THE METHOD OF QUASILIKELIHOOD

Yan-xia Lin

B.Sc. Fujian Normal University, P.R.China

M.Sc. The University of Jordan, Jordan

A thesis submitted for the degree of

Doctor of Philosophy

in the Australian National University

March 1992

DECLARATION

In accordance with the regulations of the Australian National University, I wish to state that the work described in this thesis was carried out by myself except where due reference is made.

Lin Yan-xia

Yan-xia Lin

ACKNOWLEDGMENTS

The research work for this thesis was carried out at the Australian National University and financial assistance was provided by an Australian National University Ph.D scholarship.

Firstly, I would like to thank my supervisor Prof. C.C.Heyde for his many invaluable comments and suggestions, for his unfailing interest and encouragement and for his careful reading the original draft.

I would also like to express my appreciation to members of the Statistics Research Section, School of Mathematics Science for their help and for providing a congenial environment in which to work.

Most of the research of Chapters 4 and 6 was carried out jointly with Professor Heyde. A large part of the work of Chapter 4 appears in [50] and of Chapter 6 in [51]. The remaining chapters are my own work.

To Yi, Chong and My Parents

NOTATIONS

In this thesis, we always discuss our questions in the framework of a probability space $(\Omega, \mathcal{F}, P)$, although we do not always mention this specifically.

R^k	k-dimensional Euclidean space
B^c	the complement of the set B
$I_{(A)}$	the indicator function of the set B
$\{\mathcal{F}_t\}$	σ-fields
natural σ-field $\{\mathcal{F}_t\}$	$\mathcal{F}_t = \sigma\{X_s, s \leq t\}$
$\mathcal{F}_t \bigvee \mathcal{F}_s$	the $\sigma-$field generated by $\mathcal{F}_t$ and$\mathcal{F}_s$
$\mathcal{F}_t \bigwedge \mathcal{F}_s$	the maximum $\sigma-$field contained in $\mathcal{F}_t$ and $\mathcal{F}_s$
Θ	parameter space
$\dot{G}_T(\theta)$	$\dot{G}_T(\theta) = (\partial G_{T,i}/\partial \theta_j)$
$\mathcal{G}_T$ or $\mathcal{G}_n$	an estimating function space in which every element G satisfies the conditions that $E\dot{G}$ and EGG' are nonsingular
$\lim_{n\to\infty} X_n = X$, (P); or $X_n \xrightarrow{p} X$, as $n \to \infty$	X_n $P-$converges to X
$\lim_{n\to\infty} X_n = X$ (a.s.); or $X_n \xrightarrow{a.s.} X$, as $n \to \infty$	X_n converges to X almost surely
diag (A)	diagonal matrix of A
diag $\{A\}$	the set of diagonal elements of A
$tr(A)$	the trace of the matrix A
X'	the transpose of the vector or matrix X
$\lambda_{max}(A)$	the maximum eigenvalue of A
$\lambda_{min}(A)$	the minimum eigenvalue of A
$\|A\|$	$\|A\| = (tr(AA'))^{1/2}$
$\|A\|_1$	$\|A\|_1 = (\lambda_{max}(AA'))^{1/2}$
$\lvert a \rvert$	for a vector a, the Euclidean norm

p.d.	positive definite
i.i.d.	independent and identically distributed
$N(\mu, B)$	the multivariate normal distribution with mean vector μ and variance matrix B
$O(a_n)$	$\lvert O(a_n)/a_n \rvert$ is bounded, as $n \to \infty$
$o(a_n)$	$\lvert o(a_n) \rvert \to 0$, as $n \to \infty$
A^+	the Moore-Penrose generalized inverse of A i.e. $AA^+A = A$, $A^+AA^+ = A^+$, $A^+A = AA^+$
$f \hat{=} g$	the left side is definited by the right side

CONTENTS

REFERENCES 171

CHAPTER 1
INTRODUCTION AND SUMMARY

1 Historical Review

The methods of "Maximum Likelihood " (ML) and "Least Squares" (LS) are two central methods in the theory of inference. They have being developed and studied for a long time, and they still play an important role in statistics. However, as time goes on, more and more understanding of natural phenomena has been obtained and more diverse requirements have been introduced into statistical practice. Sometimes those phenomena and requirements cause some difficulties in the use of the ML and LS methods. Therefore statisticians have to look in detail into the advantages and weaknesses of each of the methods.

