Ann. Inst. Statist. Math.
Vol. 51, No. 4, 603–619 (1999)

# NONCONSERVATIVE ESTIMATING FUNCTIONS AND APPROXIMATE QUASI-LIKELIHOODS

JINFANG WANG

*The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan*

(Received August 19, 1997; revised September 3, 1998)

**Abstract.** The estimating function approach unifies two dominant methodologies in statistical inferences: Gauss's least square and Fisher's maximum likelihood. However, a parallel likelihood inference is lacking because estimating functions are in general not integrable, or nonconservative. In this paper, nonconservative estimating functions are studied from vector analysis perspective. We derive a generalized version of the Helmholtz decomposition theorem for estimating functions of any dimension. Based on this theorem we propose locally quadratic potentials as approximate quasi-likelihoods. Quasi-likelihood ratio tests are studied. The ideas are illustrated by two examples: (a) logistic regression with measurement error model and (b) probability estimation conditional on marginal frequencies.

*Key words and phrases*: Divergence-free vector fields, generalized Helmholtz decomposition, gradient vector fields, logistic regression with measurement error, potentials, quasi-likelihood ratio test, quasi-scores.

## 1. Introduction

An estimating function

$$u(\theta;Y)=\sum_{i=1}^{p}u_i(\theta;Y)d\theta_i \tag{1.1}$$

is a differential 1-form defined on $\Theta$, where $\Theta$ is a parameter space. The estimating function $u(\theta;Y)$ defines a random vector field on $\Theta$. This view of estimating functions emphasizes the important fact that the components $u_i(\theta;Y)$ of $u(\theta;Y)$ are ordered in the sense of (1.1), thus play an asymmetrical role in the theory of estimating functions. It is this asymmetry of the components that we set out to study in this paper. Note that it is essential for us to assume that $u(\theta;Y)$ and $\theta$ have the same dimension. We assume throughout that $\dim u(\theta;Y)=\dim\theta=p\geq 2$, since asymmetrical problem does not arise in the case $p=1$.

Construction of $u(\theta; Y)$ is usually based on assumptions about lower order moments of the underlying distribution. A typical example is the quasi-score (Wedderburn (1974); McCullagh (1983))

$$u(\theta; Y) = D'(\theta)V^{-1}(\theta)(Y - \mu(\theta)), \tag{1.2}$$

where $\mu(\theta)$ and $V(\theta)$ are mean and covariance matrix of random vector $Y = (Y_1, \ldots, Y_n)$ and $D(\theta) = (\partial/\partial\theta)\mu(\theta)$. McCullagh and Nelder ((1989), Section 9.4) call $Y - \mu(\theta)$ elementary estimating functions and argue that the weight $D'(\theta)V^{-1}(\theta)$ is optimal in combining these elementary estimating functions. We shall tacitly assume that our estimating function $u(\theta; Y)$ of (1.1) is optimized in the sense of McCullagh and Nelder ((1989), Section 9.5).

The estimating function approach (Godambe (1960)) unifies two dominant methodologies in statistical inferences: Gauss's least square and Fisher's maximum likelihood. Substantial theories have been developed in this area (e.g. Godambe (1991)); see also a new and interesting theory fundamental to both probability and statistics recently proposed by Small and McLeish (1994).

In the maximum likelihood theory, $u(\theta; Y)$ corresponds to score functions. Scores are symmetrical in the sense that the observed Fisher information matrix is symmetrical, due to the fact that scores are the gradient of a log-likelihood. Scores thus form a gradient (or potential) field, where the log-likelihood plays the role of a potential. By definition, an estimating function needs not be a gradient field. That is, there may exist no scalar function $\phi(\theta)$ such that

$$u(\theta) = d\phi(\theta) \tag{1.3}$$

where $d$ denotes the exterior differentiation operator. We dropped the dependence of $u(\theta; Y)$ on $Y$ in (1.3) and will do so when we wish to stress that $u(\theta; Y)$ is a vector field of $\theta$. When (1.3) holds for no $\phi(\theta)$, we say that estimating function $u(\theta)$ is nonconservative or non-integrable. Two examples will be studied in Section 4. Note that an equivalent condition for $u(\theta)$ being conservative is that $du(\theta) \equiv 0$. This condition says that, in matrix terminology, the Hessian of $u(\theta)$ is symmetrical. We note in passing that while the expected Hessian of quasi-score (1.2), $D'(\theta)V^{-1}(\theta)D(\theta)$, is symmetrical, the observed Hessian needs not be so. Thus quasi-scores are in general nonconservative. The voter transition probability problem studied in Section 4 provides such an example. Note that a nonconservative estimating function $u(\theta)$ may be transformed into a conservative one $v(\theta) = T(\theta)u(\theta)$ by a nonsingular matrix $T(\theta)$. While the estimating functions $u(\theta)$ and $v(\theta) = T(\theta)u(\theta)$ do provide the same estimators, they might play quite different roles as more general inference functions. This can be appreciated by considering a vector field obtained by permuting a gradient field such as the score function.

Inferences based on estimating functions for which there exists no potential function deviate from theories based on likelihood. A Baysian viewpoint is difficult in the absence of a potential. There are also practical difficulties associated with nonconservativeness: constructing goodness-of-fit statistics; distinguishing consistent root from among multiple roots; constructing confidence intervals when multiple roots exist, etc.

Nonconservativeness of estimating functions has been studied by a number of authors. McCullagh and Nelder ((1989), pp. 334–336) study conditions on $V(\cdot)$ of (1.2), under which quasi-score is integrable; Li and McCullagh (1994) study conditions under which a linear and unbiased estimating function is integrable. Their ideas are based on restricting either the choice of variance functions or the class of estimating functions. Unlike these authors we begin with nonconservative estimating functions and proceed to develop a theory for choosing approximate potentials, or quasi-likelihoods. We shall derive a generalized version of the Helmholtz decomposition theorem for estimating functions of any dimension $p \geq 2$. This is studied in Section 2. The generalized Helmholtz decomposition theorem says that any vector field can be decomposed into the sum of a gradient vector field and a divergence-free vector field. We note that relevance of the Helmholtz decomposition theorem has been pointed out by McCullagh ((1991), pp. 284–285).

Two difficulties arise here. First, the decomposition is not unique. There exists a class of potential functions for a given estimating function. Second, it is usually impossible to express potentials in closed forms even for very simple estimating functions. These issues are settled in Section 3 by linearizing estimating function $u(\theta)$ properly. We therefore propose a locally quadratic potential as a (log) quasi-likelihood (cf. (3.6)), based on which we study quasi-likelihood ratio test and so forth. Two examples are studied in Section 4. The first concerns the so-called Neyman-Scott paradox (Neyman and Scott (1948)), the second on probability estimation based on marginal frequencies.

For related works on approximate quasi-likelihoods, see also McLeish and Small (1992), Li (1993, 1996), Barndorff-Nielsen (1995) and Hanfelt and Liang (1995, 1997), etc.

