Institute of Mathematical Statistics

LECTURE NOTES — MONOGRAPH SERIES

AVOIDING THE LIKELIHOOD

C.C.Heyde
Columbia University and Australian National University

ABSTRACT

For the estimation of a finite dimensional parameter in a stochastic model it has become increasingly clear that it is usually possible to replace likelihood based techniques by quasi-likelihood alternatives in which only assumptions about means and covariances are made in order to obtain estimators. If it is available, the likelihood does provide a basis for benchmarking of alternative approaches but not more than that. The challenge is to see whether everything that can be done via likelihoods has a corresponding quasi-likelihood approach from which the likelihood based results can be recovered, if they are available. It is conjectured that this is the case. In this paper, various illustrations are sketched of avoiding the likelihood in contexts where alternative approaches have not been obvious.

Key Words: Quasi-likelihood; E-M algorithm; constrained estimation; nuisance parameters; diffusions; REML estimation.

1 Introduction

This paper is concerned with promoting the thesis that:

For parameter inference

(1) it is advantageous to *make minimalist assumptions* on models (initially concerning only means and covariance structure), and

(2) *there is a sensible quasi-likelihood (QL) alternative/ generalization of any likelihood based methodology*, at least to the first order of asymptotics.

We have also come to rely on the full distribution theory as a basis for a wide range of statistical procedures. Indeed, questions of appropriateness of the model are often supressed in order to make use of easy analytical methods (as with the Black-Scholes model in Finance). However, many ostensibly likelihood based methods do not actually require full distributional assumptions. They can readily be extended to the estimating functions context when there is a conservative quasi-score. That is, an estimating function which is the gradient of a scalar objective function which plays the role of the likelihood if and when it exists. However, it is argued that a scalar objective function for which the quasi-score is the gradient is inessential.

These pronouncements may be regarded as controversial. However, they are motivated by a wish to promote serious consideration and debate and not out of intrinsic dogmatism.

We shall give a smorgasbord of examples to elucidate the point of view described above. After discussion the general QL framework (Section 2), we shall describe some Projection-Solution (P-S) methods, namely in the contexts of constrained parameter estimation, nuisance parameters and the E-M algorithm (Section 3). Then we shall discuss bypassing the likelihood through examples for diffusion processes and REML estimation (Section 4). There are, of course, many other areas in which there is substantial progress towards the use of estimating functions without direct recourse to likelihood ideas. These include the areas of multiple roots (e.g. Heyde and Morton (1996b)), likelihood ratio tests (e.g Li (1993)) and Bayesian analysis (e.g. Godambe (1994)). A much broader perspective will soon be available in Heyde (1997).

2 General QL Principles

Suppose we have a sample $\{X_t, t \in T\}$ of vectors of dimension r, T possibly being discrete, continuous or lattice. The possible probability measures $\{P_\theta\}$ for $\{X_t\}$ are the union of families of models and the $\theta = (\theta_1, ..., \theta_p)'$ to be estimated is a vector of dimension p.

The approach is via the set of p dimensional vector estimating functions

$$\mathcal{G} = \{G_T(\theta) = G_T(\{X_t, t \in T\}, \theta)\}$$

which are functions of the data and θ for which $EG_T = 0$ for each P_θ and the matrices

$$E\dot{G}_T(\theta) = (E\partial G_{T,i}(\theta)/\partial\theta_j)$$

and $EG_T(\theta)G_T(\theta)'$ are nonsingular, the prime denoting transpose.

The QL theory focuses on suitably chosen subsets of $\mathcal{G}$ and involves choice of an estimating function G_T to maximize, in the partial order of non-negative definite (nnd) matrices, the *information criterion*:

$$\mathcal{E}(G_T) = (E\dot{G}_T)'(EG_TG_T')^{-1}(E\dot{G}_T)$$

which is a natural generalization of Fisher information. We have the following definition.