The ML method was introduced by Fisher in 1921. Let X have density $f(x;\theta)$ for $x \in R^m$, $\theta \in \Theta \subseteq R^d$; let X_1, X_2, ..., X_n be random samples of X. The joint density of the X_i's, as a function of θ, is called the likelihood function:

$$L_n(x_1,\ldots,x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

The maximum likelihood estimator(MLE) of θ is defined to be a value $\hat{\theta}_n(x_1,\ldots,x_n)$ (if it exists) for which

$$L_n(x_1,\ldots,x_n;\hat{\theta}_n) = \max_{\theta\in\Theta} L_n(x_1,\ldots,x_n;\theta).$$

Therefore, in order to construct a likelihood function it is usually necessary to assume a probabilistic mechanism specifying, for a range of parameter values, the probabilities of all relevant samples that might possibly have been observed. Such a specification implies either knowledge of the mechanism by which the data were generated or substantial experience of similar data from previous experiments. Some of this knowledge is not easy to obtain and, in practice, statisticians frequently have difficulty with it. An inference method which is free from the full distribution of the observations is desirable. The LS method satisfies this requirement. The LS method is much more widely applicable than the ML method. The latter requires full distributional specification while the former is usually applicable under assumptions on only the first two moments.

The LS method was introduced by Legendre in 1805 on more or less intuitive grounds. He also coined the term " LS method". The principle of least squares can be stated as follows: choose as the " best-fitting line", the " line" that minimizes the sum of squares of the deviations of the observed value from those predicted. Gauss(1809, 1823) provided two statistical justifications to the LS method. (1) He related it to ML estimation and the normal model by showing that within the location family of distributions, the two methods LS and ML coincide uniquely for the normal family. This indicated that the acceptance of both LS and ML methods was in line with the acceptance of the normal model. (2) He showed that with conditions only on the first two moments of the distribution but otherwise irrespective of its distributional form, the LS estimate has the minimum variance within the class of " linear unbiased estimates". This is the well-known Gauss-Markov(GM) theorem. Indeed, concentrating on the smallest sum of squares of errors, the LS method is a powerful method for estimating parameters if the deviations of the observed value from those predicted satisfy some conditions such as independence, constant variance and so on. But when the error variance depends on the parameters, the LS method will lose some of its precision. Here is an example from Godambe and Kale(1991).

Consider independent random variates $y_1, y_2,, \ldots, y_n$ such that

$$E(y_i) = \alpha_i(\theta) \qquad \text{and} \qquad var(y_i) = c\sigma_i^2(\theta),$$

where $\{\alpha_i(\theta)\}$ and $\{\sigma_i^2(\theta)\}$ are specified differentiable functions of θ and c is an unknown positive constant not depending on θ. To estimate θ, the LS approach requires minimization of $\sum(y_i - \alpha_i(\theta))^2/\sigma_i^2(\theta)$, leading to the LS estimating equation

$$\tilde{g}_1^* + B = 0,$$

where $\tilde{g}_1^* = \sum(y_i - \alpha_i(\theta))(\partial\alpha_i/\partial\theta)\sigma_i^2(\theta)$ and $B = \sum(y_i - \alpha_i(\theta))^2(\partial\sigma_i/\partial\theta)\sigma_i^{-3}(\theta)$. We first note that whereas $E(\tilde{g}_1^*) = 0$, $E(B) = \sum(\partial\sigma_i/\partial\theta)\sigma_i^{-1}(\theta) = \sum(\partial log\sigma_i/\partial\theta)$ is in general non-zero and thus $\tilde{g}_1^* + B$ is not an unbiased estimating function . Even for large samples, although $\tilde{g}_1^*/n$ converges in probability to zero this may not be the case for B/n, and in fact B/n could diverge to $\pm\infty$ depending on the nature of $1/n \sum(\partial log\sigma_i/\partial\theta)$. Thus for large n, the solution of the equation $\tilde{g}_1^* = 0$, under some regularity conditions, will converge in probability to the true value whereas the solution of the LS estimating equation may not. We notice that $\tilde{g}_1^* = 0$ is just the weighted least squares estimating equation. That means that weighting may becomes necessary to remedy the disadvantage of the least squares method in some circumstances.