## 2. Nonconservative estimating functions and Helmholtz-type potentials

### 2.1 *Nonconservative estimating functions*

Let $Y = (Y_1, \ldots, Y_n)$ be a random vector with possibly dependent components. For a parameter of interest $\theta \in \Theta \subset \Re^p$, consider estimating function $u(\theta; Y)$ of form (1.1), which is assumed optimal in the sense of McCullagh and Nelder ((1989), Section 9.5). Parameter $\theta$ often is a structure parameter relating the mean of $Y$ with the linear predictor via a link function, such as in the generalized linear model (Nelder and Wedderburn (1972); McCullagh and Nelder (1989)). With $\dim(\theta) = p = 1$, we can always integrate $u(\theta)$ back to get a potential function, which, when normalized, serves as a quasi-likelihood.

By definition, an estimating function $u(\theta)$ is conservative, or integrable, if there exists a scalar function $\phi(\theta)$ such that $u(\theta) = d\phi(\theta)$; or equivalently, by Poincaré's Lemma, $du(\theta) \equiv 0$. Otherwise $u(\theta)$ is said nonconservative, or non-integrable. Score functions are conservative estimating functions, while quasi-scores are generally not. Note that a theory for conservative quasi-scores can essentially be reduced to a theory for exponential families.

Definition of conservativeness applied to linear estimating function, $u(\theta) = A\theta + b$, say, is simply the requirement that $A$ is symmetrical; where $A$ and $b$ are constant matrix and column vector not depending on $\theta$. We shall express estimating functions in usual matrix notations and in differential 1-forms interchangeably,

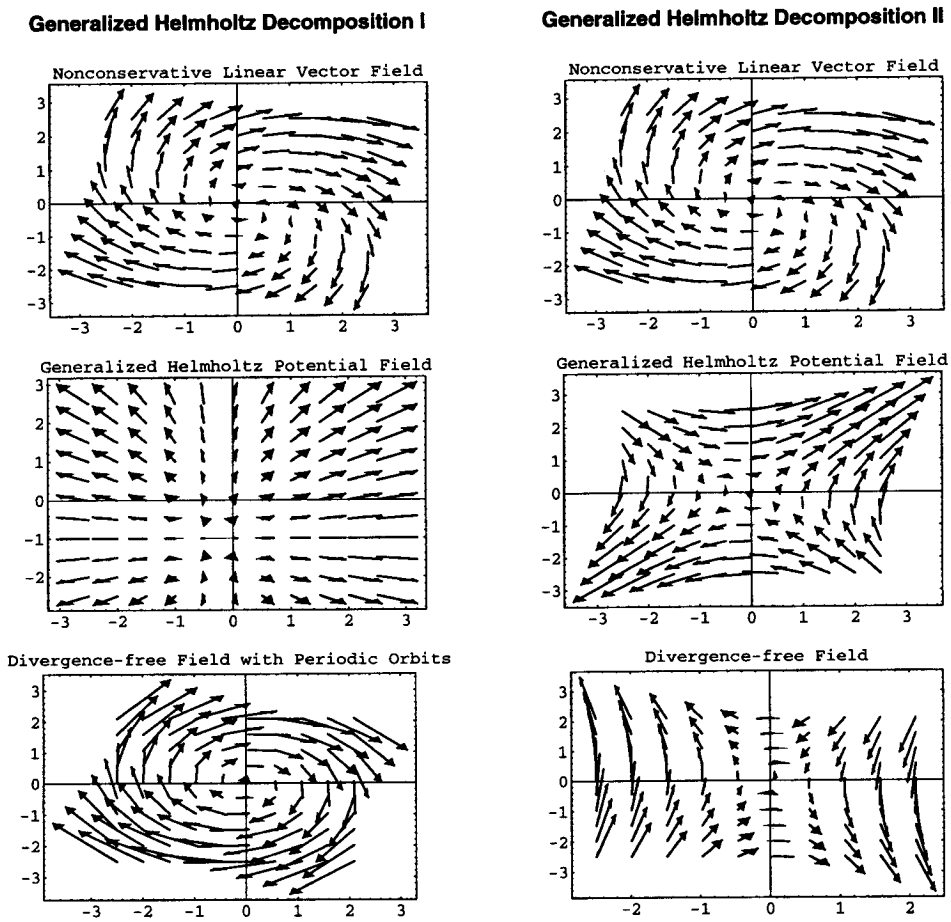


Fig. 1. Helmholtz-type decompositions for a two-dimensional linear vector field $u(\theta) = A\theta + b$; where $A = (a_{ij})$, $b = (1.5, 3)'$ and $a_{11} = 5$, $a_{22} = 3$, $a_{12} = -a_{21} = 10$. The left panel corresponds to quadratic potential $\frac{1}{2}\theta' A\theta + b\theta$; the right panel corresponds to potential $\theta' B\theta + b\theta$, where $B = (b_{ij})$, $b_{ii} = a_{ii}/2$ and $b_{ij} = a_{ij}$ for $i \neq j$; cf. Section 3 for details.

whenever we feel convenient and where no confusion is anticipated. Top of Fig. 1 displays an artificial linear nonconservative vector field, for $p = 2$, $A = (a_{ij})$ and $b = (1.5, 3)'$; where $a_{11} = 5$, $a_{22} = 3$, $a_{12} = -a_{21} = 10$. More complicated nonconservative vector fields are shown in top of Fig. 2, which are studied in Section 4. For a general estimating function $u(\theta)$, conservativeness is equivalent to symmetry of the Hessian, which is convenient for checking.

### 2.2 *Preliminaries on vector analysis*

Let $\Theta \subset \Re^p$ be a parameter space of dimension $p$. We may alternatively regard parameter $\theta = (\theta_1, \ldots, \theta_p)$ as a local coordinate in $\Theta$. A differential $k$-form $(k = 0, 1, \ldots, p)$ on $\Theta$ is a formal sum of terms $f(\theta) d\theta_{i_1} \wedge \cdots \wedge d\theta_{i_k}$, where $f(\theta)$ is a scalar function on $\Theta$ and $\{i_1, \ldots, i_k\} \subset \{1, \ldots, p\}$. Differential 0- and $p$-forms

