf{B}$.

Now put $\mathbf{B} = \beta\mathbf{B}_1$ where β is a scalar. A similar argument to that used to show that

Criterion 3 implies Criterion 2 first gives

$$\mathbf{B}(E(\mathbf{G}_f^* - \mathbf{G}_f)\mathbf{G}_f^{*\prime} + E\mathbf{G}_f^*(\mathbf{G}_f^* - \mathbf{G}_f)')\mathbf{B} = \mathbf{0},$$

and then

$$E((\mathbf{G}_f^* - \mathbf{G}_f)\mathbf{G}_f^{*\prime}) = E(\mathbf{G}_f^*(\mathbf{G}_f^* - \mathbf{G}_f)') = \mathbf{0}. \tag{A2}$$

But, if $\boldsymbol{\alpha} = -(E\dot{\mathbf{G}}^*)E\mathbf{G}_f^*\mathbf{G}_f^{*\prime}$, then

$$E((\mathbf{G}_f^* - \mathbf{G}_f)\mathbf{G}_f^{*\prime}) = E\mathbf{G}_f\mathbf{U}'\boldsymbol{\alpha}'((E\dot{\mathbf{G}}^*)^{-1})' - E\mathbf{G}_f\mathbf{G}_f^{*\prime} = E(\mathbf{G}_f(\boldsymbol{\alpha}\mathbf{U} - \mathbf{G}^*)')((E\dot{\mathbf{G}}^*)^{-1})'$$

and hence (A2) yields Criterion 2. Thus Criteria 1 and 2 are equivalent and the proof is complete.

References

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.* **6,** 701–726.

Andersen, P.K., Borgan, O., Gill, R. & Keiding, N. (1982). Linear nonparametric tests for comparison of counting processes, with applications to censored survival data. *Int. Statist. Rev.* **50,** 219–258.

Basawa, I.V. (1985). Neyman–Le Cam tests based on estimating functions. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer,* **2,** Ed. L. Le Cam and R.A. Olshen, pp. 811–825. Belmont, CA.: Wadsworth.

Bertrand, J. (1889). *Calcul des Probabilités.* Paris: Gauthier–Villars; 2nd ed. 1907. Reprinted (1973), New York: Chelsea.

Bhapkar, V.P. (1972). On a measure of efficiency in an estimating equation. *Sankhyā* A **34,** 467–472.

Bradley, E.L. (1973). The equivalence of maximum likelihood and weighted least squares estimates in the exponential family. *J. Am. Statist. Assoc.* **68,** 199–200.

Chandrasekar, B. & Kale, B.K. (1984). Unbiased statistical estimation functions in presence of nuisance parameters. *J. Statist. Plan. Inf.* **9,** 45–54.

Durbin, J. (1960). Estimation of parameters in time-series regression models. *J. R. Statist. Soc.* B **22,** 139–153.

Feigin, P.D. (1977). A note on maximum likelihood estimation for simple branching processes. *Aust. J. Statist.* **19,** 152–154.

Feigin, P.D. (1985). Stable convergence of semimartingales. *Stoch. Processes Applic.* **19,** 125–134.

Ferreira, P.E. (1982). Multiparametric estimating equations. *Ann. Inst. Statist. Math.* A **34,** 423–431.

Gauss, C.F. (1880). Theoria Combinationis Observationum Erroribus Minimus Obnaxiae. Part 1, 1821; Part 2, 1823; Suppl., 1826. In *Werke* **4,** 1–108. Göttingen.

Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* **31,** 1208–1212.

Godambe, V.P. (1985). The foundations of finite sample estimation in stochastic processes. *Biometrika* **72,** 419–428.

Godambe, V.P. (1987). The foundations of finite sample estimation in stochastic processes—II. In *Proc. First World Congress of Bernoulli Soc., Tashkent, 1986.* To appear.

Godambe, V.P. & Thompson, M.E. (1985). Logic of least squares revisited. Preprint.

Godambe, V.P. & Thompson, M.E. (1986). Parameters of superpopulation and survey population: their relationship and estimation. *Int. Statist. Rev.* **54,** 127–138.

Hall, P.G. & Heyde, C.C. (1980). *Martingale Limit Theory and its Application.* New York: Academic Press.

Heyde, C.C. (1975). Remarks on efficiency in estimation for branching processes. *Biometrika* **62,** 49–55.

Heyde, C.C. (1987). On combining quasi-likelihood estimating functions. *Stoch. Processes Applic.* **25.** To appear.

Heyde, C.C. & Seneta, E. (1977). *I.J. Bienaymé: Statistical Theory Anticipated.* New York: Springer.

Hotelling, H. (1936). Relations between two sets of variables. *Biometrika* **28,** 321–377.

Hutton, J.E. & Nelson, P.I. (1984). A mixing and stable central limit theorem for continuous time martingales. Preprint.

Hutton, J.E. & Nelson, P.I. (1986). Quasi-likelihood estimation for semimartingales. *Stoch. Processes Applic.* **22,** 245–257.

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

McCullagh, P. & Nelder, J.A. (1983). *Generalized Linear Models.* London: Chapman and Hall.

Morton, R. (1981). Efficiency of estimating equations and the use of pivots. *Biometrika* **68,** 227–233.

Rao, C.R. (1965). *Linear Statistical Inference and its Applications.* New York: Wiley.

Thavaneswaran, A. & Thompson, M.E. (1986). Optimal estimation for semimartingales. *J. Appl. Prob.* **23,** 409–417.

Shiryaev, A.N. (1981). Martingales: recent developments, results and applications. *Int. Statist. Rev.* **49,** 199–233.

Sprott, D.A. (1983). Gauss, Carl Frederich. In *Encyclopedia of Statistical Sciences,* **3,** Ed. S. Kotz and N.L. Johnson, pp. 305–308. New York: Wiley.

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61,** 439–447.
Wilks, S.S. (1938). Shortest average confidence intervals from large samples. *Ann. Math. Statist.* **9,** 166–175.

Résumé

Cet article étudie l'estimation des paramètres optimaux pour les processus stochastiques. Les équations d'estimation sont utilisées. Une définition très générale est présentée pour l'estimateur de quasi-vraisemblance. Ce travail généralise les résultats de Godambe (1985) et de Hutton & Nelson (1986).

[*Received November* 1986, *revised May* 1987]