It is well known that, if the likelihood function has the exponential family form, maximum likelihood estimates of regression parameters can often be found using the method of weighted least squares. In the seventies and eighties, people understood the fact that the method of weighted least squares can be used to find maximum likelihood estimates even in some case where the likelihood function does not have the exponential family form(see, Nelder and Wedderburn, 1972; Bradley, 1973; Wedderburn, 1974; Jennrich and Moore, 1975 and Jørgensen, 1983). This led people to seek the application of the weighted least squares method to more general problems(e.g. McCullagh, 1983). Here is an example concering regression with constant coefficient of variation. Suppose that the random variables $Y_1, \cdots, Y_n$ are uncorrelated, that $E(Y_i) = \mu_i < \infty$, and that $var(Y_i) = \sigma^2\mu_i^2$, so that the coefficient of variation, σ, rather than the variance, is constant over all observations. Suppose further that inference is required for (θ_0, θ_1), where

$$log(\mu_i) = \theta_0 + \theta_1(x_i - \bar{x}), \qquad i = 1, 2, \cdots, n,$$

and $x_1, \cdots, x_n$ are known constants, $\bar{x}$ being their mean.

In this example, the regression is not linear; the covariance matrix of $\boldsymbol{Y}$ is given by

$$cov(\boldsymbol{Y}) = \sigma^2\boldsymbol{V}(\boldsymbol{\mu}),$$

where $\boldsymbol{V}(\cdot)$ is a matrix of known functions and σ^2 is known as the dispersion parameter; the model is specified entirely in terms of first and second moments of $\boldsymbol{Y}$, which is called a semiparameter model or nonparametric model, so that it is not possible to write down the likelihood function. As we have said before, it is impossible to apply the ML method to it and it is not good to apply the LS method either, since the variation contains an unknown parameter. We require a method of estimation that is reliable under the second moment assumptions just described. This is one of the important reasons to motivate statisticians to establish the quasi-likelihood method.

2 The Quasi-likelihood Method

The term "quasi-likelihood " was first proposed by Wedderburn(1974) for the estimation of parameters in regression models. He observed that, from a computational point of view, the only assumptions on a generalized linear model necessary to actually fit the model was a specification of the mean (in terms of the regression parameters) and the relationship between the mean and the variance and not necessarily a fully specified likelihood. Therefore he replaced the assumptions on the probability distribution by defining a function based solely on the mean-variance relationship which had algebraic and frequency properties similar to those of log-likelihoods. For the arbitrary nonlinear regression,

$$\boldsymbol{Y} = \boldsymbol{\mu}(\boldsymbol{\theta}) + \boldsymbol{e},$$

where $E\ \boldsymbol{Y} = \boldsymbol{\mu}(\boldsymbol{\theta})$ and $cov\{\ (\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\theta}))\} = \boldsymbol{\sigma}^2\boldsymbol{V}$. Under regularity conditions, a function Q is defined by the differential equations

$$\frac{\partial Q}{\partial \boldsymbol{\theta}} = \boldsymbol{U}(\boldsymbol{\theta}) = \boldsymbol{D}'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\theta})),$$

where $\boldsymbol{D}=\{\partial\mu_i/\partial\theta_j\}$ is a matrix and is a function of $\boldsymbol{\theta}$. It may be verified that

$$E\{\boldsymbol{U}(\boldsymbol{\theta})\} = 0,$$

$$E\{\frac{\partial}{\partial \boldsymbol{\theta}}(\boldsymbol{U}(\boldsymbol{\theta}))\} = -\boldsymbol{D}'\boldsymbol{V}^{-1}\boldsymbol{D}$$

$$cov\{\ \boldsymbol{U}(\boldsymbol{\theta})\} = \sigma^2\boldsymbol{D}'\boldsymbol{V}^{-1}\boldsymbol{D}.$$