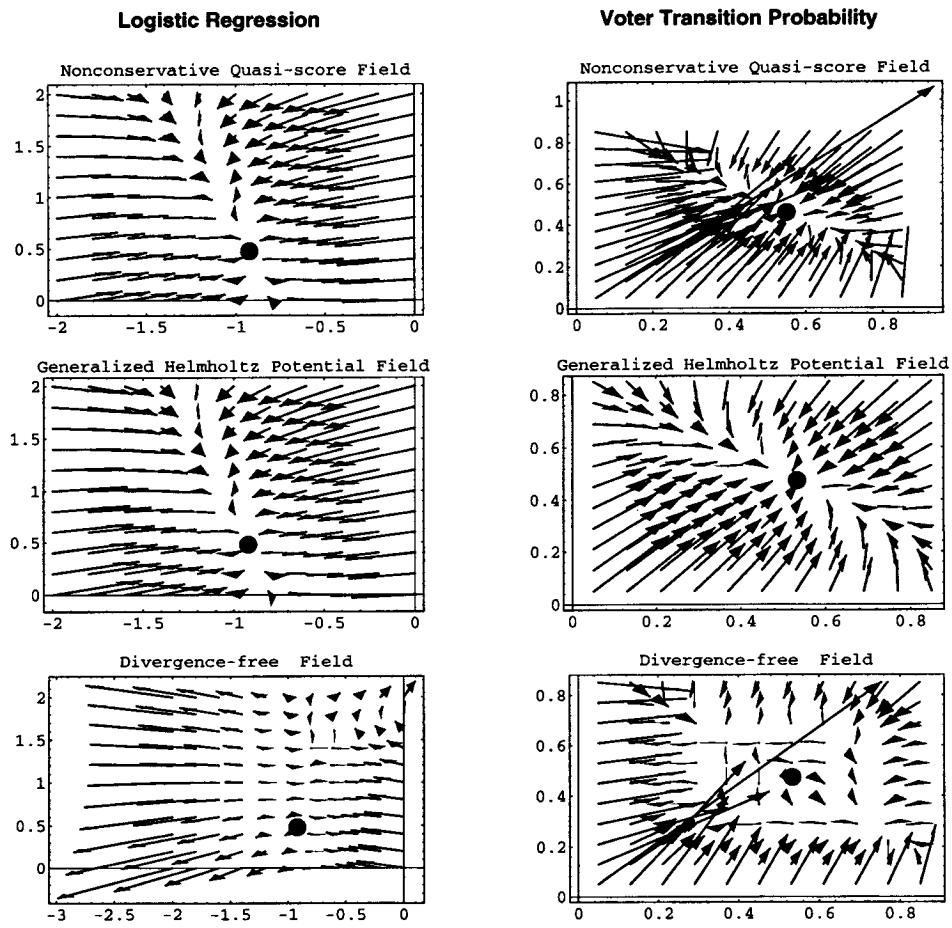


Fig. 2. Helmholtz-type decompositions for non-conservative estimating functions (4.1) and (4.2) studied in Section 4. The left panel corresponds to logistic regression with measurement error; the right panel corresponds to voter transition probability estimation problem. In both cases, the gradient fields correspond to quadratic potentials (3.6); these vector fields are plotted for the same fixed observations in each case; cf. Section 4 for details.

correspond to scalars; differential 1-forms correspond to vectors. An elegant and formal treatment of differential forms would involve alternating tensor fields or the Grassmann algebra.

The collection of all differential $k$-forms, $\Omega^k$ (say), is a linear space of dimension $\binom{p}{k}$. The set $\{d\theta_{i_1} \wedge \cdots \wedge d\theta_{i_k}; i_1 < \cdots < i_k\}$ forms a basis of $\Omega^k$ (it is in fact an orthonormal basis in the Riemannian sense (Kobayashi (1990), p. 121)). Note that $\Omega^{p-k}$ has the same dimension $\binom{p}{k}$. Hodge's star operator is an isomorphic linear mapping

$$* : \Omega^k \to \Omega^{p-k}$$

satisfying

$$\omega \wedge *\omega = d\theta_1 \wedge \cdots \wedge d\theta_p$$

for any element of the orthonormal basis $\omega = d\theta_{i_1} \wedge \cdots \wedge d\theta_{i_k}$ $(i_1 < \cdots < i_k)$. Or, equivalently, $*(d\theta_{i_1} \wedge \cdots \wedge d\theta_{i_k}) = \pi d\theta_{i_{k+1}} \wedge \cdots \wedge d\theta_{i_p}$, where $i_1 < \cdots < i_k; i_{k+1} < \cdots < i_p, \{i_1, \ldots, i_p\} = \{1, \ldots, p\}$ and $\pi$ is the sign of the permutation from $(1, \ldots, p)$ to $(i_1, \ldots, i_p)$.

For example, if $p = 2$, then $*(d\theta_1) = d\theta_2$, $*(d\theta_2) = -d\theta_1$; if $p = 3$, then $*(d\theta_1) = d\theta_2 \wedge d\theta_3$, $*(d\theta_2) = d\theta_3 \wedge d\theta_1$, $*(d\theta_3) = d\theta_1 \wedge d\theta_2$. For details on differential forms and dynamic systems see Abraham and Marsden (1978), Kobayashi (1990, Section 3.7), Siegel and Moser (1991) and Fukaya (1996).

We list some basic properties of the star operator, proofs of which are straightforward, and hence omitted.

PROPOSITION 2.1. *Denote by $*_{\Omega^k}$ the star operator from $\Omega^k$ to $\Omega^{p-k}$ $(k = 0, 1, \ldots, p)$. The following properties hold.*

(i) *Operator $*_{\Omega^{p-k}}$ is the inverse of $*_{\Omega^k}$, except for $k$ odd and $p$ even for which the inverse is $(-1)*_{\Omega^{p-k}}$. That is*

$$(*_{\Omega^{p-k}} \circ *_{\Omega^k})\omega = (-1)^{k(p-k)}\omega, \qquad \omega \in \Omega^k.$$

(ii) *Let $d_* \equiv *d*$. Then $d_* \circ d_* \equiv 0$.*

(iii) *Let $\Delta$ be the Laplacian operator, and $\phi(\theta) \in \Omega^0$. We have $d_* \circ d\phi(\theta) \equiv \Delta\phi(\theta)$.*

(iv) *Let $\omega(\theta) \in \Omega^1$. We have $d_*\omega = \mathrm{div}(\omega)$, where* div *is the divergence operator, i.e. $\mathrm{div}(\omega(\theta)) = \sum_{i=1}^{p}(\partial/\partial\theta_i)\omega_i(\theta)$ for $\omega(\theta) = \sum_{i=1}^{p}\omega_i(\theta)d\theta_i$.*

(v) [Poincaré's Lemma] *For any $\omega \in \Omega^k (k < p)$, there exists $\varsigma \in \Omega^{k+1}$ such that*

$$\omega = d_*\varsigma \quad \text{if and only if} \quad d_*\omega = 0.$$

2.3 *Generalized Helmholtz decomposition*

Gradient, or potential, vector fields and divergence-free vector fields are two kinds of most important vector fields in physics. Many natural fields turn out to be potential fields. The gravitational field and electric field of particles at rest are well-known examples (Fukaya (1995), p. 111). On the other hand, divergence-free vector fields can be used to model, for example, an incompressible flow of gas. Divergence can be visualized in terms of *source* and *sink* of a vector field (Irwin (1980)). Gauss's law is one of the facts which makes the concept of divergence important in physics (Fukaya (1995), p. 105). While bottoms of Figs. 1 and 2 display examples of divergence-free vector fields, plots at the center show four gradient fields.