Definition. Suppose that $G_T^* \in \mathcal{H} \subset \mathcal{G}$. If

$$\mathcal{E}(G_T^*) - \mathcal{E}(G_T)$$

is nnd for all $G_T \in \mathcal{H}$ we say that G_T^* is a *quasi-score estimating function* (QSEF) within $\mathcal{H}$.

The choice of the family $\mathcal{H}$ is completely open and should be tailored to the particular application.

The estimator θ_T^* obtained from $G_T^*(\theta_T^*) = 0$ which is termed a *quasi-likelihood estimator* has, under broad conditions, certain minimum size asymptotic confidence zone properties for θ, at least within $\mathcal{H}$. Indeed, the basic properties are those of the maximum likelihood estimator, but restricted to the class $\mathcal{H}$.

The theory does not require a parametric setting, let alone the existence of a likelihood score function $U_T(\theta)$. However, if $U_T \in \mathcal{H}$, as can ordinarily be arranged in exponential family problems, then U_T is the QSEF within $\mathcal{H}$ *and can easily be calculated without using likelihoods.*

It is not usually practicable to find a quasi-score estimating function directly from the definition. However, the criterion given in the following proposition is easy to use in practice.

Proposition 1. Let $\mathcal{H} \in \mathcal{G}$. Then $G_T^* \in \mathcal{H}$ is a quasi-score estimating function within $\mathcal{H}$ if

$$(E\dot{G}_T)^{-1} E G_T G_T^{*\prime} = C_T \qquad (2.1)$$

for all $G_T \in \mathcal{H}$, where C_T is a fixed matrix. Conversely, if $\mathcal{H}$ is convex and G_T^* is a quasi-score estimating function then (2.1) holds.

3 Projection Based Methods

Many problems for which the use of likelihood based ideas is standard, and for which there is not an obvious quasi-likelihood analogue in the absence of a conservative quasi-score, can be dealt with via projection based methods. We give three illustrations in this section.

3.1 Constrained Parameter Estimation

Here we wish to estimate θ subject to the constraint $F'\theta = d$, F being a $q \times p$ matrix which does not depend on the data or θ.

Suppose we have an unconstrained quasi-score $Q(\theta) \in \mathcal{H}$ and, using a minus to denote generalized inverse, define the projection matrix

$$P = F(F'V^{-1}F)^{-}F'V^{-1}$$

for $V = EQQ'$.

In the case where a likelihood $L(\theta)$ is available, the usual procedure is to use the method of Lagrange multipliers and maximize $L(\theta) + \lambda'(F'\theta - d)$ where λ is determined by the constraint. Thus, we differentiate with respect to θ and solve the equations

$$U(\tilde{\theta}) + F\lambda = 0, F'\tilde{\theta} = d$$

for λ and $\tilde{\theta}$, U being the score function.

The striking thing is that this procedure works in general for quasi-likelihood, even in the non-conservative case. We solve the equations

$$Q(\tilde{\theta}) + F\lambda = 0, F'\tilde{\theta} = d$$

for λ and $\tilde{\theta}$, that is,

$$(I - P)Q(\tilde{\theta}) = 0, F'\tilde{\theta} = d.$$

Optimality is preserved. For details see Heyde and Morton (1993).

3.2 Nuisance Parameters

Here we have $\theta' = (\phi', \psi')$ where ϕ is the parameter of interest and ψ is a nuisance parameter. Then, supposing that we have a quasi-score $Q(\theta) \in \mathcal{H}$, we use the partitioned forms

$$Q = \begin{pmatrix} Q_\phi \\ Q_\psi \end{pmatrix}$$

$$V = EQQ' = \begin{pmatrix} V_{\phi\phi} & V_{\phi\psi} \\ V_{\psi\phi} & V_{\psi\psi} \end{pmatrix}$$