So that, apart from the factor σ^2 above, $\boldsymbol{U}(\boldsymbol{\theta})$ behaves like the derivative of a log-likelihood and is termed a quasi-score or quasi-score estimating function (estimating functions being introduced below), while Q itself is called a quasi-(log)likelihood. This is an objective definition. It may be difficult to decide what distribution one's observations follow, but the form of the mean-variance relationship is often much easier to postulate. This is what makes quasi-likelihood useful and why the quasi-likelihood method has received considerable attention from statisticians. After Wedderburn(1974) put forward this new concept, McCullagh(1983), Firth(1987), Nelder and Pregibon(1987), Wiley(1987), Davidian and Carroll(1988), McCullagh and Nelder(1989) and others studied different aspects of properties of the quasi-likelihood method based on Wedderburn's interpretation. We notice that the quasi-likelihood concept given by Wedderburn is more or less restricted to the weighted least squares method and the generalized linear model framework. When that quasi-likelihood concept was applied to a model, it was always assumed that the errors in the model were independent or uncorrelated, the variances of the errors were specified functions of the mean together with a possibly unknown scalar multiplicative parameter, something like the model at the end of Section 1, and the quasi-score equation was always of the form $\boldsymbol{D}'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\theta}))$, where $\boldsymbol{\mu}(\boldsymbol{\theta}) = E\boldsymbol{Y}$.

While Wedderburn's quasi-likelihood concept was investigated and applied by some statisticians, there were separate developments in the field of estimating functions. For example, Godambe(1960), Godambe and Thompson(1974), Godambe(1976), Hutton and Nelson(1986) and Godambe and Heyde(1987) also defined and investigated a quasi-likelihood concept. The complete and detailed definition of the quasi-score estimating function under the view-point of estimating function was given in Godambe and Heyde(1987). In that paper a quasi-score estimating function is interpreted as an optimal estimating function in a given estimating function space under some criterion. The prime idea of the criterion is linked with the Gauss-Markov (GM) theorem, the Cramér-Rao inequality and the Fisher information(also see Godambe and Kale (1991)). The detailed explanation will appear later in this section. Here we first state the definition of the quasi-score estimating function given by

Godambe and Heyde(1987).

Let $\{X_t, 0 \leq t \leq T\}$ 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 θ taking value in an open subset Θ of p-dimensional Euclidean space. Suppose that the possible probability measures for $\{X_t\}$ are $\mathcal{P} = \{\mathcal{P}_\theta\}$, a union of parametric families each family being indexed by the same parameter θ and that each $(\Omega, \mathcal{F}, P_\theta)$ is a complete probability space. Let $(\mathcal{F}_t, t \geq 0)$ denote a standard filtration, $\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F}$ for $s \leq t$; $\mathcal{F}_0$ is augmented by sets of measure zero of $\mathcal{F}$ and $\mathcal{F}_t = \mathcal{F}_{t+}$, where $\mathcal{F}_{t+} = \bigcap_{s>t} \mathcal{F}_s$.

Let Ψ_T be a class of zero mean, square integrable estimating functions $G_T = G_T(\{X_t, \quad 0 \leq t \leq T\}, \theta)$ for which $EG_T(\theta) = 0$ for each $P_\theta \in \mathcal{P}$ with index θ and $\mathcal{G}_T$ be a subset of Ψ_T in which every element G_T is almost surely differentiable with respect to the components of θ and for which $E\dot{G}_T(\theta) = (E\partial G_{T,i}(\theta)/\partial\theta_j)$ and $E(G_T(\theta)G_T{}'(\theta))$ are nonsingular. The expectations are always with respect to P_θ.

Following the above notation, Godambe and Heyde (1987) gave the following definition of the quasi-score estimating function .

<u>**Definition**</u> **1** *G_T^* is the quasi-score estimating function in $\mathcal{G}_T \subseteq \Psi_T$, iff,*

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

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

The root θ_T^* of the equation $G_T^*(\theta) = 0$ is called the quasi-likelihood estimator . The quasi-likelihood method is a procedure for (1)finding the quasi-score estimating function from the estimating function space ; (2) using the quasi-likelihood estimator as the estimator of the true parameter θ.