Now we give a generalized version of the Helmholtz decomposition theorem.

THEOREM 2.1. *For any given $u(\theta) \in \Omega^1$, there exist $\phi(\theta) \in \Omega^0$ and $\mathcal{P}(\theta) \in \Omega^2$ such that*

$$u(\theta) = d\phi(\theta) + d_*\mathcal{P}(\theta). \tag{2.1}$$

*Or equivalently, for any* $u(\theta) \in \Omega^1$, *there exist* $\phi(\theta) \in \Omega^0$ *and* $\mathcal{Q}(\theta) \in \Omega^{p-2}$ *such that*

$$u(\theta) = d\phi(\theta) + *d\mathcal{Q}(\theta). \tag{2.2}$$

PROOF. Equivalence of (2.1) and (2.2) follows immediately from definition of star operator, thus we shall only prove (2.1). Since $u(\theta) \in \Omega^1$, $d_*u(\theta)$ is a scalar function. By the fact that Poission equation $\Delta\phi(\theta) = d_*u(\theta)$ admits a solution $\phi(\theta)$ for any given $u(\theta)$, we conclude that $d_*(u(\theta) - d\phi(\theta)) = 0$. The last equation, by Poincaré's lemma, implies the existence of $\mathcal{P}(\theta) \in \Omega^2$ such that $u(\theta) - d\phi(\theta) = d_*\mathcal{P}(\theta)$, or equivalently $u(\theta) = d\phi(\theta) + d_*\mathcal{P}(\theta)$, as was claimed.

For estimating equation $u(\theta)$, we shall call $\phi(\theta)$ a scalar potential, or simply a potential, of $u(\theta)$, and $\mathcal{P}(\theta)$ a vector potential. Note that divergence of $d_*\mathcal{P}(\theta)$ vanishes because $\operatorname{div}(d_*\mathcal{P}(\theta)) = d_* \circ d_*\mathcal{P}(\theta) = 0$, by (iii) and (iv) of Proposition 2.1. For $p = 3$, Theorem 2.1 says that any vector field can be decomposed as the sum of an irrotational and a solenoidal vector field, which is the well-known Helmholtz decomposition theorem. We state this as a corollary together with the case $p = 2$.

COROLLARY 2.1. *Let* $u(\theta)$ *be an estimating function.*

(i) *If* $\dim\theta = 2$, *then there exist scalar functions* $\phi(\theta)$ *and* $\psi(\theta)$ *satisfying*

$$u(\theta) = \left(\frac{\partial}{\partial\theta_1}, \frac{\partial}{\partial\theta_2}\right)'\phi(\theta) - \left(\frac{\partial}{\partial\theta_2}, -\frac{\partial}{\partial\theta_1}\right)'\psi(\theta). \tag{2.3}$$

(ii) *If* $\dim\theta = 3$, *then there exist a scalar function* $\phi(\theta)$ *and a vector* $\mathcal{Q}(\theta)$ *such that*

$$u(\theta) = grad\,\phi(\theta) + Curl\,\mathcal{Q}(\theta). \tag{2.4}$$

*Remark.* In case (i) when $p = 2$, the divergence-free vector field forms a Hamilton vector field, where $-\psi(\theta)$ plays the role of a Hamiltonian.

Helmholtz-type decomposition is not unique. There exist a class of potentials generated by the class of harmonic functions, exact meaning of which is summarized in the following theorem. We omit the proof.

THEOREM 2.2. *Let* $u(\theta)$ *be an estimating function.*

(i) *If* $u(\theta) = d\phi(\theta) + d_*\mathcal{P}(\theta)$, *then for any harmonic function* $h(\theta)$, *we also have* $u(\theta) = d\phi'(\theta) + d_*\mathcal{P}'(\theta)$, *where* $\phi'(\theta) = \phi(\theta) + h(\theta)$ *and* $d_*\mathcal{P}'(\theta) = d_*\mathcal{P}(\theta) - dh(\theta)$.

(ii) *Conversely, if*

$$u(\theta) = d\phi(\theta) + d_*\mathcal{P}(\theta) \quad \text{and} \quad u(\theta) = d\phi'(\theta) + d_*\mathcal{P}'(\theta),$$

*then* (a) $h(\theta) = \phi'(\theta) - \phi(\theta)$ *is harmonic and* (b) $d_*\mathcal{P}'(\theta) = d_*\mathcal{P}(\theta) - dh(\theta)$.

Thus, if $u(\theta) = d\phi(\theta) + d_*\mathcal{P}(\theta)$, then any member in the class $\{\phi(\theta) + h(\theta) \mid h(\theta)$ : harmonic function$\}$, serves as a potential for $u(\theta)$. A theory of quasi-likelihoods therefore intrinsically depends first on (i) statistical informativeness of estimating function $u(\theta)$ and then on (ii) the choice of a statistically meaningful harmonic function $h(\theta)$. The first requirement, viz. statistical informativeness, is guarded by the theory on optimal estimating functions. Next, choosing a statistically meaningful harmonic function is, equivalently, the requirement that the divergence-free vector field should contain as less as possible statistical information contained in the original estimating function. This is because scalar and vector potentials are paired according to (2.1) or (2.2). In Section 3 we shall focus our attention on the choice of a vector potential for a particular kind of vector fields. There arises a third difficulty. It is usually impossible to obtain in a closed form a potential function for even very simple estimating functions. These considerations coerce us into compensation for an approximate theory.

## 3. Approximate quasi-likelihoods

### 3.1 *Quadratic potential functions*

We have seen that choosing a proper scalar potential is the same as choosing a proper vector potential. Now we study vector potentials for linear vector fields $u(\theta) = A\theta + b$, where $A$ and $b$ are $p \times p$ matrix and $p \times 1$ vector not depending on $\theta$. For linear vector fields we are able to identify a particular kind of vector potentials.

THEOREM 3.1. *For any linear vector field* $u(\theta) = A\theta + b$, *there exists a scalar function* $\psi(\theta)$ *such that the generalized Helmholtz vector potential* $\mathcal{Q}(\theta) \in \Omega^{p-2}$ *can be written* $\mathcal{Q}(\theta) = \psi(\theta) * du(\theta)$, *namely*

$$u(\theta) = d\phi(\theta) + *d\{\psi(\theta) * du(\theta)\}. \tag{3.1}$$

PROOF. Decomposition (3.1) can be proved by taking

$$\psi(\theta) = (-1)^p \frac{1}{4}\theta'\theta. \tag{3.2}$$

We omit the details.

Decomposition (3.1) enjoys a good property that $u(\theta)$ is integrable if and only if the divergence-free part vanishes. This means that if $u(\theta)$ happens to be conservative, then scalar potential chosen according to (3.1) coincides with the original estimating function (up to a constant).