and write

$$F_\phi = \begin{pmatrix} V_{\phi\phi} \\ V_{\psi\phi} \end{pmatrix}$$

$$F_\psi = \begin{pmatrix} V_{\phi\psi} \\ V_{\psi\psi} \end{pmatrix}.$$

The projection

$$P_\psi = F_\psi(F'_\psi V^{-1} F_\psi)^{-} F'_\psi V$$

identifies the information about ψ for ϕ given and the estimating equation

$$(I - P_\psi)Q = 0$$

is optimal for the estimation of ϕ in the presence of the nuisance parameter ψ. The sensitive dependence of Q on ψ has been removed in the sense that

$$E(\frac{\partial}{\partial \psi'}(I - P_\psi)Q) = 0.$$

This is a first order approach. In the language of McLeish and Small (1988), $(I - P_\psi)Q$ is locally E-ancillary for ψ and $P_\psi Q$ is locally E-sufficient for ψ.

3.3 E-M Algorithm Generalization

The E-M method is used for parameter estimation where there is missing data. In the first (E) step, one takes the conditional expectation of the complete data likelihood with respect to the available data. In the second (M) step, one maximizes over possible distributions. However, it is possible to avoid the likelihood completely by introducing a project-solve (P-S) method.

Suppose that the full data is denoted by x, the observed data by y and θ is the parameter of interest. We seek to adapt a quasi-score $Q(\theta; x) \in \mathcal{H}_x$ to obtain a quasi-score $Q(\theta; y) \in \mathcal{H}_y$. In fact

$$E(Q^* - Q)(Q^* - Q)' = \inf_{G \in \mathcal{H}_y} E(G - Q)(G - Q)'$$

and Q^* is the element of $\mathcal{H}_y$ with minimum dispersion distance from $Q \in \mathcal{H}_x$.

If the likelihood score $U \in \mathcal{H}_x$, then $Q = E(U|y)$ provided this belongs to $\mathcal{H}_y$ as in the E-M case. However, Q^* is given in general just as a least squares predictor and mostly

$$Q^*(\theta; y) \neq E_\theta(Q(\theta, x)|y).$$

A detailed discussion of the method can be found in Heyde and Morton (1996a).

There is an algorithm for solving $Q^*(\theta; y) = 0$ along the lines of the E-M algorithm. This usually gives a first order rate of convergence. However, the equation can often be solved directly with a second order rate of convergence, for example using Fisher's method of scoring.

4 Bypassing the Likelihood

In this section we give examples of the derivations of score functions without having first to find a likelihood to differentiate. In such cases there is the added advantage of being useful under distinctly broader distributional conditions than are imposed by the need to prescribe a likelihood.

4.1 Parameters in Diffusion Type Models

Here we have a model described by the stochastic differential equation

$$dX_t = a(t, X_t, \theta)dt + b^{\frac{1}{2}}(t, X_t)dW_t$$

where a, b are known functions and W_t is standard Brownian motion.

The usual approach to estimation of θ is to obtain an appropriate Radon-Nikodym derivative. This is tedious from first principles. Differentiation with respect to θ then gives the likelihood score.

Alternatively, one may consider the family of martingale estimating functions

$$\mathcal{H} = \{\int_0^T k_t(\theta)(dX_t - a(t, X_t, \theta)dt), k_t predictable\}.$$

The quasi-score estimating function from this family can be written down almost immediately using Proposition 1 as

$$\int_0^T (\dot{a}(t, X_t, \theta))'(b(t, X_t))^{-1}(dX_t - a(t, X_t, \theta)dt),$$

and *this is equivalent to the likelihood score.*

The explanation is straightforward. Note that the elements of $\mathcal{H}$ are (martingale) stochastic integrals with respect to W_t. Also, a likelihood score is a martingale under modest regularity conditions. Furthermore, all square integrable martingales living on the same probability space as this process can be described as stochastic integrals with respect to the Brownian motion. Thus $\mathcal{H}$ contains the likelihood score and the QSEF will pick it out.