The properties of the quasi-score estimating function under this definition can be found from Godambe and Heyde(1987), Godambe and Thompson(1985, 1989), Heyde(1988, 1987, 1989a, 1989b), Heyde and R.Gay (1989), Heyde and Y.-X. Lin (1991, 1992), Hutton and P.I.Nelson (1986), Sørensen (1990) and this thesis.

The research on the quasi-likelihood concept or quasi-likelihood method from the two different fields mentioned above was more or less unlinked until about 1985 (see, Desmond (1989, 1991)). However, Godambe and Heyde's definition and Wedderburn's approach overlap and complement each other.

Let us consider a generalized linear models $Y_i = \mu_i + e_i$, $i = 1, 2, \cdots, n$ where $Y_i \in R^m$ and $e_i \in R^m$, as discussed in McCullagh and Nelder (1989). Suppose that the components of the response vector $\boldsymbol{Y} = (Y_1, \cdots, Y_n) \in R^{m\times n}$ are independent with mean $\boldsymbol{\mu}$ and the covariance matrix $\sigma^2 \boldsymbol{V}(\boldsymbol{\mu})$, where σ^2 may be unknown and $\boldsymbol{V}(\boldsymbol{\mu})$ is a matrix of known functions. Let $\boldsymbol{\theta}$ is a parameter which relates to the dependence of $\boldsymbol{\mu}$ on covariates $\boldsymbol{x}$. This relation need not concern us for this moment, so we write $\boldsymbol{\mu}(\boldsymbol{\theta})$, thereby absorbing the covariates into the regression function. Here σ^2 is assumed constant and does not depend on $\boldsymbol{\theta}$.

Let $\mathcal{F}_i = \sigma\{Y_j, j \le i\}$. Then $\{e_i\}$ is a martingale difference sequence with respect to $\{\mathcal{F}_i\}$. Godambe and Heyde(1987) gave an important formula for determining a quasi-score estimating function from an estimating function space $\mathcal{G} \subseteq \mathcal{H} = \{G = \sum \alpha_i m_i\}$ where $\{m_i\}$ is a martingale difference sequence with respect to some σ-fields $\{\mathcal{F}_i\}$, $\{\alpha_i\}$ is a predictable process with respect to the same σ-fields $\{\mathcal{F}_i\}$ and the elements in $\mathcal{G}$ satisfy the conditions that $E\dot{G}$ and EGG' are nonsingular. They pointed out that, in that estimating function space $\mathcal{G}$, the quasi-score estimating function has the form

$$G^* = \sum (E(\dot{m}_i|\mathcal{F}_{i-1}))'(E(m_i m_i'|\mathcal{F}_{i-1}))^+ m_i,$$

where " +" denotes the Moore-Penrose generalized inverse. Therefore applying this formula here, we can easily obtain that the quasi-score estimating function in $\{\sum \alpha_i e_i | \alpha_i \in \mathcal{F}_{i-1}\}$ is

$$\begin{aligned} G^* &= \sum_i (E(\dot{e}_i|\mathcal{F}_{i-1}))'(E(e_i e_i'|\mathcal{F}_{i-1}))^+ e_i \\ &= \sum_i (\frac{\partial \mu_i}{\partial \theta})' V_i^{-1}(Y_i - \mu_i), \end{aligned}$$

where $V_i = E(e_i e_i') = E((Y_i - \mu)(Y_i - \mu)')$. This is exactly the quasi-score estimating function obtained from Wedderburn's approach. From this example, we feel that, in some sense, using the view-point of estimating functions to define the quasi-likelihood estimating function can make the concept of the quasi-score estimating function clear.

Furthermore, as we shall notice from the definition of Godambe and Heyde, the quasi-score estimating function should be associated with an estimating function space. When we say a quasi-score estimating function , we should remember that there is an estimating function space behind it, otherwise we shall misuse the quasi-likelihood method. In fact the quasi-score estimating function which cames from Wedderburn's approach is just an optimal estimating function in an estimating function space of the form $\{\sum \alpha_i(y_i-\mu_i)\}$ where $\{\alpha_i\}$ are nonrandom. Let us consider the model in the end of Section 1 again. There, the model just assumes that $\{y_i\}$ are uncorrelated. In this case, we usually can not assume that $\{y_i - \mu_i = y_i - Ey_i\}$ is a martingale difference sequence. But anyway we can always find a sequence of σ-fields $\{\mathcal{F}_i\}$ such that