COROLLARY 3.1. *When* $p = 3$, *decomposition* (3.1) *can be rewritten, using notations of classical vector analysis, as*

$$u(\theta) = d\phi(\theta) + \frac{1}{2}(Curl\, u(\theta)) \times \theta, \tag{3.3}$$

*where* $\times$ *denotes the outer product of two vectors.*

Decomposition (3.3) has a simple physical interpretation. Let $u(\theta)$ represent a velocity field, and $\theta$ a location vector. Decomposition (3.3) says that any linear velocity field can be 'orthogonally' decomposed as the sum of a potential field and its maximum circulation. If, further, $u(\theta)$ happens to be generated by a constant angular velocity field $\omega$, i.e. $u(\theta) = \omega \times \theta$, then the gradient part of (3.3) vanishes.

For decomposition (3.1) with vector potential determined by (3.2), the scalar potential is uniquely (up to a constant) given by

$$\phi(\theta) = \frac{1}{2}\theta' A\theta + b'\theta. \tag{3.4}$$

This is the potential function we shall use for our purpose of statistical inferences. Note that since $A$ is asymmetrical, the first term reduces to the quadratic form $\theta'(A + A')\theta/4$. With quadratic potential (3.4), the divergence-free vector field has the form $\frac{1}{2}(A - A')\theta$. This divergence-free vector field defines a dynamic system with only periodic orbits; cf. Fig. 1 (left panel) for an artificial example. The right panel of Fig. 1 displays an alternative decomposition with potential $\theta' B\theta + b\theta$, where $A = (a_{ij})$, $B = (b_{ij})$, $b_{ii} = a_{ii}/2$ and $b_{ij} = a_{ij}$ for $i \neq j$. The potential and divergence-free vector fields are given by $(A + A_0')\theta + b$ and $-A_0'\theta$, respectively, where $A_0$ is $A$ with diagonal elements replaced by zeros.

3.2 *Approximate quasi-likelihoods*

To apply previous results to the theory of estimating functions, we first consider linearization of $u(\theta)$ in a neighborhood $\Theta_0$ of $\theta_0$. Let $\xi \in \Theta_0$, $A_\xi = (\partial/\partial\theta)u(\theta; Y)\,|_{\theta=\xi}$ and $b_\xi = u(\xi; Y)$. One term Taylor expansion of $u(\theta; Y)$ at $\theta = \xi$ yields

$$\tilde{u}(\theta, \xi; Y) = A_\xi(\theta - \xi) + b_\xi. \tag{3.5}$$

For this linear vector field we apply Theorem 3.1 with scalar potential (3.4) to obtain the potential

$$\phi(\theta, \xi; Y) = \frac{1}{2}(\theta - \xi)' A_\xi(\theta - \xi) + b_\xi'(\theta - \xi) \tag{3.6}$$

depending on $\xi$ at which $u(\theta)$ is linearized. In a specific problem $\xi$ has to be chosen by data or the specific statistical problem at hand. For instance, we may choose $\xi$ to be a consistent root to $u(\theta) = 0$; or choose $\xi = \theta_0$ for testing a null hypothesis and so on. Our simulation studies (see Section 4) show however asymptotic distributions of quasi-likelihood ratio statistics based on (3.6) are quite robust against the choice of $\xi$, as long as it belongs to a neighborhood of the null $\theta = \theta_0$.

We call potential $\phi(\theta; Y)$ of (3.6) an approximate (log) quasi-likelihood of estimating function $u(\theta; Y)$. Center of Fig. 2 shows quadratic gradient fields for the examples studied in Section 4.

So far our theory is general for arbitrary estimating functions. Now we apply (3.6) to the most important special case of quasi-score (1.2). First note that

$$b_\xi = D'(\xi)V^{-1}(\xi)(Y - \mu(\xi)),$$
$$A_\xi \approx -D'(\xi)V^{-1}(\xi)D(\xi),$$
$$\mu(\theta) - \mu(\xi) \approx D(\xi)(\theta - \xi).$$

So by (3.6) we have approximation

$$\begin{aligned}(3.7)\qquad \phi(\theta,\xi;Y) &= \frac{1}{2}(\theta-\xi)'A_\xi(\theta-\xi) + b_\xi'(\theta-\xi)\\ &= -\frac{1}{2}(\theta-\xi)'D_\xi'V^{-1}(\xi)D(\xi)(\theta-\xi)\\ &\quad + (Y-\mu(\xi))'V^{-1}(\xi)D(\xi)(\theta-\xi)\\ &= -\frac{1}{2}(\mu(\theta)-\mu(\xi))'V^{-1}(\xi)(\mu(\theta)-\mu(\xi))\\ &\quad + (\mu(\theta)-\mu(\xi))'V^{-1}(\xi)(Y-\mu(\xi)).\end{aligned}$$

Quasi-likelihood (3.7) is invariant under reparameterization and enjoys (local) likelihood properties. Denote $u_\xi(Y) = \partial\phi(\theta,\xi;Y)/\partial\theta \mid_{\theta=\xi}$ and $u_{\xi\xi'}(Y) = \partial^2\phi(\theta,\xi;Y)/\partial\theta\theta' \mid_{\theta=\xi}$. First we have (local) zero-unbiasedness since expectation of

$$u_\xi(Y) = D'(\xi)V^{-1}(\xi)(Y-\mu(\xi))$$

vanishes at $\theta = \xi$. Information-unbiasedness follows by noting that identities

$$E_\xi[u_\xi(Y)u_\xi'(Y)] = D'(\xi)V^{-1}(\xi)D(\xi)$$

and

$$E_\xi[u_{\xi\xi'}(Y)] = -D'(\xi)V^{-1}(\xi)D(\xi).$$

Note that the form of quasi-likelihood (3.7) closely resembles the deviance function of Li ((1993), equation (3)).

3.3 *Quasi-likelihood ratio tests*

Based on a general nonconservative estimating function $u(\theta;Y)$ we now consider the problem of hypothesis testing with null hypothesis $H_0 : \theta = \theta_0$. One obvious choice of $\xi$ on which our quadratic potential $\phi(\theta,\xi,Y)$ depends is $\theta_0$. Let $\hat\theta(\theta_0) = \operatorname{argmax}\phi(\theta,\theta_0;Y)$ be the maximizer of approximate (log) quasi-likelihood. Since $\phi(\theta_0,\theta_0;Y) \equiv 0$, definition of the usual likelihood ratio suggests the definition of quasi-likelihood ratio statistic

$$(3.8)\qquad \gamma = 2\phi(\hat\theta(\theta_0),\theta_0;Y).$$

Let $\tilde u(\theta,\theta_0;Y)$ be the linearized version defined by (3.5) in the previous section for a general estimating function $u(\theta;Y)$. Assume that $\Sigma = -E_{\theta_0}(\partial/\partial\theta)\tilde u(\theta,\theta_0;Y)$ and $\Gamma = E_{\theta_0}\tilde u'(\theta,\theta_0;Y)\tilde u(\theta,\theta_0;Y)$ exist and be positive definite. Note that since

quasi-scores are information-unbiased we have $\Sigma = \Gamma$. Under the null hypothesis $\hat{\theta}(\theta_0)$ will be consistent for $\theta_0$ and normally distributed with covariance matrix $\Sigma^{-1}\Gamma\Sigma^{-1}$. Standard arguments for quadratic forms of normal variables (e.g. Johnson and Kotz (1970), Chapter 29) therefore lead to the asymptotic result

$$\gamma \xrightarrow{\mathcal{L}} \sum_{i=1}^{p} \lambda_i Z_i^2 \tag{3.9}$$

where the $Z$'s are independent unit normal variables and $\lambda$'s eigenvalues of $\Sigma^{-1}\Gamma$. The above arguments are similar to that in Li and McCullagh (1994). Figure 4 compares asymptotic distributions (3.9) with simulated true distributions of $\gamma$ for the two examples studied in Section 4.