Now the quasi-likelihood method goes much further. One does not have to perturb a diffusion type model much to destroy the likelihood. For example, in the Cox-Ingersoll-Ross model used for interest rates in financial modelling,

$$dX_t = \alpha(\beta - X_t)dt + \sigma X_t^{\frac{1}{2}} dW_t,$$

the Radon-Nikodym derivative will not exist if the volatility σ is rate dependent on α. However, the QSEF is unaffected. For more details see Heyde (1994a).

4.2 Restricted (or Residual) Maximum Likelihood

Here the problem is of estimating dispersion in a linear model; the $n \times r$ vector y has the multivariate normal distribution $MVN(X\beta, V(\theta))$ with mean $X\beta$, covariance $V(\theta)$ and θ is to be estimated.

Take the rank of X as r, the dimension of β and let A be any matrix with n rows and rank $n-r$ satisfying $A'X = 0$. Then $A'y$ has the $MVN(0, A'VA)$ distribution.

The striking thing here is that the likelihood function does not depend on A. Indeed, for all A,

$$A(A'VA)^{-}A' = V^{-1}Q$$

where $Q = I - P$, P being the projector onto the subspace $R(X)$ (the range space of X) with respect to the inner product $a'Vb$.

The likelihood function of θ based on $A'y$ is, omitting a constant multiplier,

$$(\prod_{i=1}^{n-r} l_i)^2 exp(-\frac{1}{2}y'V^{-1}Qy)$$

where the l_i are the non-zero eigenvalues of $V^{-1}Q$. This is not a straightforward calculation, nor is the differentiation with respect to θ required to obtain the REML estimating equations

$$tr(V^{-1}Q\frac{\partial V}{\partial \theta_i}) = y'(V^{-1}Q\frac{\partial V}{\partial \theta_i}V^{-1}Q)y, i = 1, 2, ...p,$$

tr denoting trace.

For the quasi-likelihood approach we no longer require multivariate normality but instead that y has mean vector $X\beta$, covariance matrix $V(\theta)$, and that each y_i has kurtosis 3.

The crucial step is taken by noting that *we expect to use quadratic functions of the data to estimate covariances.*

For fixed A, let $z = A'y$ and take θ as a scalar for clarity. Now introduce the family of estimating functions

$$\mathcal{H} = \{G(S) = z'Sz - Ez'Sz, S symmetric\}.$$

Write $W = A'VA$. Then

$$EG(S)G(S^*) = 2tr(WSWS^*)$$

$$Ez'Sz = tr(WS)$$

$$E\dot{G}(S) = -tr(\frac{\partial W}{\partial \theta}S),$$

and we see via Proposition 1 that S^* for the QSEF is given by

$$S^* = W^-\frac{\partial W}{\partial \theta}W^-.$$

The REML estimating equations then follow since

$$AW^-A' = V^{-1}Q.$$

For more details see Heyde(1994b).

References

Godambe, V.P. (1994). Linear Bayes and optimal estimation. Technical Report STAT-94-11, University of Waterloo, Canada.

Heyde, C.C. (1994a). A quasi-likelihood approach to estimating parameters in diffusion-type processes. *Studies in Applied Probability. J. Applied Prob.*, 31A, 283-290.

Heyde, C.C. (1994b). A quasi-likelihood approach to the REML estimating equations. *Statistics and Probability Letters*, 21, 381-384.

Heyde, C.C. (1997). *Quasi-Likelihood Theory and its Application.* Springer, New York.

Heyde, C.C. and Morton, R. (1993). On constrained quasi-likelihood estimation. *Biometrika*, 80, 755-761.

Heyde, C.C. and Morton, R. (1996a). Quasi- likelihood and generalizing the E-M algorithm. *J. R. Statist. Soc. Ser. B*, 58, 317-327.

Heyde, C.C. and Morton, R. (1996b). Multiple roots and dimension reduction issues for general estimating equations. Unpublished manuscript.

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

McLeish, D.L. and Small, C.G. (1988). *The Theory and Applications of Statistical Inference Functions*, Lecture Notes in Statistics Vol.44, Springer, New York.