$$E(y_i - \mu_i|\mathcal{F}_i) = E(y_i - \mu_i) = 0.$$

Let $\mathcal{G}$ be an estimating function space and $\mathcal{G} = \{G = \sum_1^n \alpha_i(y_i - \mu_i)\}$, where α_i are nonrandom; $((\alpha_1, \cdots, \alpha_n)\boldsymbol{D}))$ and $(\alpha_1, \cdots, \alpha_n)\boldsymbol{V}(\alpha_1, \cdots, \alpha_n)'$ are nonsingular. Let

$$\begin{aligned} G^* &= \boldsymbol{D}'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\theta})) \\ &= \sum_i \alpha_i^*(y_i - \mu_i). \end{aligned}$$

Then $G^* \in \mathcal{G} \subseteq \mathcal{H}$. For any $G \in \mathcal{H}$ denoted by

$$G = \boldsymbol{P}(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\theta})) = \sum \alpha_i(y_i - \mu_i),$$

where $\boldsymbol{P} = (\alpha_i, \cdots, \alpha_n)$, we have

$$\begin{aligned} EGG^{*\prime} &= E\boldsymbol{P}(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\theta}))(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\theta}))'\boldsymbol{V}^{-1}\boldsymbol{D} \\ &= \sigma^2\boldsymbol{P}\boldsymbol{V}\boldsymbol{V}^{-1}\boldsymbol{D} \\ &= \sigma^2\boldsymbol{P}\boldsymbol{D} \end{aligned}$$

and

$$\begin{aligned} E\dot{G} &= E\frac{\partial}{\partial\theta}(\boldsymbol{P}(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\theta}))) \\ &= -\boldsymbol{PD}. \end{aligned}$$

Applying Theorem 2 in Section 3 of Chapter 7, G^*, the quasi-score estimating function under Wedderburn's approach, is also a quasi-score estimating function in $\mathcal{G}$ under Godambe and Heyde's approach. Thus from this point of view, using Godambe's idea to define the quasi-score estimating function will have wider application than that defined by Wedderburn's idea, because we can consider the question of varied estimating function spaces. The obvious examples to which we can apply the quasi-likelihood method are (i) the generalized linear models in which the components of the response vector $\boldsymbol{Y}$ are dependent, (ii) stochastic processes (see e.g. Hutton and Nelson(1986), Godambe and Heyde(1987) and Sørensen(1990)), if we use the Godambe and Heyde's definition and (iii) the μ_i may be just a conditional expectation.

From now on we focus our attention on Godambe and Heyde's definition of the quasi-score estimating function . This does not mean that the Wedderburn's definition is not important; it just represents a different view-point.

The main idea of the quasi-likelihood method is to choose an estimating function from an estimating function space $\mathcal{G}$. Under Godambe and Heyde's definition(1987), the criterion for choice is that an estimating function G^* in $\mathcal{G}$ is optimal iff $(E\dot{G}_T^*)^{-1}E(G_T^*G_T^{*\prime})((E\dot{G}_T^*)^{-1})'$ is the minimum element in the partial order "$\geq$" of nonnegative definite matrices(if A and B are symmetric matrices, $A \geq B$ iff $A - B$ is nonnegative definite). The following example will well explain the idea of Godambe and Heyde using the above criterion to define the quasi-score estimating function .

Suppose that $\{Y_t, t = 1, 2, \ldots, T\}$ are independent random variables and that the likelihood function is

$$L = \prod_{t=1}^{T} f_T(Y_t; \theta).$$

The score function(derivative of the log of the likelihood with respect to the parameter), which depends on a scalar parameter θ in this case, is a sum of independent random variables with zero means,

$$U = \frac{\partial logL}{\partial \theta} = \sum_{t=1}^{T} \frac{\partial log f_t(Y_t;\theta)}{\partial \theta}$$

and when $H = \sum_{t=1}^{T} b_t(\theta)(Y_t - \alpha_t(\theta))$ we have

$$E(UH) = \sum_{t=1}^{T} b_t(\theta) E(\frac{\partial log f_t(Y_t;\theta)}{\partial \theta}(Y_t - \alpha_t(\theta))).$$