Since $\Sigma = \Gamma$ for quasi-score, $\Sigma^{-1}\Gamma$ is identity and the $\lambda$'s are all unity. So quasi-likelihood ratio statistics asymptotically follow $\chi^2_{(p)}$ under null hypothesis. If $\tilde{u}(\theta; Y)$ is approximately information-unbiased, the asymptotic distribution of quasi-likelihood ratio will be approximately $\chi^2$ with $p$ d.f. The asymptotic distributions shown in Fig. 4 are only slightly different from $\chi^2_{(p)}$; cf. Section 4 for details.

## 4. Examples

### 4.1 *Logistic regression with measurement error*

Estimating function approach is an effective way for eliminating nuisance parameters. In measurement error model the number of nuisance parameters increases with the number of observations, a situation known as the Neyman-Scott paradox, where the maximum likelihood estimators fail. Our first example concerns logistic regression with measurement error; we wish to fit the model $logit(\pi) = \alpha + \beta' x$, where $\pi$ is the mean of a binary response $Y$ and $x$ the covariate. Suppose that we can not observe $x$ but observe instead $Z = x + \epsilon$, where measurement error $\epsilon \sim N_{p-1}(0, \Psi)$ is independent of $Y$ and $\Psi$ a known covariance matrix. Conditioning on the complete sufficient statistic $A = z + y\Psi\beta$ for the nuisance parameter $x$, Stefanski and Carroll (1987) derived the conditional score $\sum(1, x_i)'(y_i - \mu_i^c)$, where $\mu_i^c = (1 + \exp\{-(\alpha + (A_i - \frac{1}{2}\Psi\beta)^T\beta)\})^{-1}$ is the conditional mean. Hanfelt and Liang (1995) suggest that $x$ be eliminated in the conditional score by forming estimating function

$$u(\theta) = \sum_{i=1}^{n} (1, d_i)'(y_i - \mu_i^c), \tag{4.1}$$

where $\theta = (\alpha, \beta)$ and $d_i = A_i + (\mu_i^c - 1)\Psi\beta$. It can be verified that $du(\theta) \neq 0$, thus $u(\theta)$ is not conservative. Stefanski and Carroll (1987) also reported the multiple roots problem, which is studied by Hanfelt and Liang (1995, 1997) by path-dependent integration approach. In general, when an estimating function has multiple roots, confidence intervals based on estimating functions tend to have separate regions (McCullagh (1991), p. 278), thus are not useful.

We shall consider the case $p = 2$ in following discussions. Figure 2, top left, depicts vector field (4.1) for a particular sample of size 10. The black dot shows one solution to the estimating equation $u(\theta) = 0$; the vector field *sinks* (cf. Section 2) into this point.

Now we consider testing the hypothesis $H_0 : \theta = (-1.4, 1.4)$, a value reported in a large cohort study by Stefanski and Carroll (1985). In this and next example we have done our simulation studies by linearizing estimating functions at null hypothesis and also at other points which slightly deviate from the null. Note that if linearization is not carried out at null, definition of quasi-likelihood ratio statistic should be adjusted accordingly as $\gamma = -2[\phi(\theta_0, \xi; Y) - \phi(\hat{\theta}(\xi), \xi, Y)]$, where $\hat{\theta}(\xi) = \operatorname{argmax} \phi(\theta, \xi; Y)$. The general asymptotic result (3.9) is still applicable where both $\Sigma$ and $\Gamma$ should also be adjusted properly. We first linearize $u(\theta)$ at $\xi = (-0.9, 0.9)$ and obtain its quadratic quasi-likelihood (3.6). Figure 2, center and bottom left, shows the gradient and the divergence-free vector fields corresponding to this quasi-likelihood for the same sample used to display the top figure. Figure 3, left panel, compares the true conditional log-likelihood surface with the quadratic quasi-likelihood for a sample of size 200.

To study quasi-likelihood ratio test, we set the variance of measurement error $\Psi = (0.1/3)^2$, a value used by Hanfelt and Liang (1995) in their simulation. In our simulation the unobservable measurements $x$'s are taken as pseudo-random numbers generated from $N(0, (0.1)^2)$. Bottom of Fig. 4 compares the asymptotic distribution of the quasi-likelihood ratio statistic, $\lambda_1 Z_1^2 + \lambda_2 Z_2^2$, $\lambda = (1.24, 0.84)$, with the true distribution for sample size 200 by 1000 simulations. We have tried various values of $\xi$ (including $\xi = \theta_0$) and similar satisfactory fits have been observed.

4.2 *Probability estimation conditional on marginal frequencies*

In the voter transition probability problem (Firth (1982)), as described by McCullagh and Nelder ((1989), pp. 336–339) and studied further by Li and McCullagh (1994), we are interested in estimating the transition probabilities $\theta_i$ of voting for Party $P_1$ by electorate previously voted for Party $P_i (i = 1, \ldots, p)$. We extended the original 2-party model to include $p$ parties. Conditional on number of voters $m_i$ for Party $P_i$ in a previous election, numbers of voters for Party $P_1$ in the next election $X_i$ are assumed independent binomial variables with index $m_i$ and transition probability $\theta_i (i = 1, \ldots, p)$; cf. Table 1. The difficulty of the problem lies in the fact that the $X$'s are hidden and we must estimate $\theta = (\theta_1, \ldots, \theta_p)$ based on the total $Y = \sum_{i=1}^{p} X_i$.

We assume that there are available records for $n$ previous elections. That is, we have data $\{y_i; m_{i1}, \ldots, m_{ip}\}_{i=1}^{n}$. To estimate $\theta$ based on quasi-score (1.2), we compute the mean and variance of $Y_i$ as

$$\mu_i(\theta) = m_{i1}\theta_1 + \cdots + m_{ip}\theta_p,$$
$$V_i(\theta) = m_{i1}\theta_1(1 - \theta_1) + \cdots + m_{ip}\theta_p(1 - \theta_p).$$

The quasi-score

$$u_k(\theta) = \sum_{i=1}^{n} m_{ik}(y_i - \mu_i(\theta))/V_i(\theta), \qquad (k = 1, \ldots, p) \tag{4.2}$$

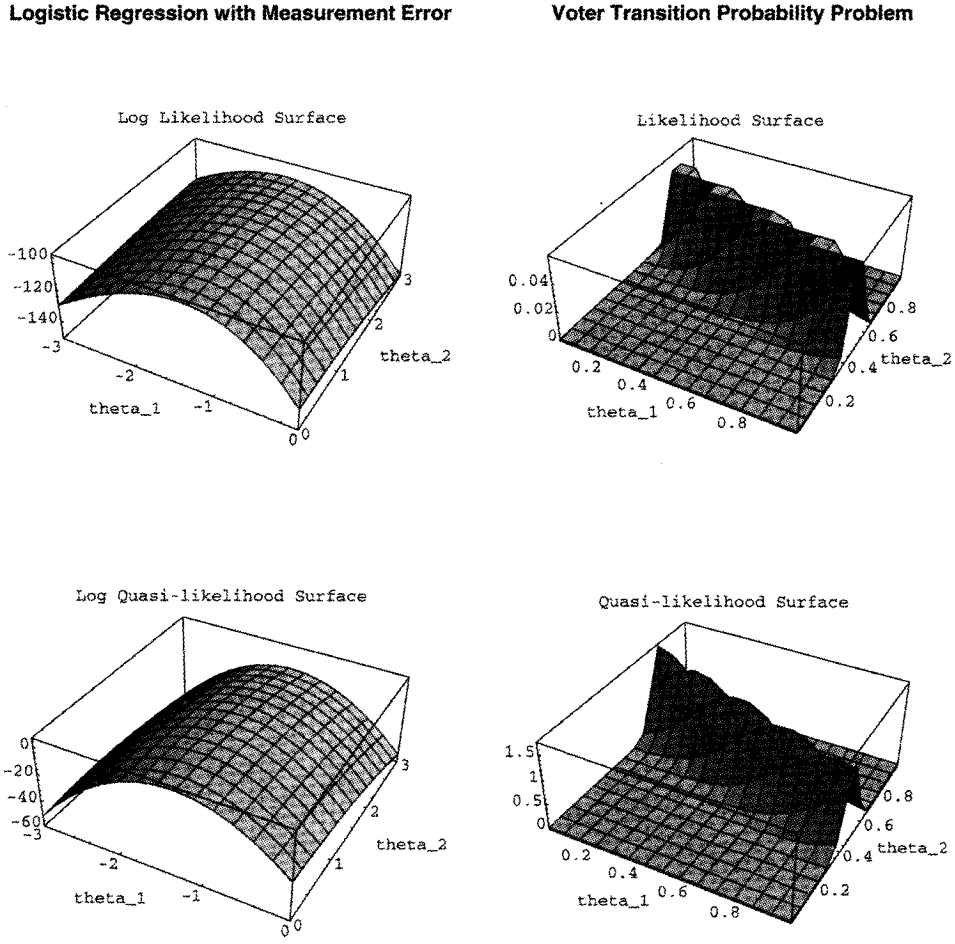


Fig. 3. True (conditional) likelihood surfaces vs. quadratic quasi-likelihood surfaces (3.6). The left and right panels correspond to logistic regression with measurement error and voter transition probability estimation problem respectively; cf. Section 4 for details.

is not conservative, because $d(u_1(\theta)d\theta_1 + \cdots + u_p(\theta)d\theta_p) \neq 0$.

Figure 2, top right, displays the nonconservative vector field (4.2) for $p = 2$, for a particular sample of size 10 and the true value of $\theta$ is assumed $(0.6, 0.4)$. The $m_{ij}$'s are taken as certain pseudo-random integers. As in the previous example, figures at center and bottom right are Helmholtz potential and divergence-free vector fields corresponding to the quadratic potential (3.6); linearization has been made at $\xi = (0.4, 0.6)$. Figure 3, right panel, displays the discrete true likelihood surface (McCullagh and Nelder (1989), p. 338) and the quadratic quasi-likelihood surface.

Now we turn to quasi-likelihood ratio tests. Simulations are carried out in two cases: (a) 2-party model with null hypothesis $H_0 : \theta = (0.6, 0.4)$ and (b) 3-party

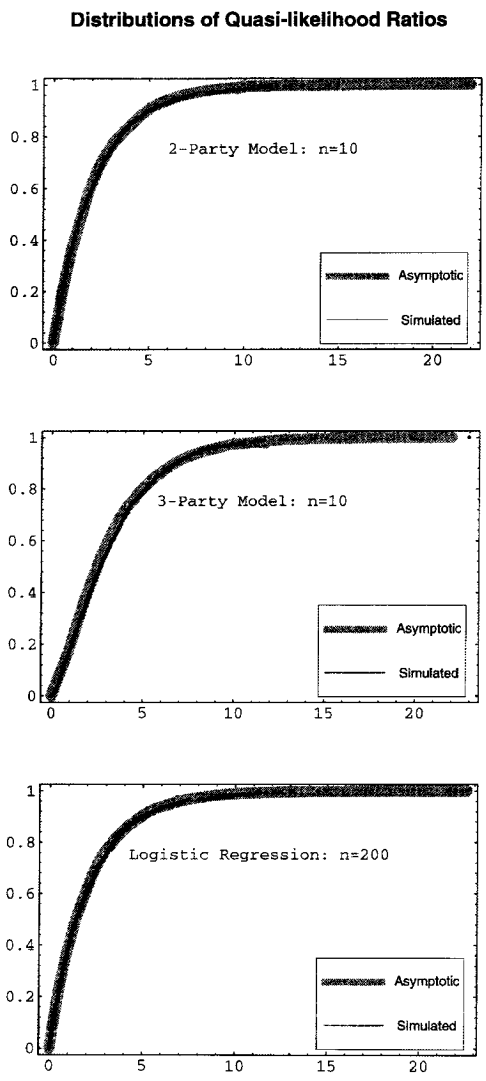


Fig. 4. Asymptotic null distributions of quasi-likelihood ratio statistics (3.8) vs. simulated true distributions. Top and center plots correspond to voter transition probability estimation problem with two and three parties respectively, the bottom plot is logistic regression with measurement error; $n$ is sample size. In each plot the simulated distribution is based on 1000 replications; cf. Section 4 for details.

model with null hypothesis $H_0 : \theta = (0.6, 0.4, 0.3)$. Quasi-score are linearized in (a) at $(0.4, 0.6)$, and in (b) at $(0.4, 0.6, 0.4)$. In both cases, sample size is taken as 10. Figure 4, top and center, compares the asymptotic null distributions with simulated true distributions of the quasi-likelihood ratio statistics. The asymptotic null distributions are in case (a) $\lambda_1 Z_1^2 + \lambda_2 Z_2^2$ with $\lambda = (1.67, 0.77)$, and (b) $\lambda_1 Z_1^2 + \lambda_2 Z_2^2 + \lambda_3 Z_3^2$ with $\lambda = (1.38, 0.98, 0.82)$. As in the first example, we also

Table 1. Voter transition probability model for $p$-parties. Conditional on number of voters $m_i$ for Party $P_i$ in a previous election, numbers of voters for Party $P_1$ in the next election $X_i$ are assumed independent binomial variables with index $m_i$ and transition probability $\theta_i (i = 1, \ldots, p)$.

| Party | $P_1$ | $P_2 + \cdots + P_p$ | Previous Votes |
|---|---|---|---|
| $P_1$ | $X_1 \sim B(m_1, \theta_1)$ | $m_1 - X_1$ | $m_1$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $P_p$ | $X_p \sim B(m_p, \theta_p)$ | $m_p - X_p$ | $m_p$ |
| Total | $Y = X.$ | $m. - X.$ | $m.$ |

tried several values of $\xi$ for linearization and again similar satisfactory fits have been observed.

## 5. Discussion

For general nonconservative estimating functions we proposed locally quadratic (log) quasi-likelihood functions. The quasi-likelihoods admit the interpretation as a particular type of potential function of the original (linearized) estimating functions. Quadratic fits instead of the original highly non-linear estimating functions or even the (true) likelihood in general statistical inference problem may have their own merits (Le Cam (1975)). The approximate quadratic-likelihoods and the quasi-likelihood ratio statistics applied to quasi-scores share a number of likelihood properties: parameter-invariance, zero-unbiasedness, information-unbiasedness, etc.

Literatures on nonconservative estimating functions have been mainly on studies of constrains or choice of integrable estimating functions (e.g. McCullagh and Nelder (1989), pp. 334–336; Li and McCullagh (1994)). The approach amounts to restricting to conservative estimating functions, with the possibility of excluding nonconservative but statistically informative ones. Some other papers (e.g. Hanfelt and Liang (1995, 1997)) study path-dependent integration. We have taken the approach of constructing quasi-likelihood for any nonconservative estimating functions. Although interpretation of our proposed solution invoked quite involved theories, the recipe is simple: (1) linearize your nonlinear nonconservative estimating function and (2) form the quadratic (approximate) quasi-likelihoods.

Among the difficulties with present approach are choice of $\xi$ at which estimating functions are linearized, loss of higher-order information due to linearization, etc. The latter problem may be solved by considering a second-order approximation. The choice of $\xi$ may be simple in certain problems such as hypothesis testing while less obvious in other situations, where one may linearize at a proper consistent root to the original equations.

We have not discussed multiple root problem in this paper. One referee pointed out the possibility of using the quasi-likelihood (3.7) for distinguishing consistent root from inconsistent roots. We have seen that if $\xi = \theta^*$, the true

parameter value, then $2\phi(\hat{\theta}(\xi), \xi; Y)$ has an asymptotic chi-square distribution. Moreover, it can be shown that if $\xi \neq \theta^*$, then $2\phi(\hat{\theta}(\xi), \xi; Y)$ is of order $O_P(n)$ with a positive value. Hence the different behaviors of this statistic reveal whether $\xi$ is near the truth. We shall leave this as our future problem and conclude this paper by noting some recent references on multiple roots; Li (1993, 1996), Hanfelt and Liang (1995, 1997), Small and Yang (1999).

## Acknowledgements

The author thanks Professors Takemi Yanagimoto and Shinto Eguchi for helpful discussions which sparked his interests in estimating functions; Hideki Tanemura and Ken'ichi Sugiyama for many technical advices. He is particularly indebted to Dr. Sugiyama for the proof of Theorem 2.1. Grateful thanks are also due to an associate editor and a referee for pointing out a mistake in a previous version and for constructive comments.

## References

Abraham, R. and Marsden, J. E. (1978). *Foundations of Mechanics*, 2nd ed., Benjamin/ Cummings, Reading, Massachusetts.

Barndorff-Nielsen, O. E. (1995). Quasi profile and directed likelihoods from estimating functions, *Ann. Inst. Statist. Math.*, **47**, 461–464.

Firth, D. (1982). Estimation of voter transition matrices, MSc Thesis, University of London.

Fukaya, K. (1995). *Electric Field and Vector Analysis*, Iwanami Shoten, Tokyo (in Japanese).

Fukaya, K. (1996). *Analytical Mechanics and Differential Forms*, Iwanami Shoten, Tokyo (in Japanese).

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

Godambe, V. P. (ed.) (1991). *Estimating Functions*, Clarendon Press, Oxford.

Hanfelt, J. J. and Liang, K.-Y. (1995). Approximate likelihood ratios for general estimating functions, *Biometrika*, **82**(3), 461–477.

Hanfelt, J. J. and Liang, K.-Y. (1997). Approximate likelihood for generalized linear errors-in-variables models, *J. Roy. Statist. Soc. Ser. B*, **59**(3), 627–637.

Irwin, M. C. (1980). *Smooth Dynamical Systems*, Academic Press, London.

Johnson, N. L. and Kotz, S. (1970). *Continuous Univariate Distributions 2*, Wiley, New York.

Kobayashi, S. (1990). *Differential Geometry of Connections and Gauge Theory*, 3rd ed., Shokabo, Tokyo (in Japanese).

Le Cam, L. (1975), Discussion on Efron (1975), *Ann. Statist.*, **3**(11), 1223–1224.

Li, B. (1993). A deviance function for the quasi-likelihood method, *Biometrika*, **80**(4), 741–753.

Li, B. (1996). A minimax approach to consistency and efficiency for estimating equations, *Ann. Statist.*, **24**(3), 1283–1297.

Li, B. and McCullagh, P. (1994). Potential functions and conservative estimating functions, *Ann. Statist.*, **22**(1), 340–356.

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

McCullagh, P. (1991). Quasi-likelihood and estimating functions, *Statistical Theory and Modelling: In Honour of Sir David Cox* (eds. D. V. Hinkley, N. Reid and E. J. Snell), 265–268, Chapman & Hall, London.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed., Chapman & Hall, London.

McLeish, D. L. and Small, C. G. (1992). A projected likelihood function for semiparametric models, *Biometrika*, **79**, 93–102.

Nelder, J. A. and Wedderburn, W. M. (1972). Generalized linear models, *J. Roy. Statist. Soc. Ser. A*, **135**(3), 370–384.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations, *Econometrika*, **16**, 1–32.

Siegel, C. L. and Moser, J. (1991). *Lectures on Celestial Mechanics*, Springer, New York.

Small, C. G. and McLeish, D. L. (1994). *Hilbert Space Methods in Probability and Statistical Inference*, Wiley, New York.

Small, C. G. and Yang, Z. (1999). Multiple roots of estimating functions, *Canad. J. Statist.*, **27**, (to appear).

Stefanski, L. A. and Carroll, R. J. (1985). Covariate measurement error in logistic regression, *Ann. Statist.*, **13**(4), 1335–1351.

Stefanski, L. A. and Carroll, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models, *Biometrika*, **74**(4), 703–716.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method, *Biometrika*, **61**(3), 439–447.