If the f_t's are such that integration and differentiation can be interchanged

$$E(\frac{\partial log f_t(Y_t;\theta)}{\partial \theta} Y_t) = \frac{\partial}{\partial \theta} EY_t = \dot{\alpha}_t(\theta),$$

the dot denoting derivative, so that

$$E(UH) = \sum_{t=1}^{T} b_t(\theta)\dot{\alpha}_t(\theta).$$

Also, using *corr* to denote the correlation,

$$corr^2(U,H) = (E(UH))^2/((EU^2)(EH^2))$$

$$= 1/((EU^2)(varH^s)),$$

where $H^s = H/E(\dot{H})$, which is maximized if $varH^s$ is minimized. Following the idea of Godambe and Heyde, the choice of an optimal estimating $H^* \in \mathcal{H}$ is given as an element of $\mathcal{H}$ which has maximum squared correlation with the generally unknown score function U.

Next, for the score function U and $H \in \mathcal{H}$ we have that

$$E(H^s - U^s)^2 = varH^s + varU^s - 2E(H^sU^s)$$

$$= (EU^2)^{-1}(EU^2 varH^s - 1)$$

because of

$$U^s = U/E\dot{U} = -U/EU^2$$

and

$$EH^sU^s = -E(H^sU)/EU^2 = 1/EU^2,$$

if differentiation and integration can be interchanged. Therefore we get

$$(E(\dot{H}))^{-1}(EH^2)(E(\dot{H}))^{-1} = varH^s \geq (EU^2)^{-1} = (E\dot{U})^{-1}(EU^2)(E\dot{U})^{-1}$$

which is the Cramér-Rao inequality. In this case, following the definition of Godambe and Heyde, the score function is the optimum estimating functions in $\mathcal{H} = \{\sum_{t=1}^{T} b_t(\theta)(Y_t - \alpha_t(\theta))\}$ if $U \in \mathcal{H}$.

Moreover we can also obtain the fact that if the likelihood function has the exponential family form, then the quasi-score estimating function in a suitable estimating function space will coincide with the score function. Let us consider an example. The Brownian motions with drift provide a particular example of a family of solutions of stochastic differential equations of the form

$$dX_t = \theta a(t, X_t)dt + b(t, X_t)dW_t, \qquad X_0 = x_0.$$

Here W is a Wiener process, a and $b > 0$ are Borel functions which satisfy some suitable conditions, and $\theta \in \Theta$ where Θ is an appropriate subset of R. We suppose that the equation has an unique strong solution for every $\theta \in \Theta$.

Applying the result in Godambe and Heyde(1987) or Hutton and Nelson(1986), the quasi-score estimating function in an estimating function space $\mathcal{G} \subseteq \mathcal{H} = \{\int_0^t f_s(\theta, X_s)dW_s |\ f_s(\theta, X_s) \in \mathcal{F}_{s-}$ and $\{\mathcal{F}_s\}$ is the σ-field associated with the Wiener process $W\}$ is

$$\begin{aligned} G_t^* &= \int_0^t a(s, X_s)b^{-2}(s, X_s)b(s, X_s)dW_s \\ &= \int_0^t a(s, X_s)b^{-2}(s, X_s)dX_s - \theta \int_0^t a^2(s, X_s)b^{-2}(s, X_s)ds. \end{aligned}$$

But from Liptser and Shiryaev(1977) or Küchler and Sørensen(1989), the likelihood function for this process, whose distribution belongs to an exponential family, is

$$L_t(\theta) = exp[\theta \int_0^t a(s, X_s)b^{-2}(s, X_s)dX_s - 1/2\theta^2 \int_0^t a^2(s, X_s)b^{-2}(s, X_s)ds]$$

i.e., the score function is a constant multiple of

$$\int_0^t a(s, X_s)b^{-2}(s, X_s)dX_s - \theta \int_0^t a^2(s, X_s)b^{-2}(s, X_s)ds,$$

which is as same as the quasi-score estimating function .

From this example we conclude that, in some circumstances and for a suitable choice of estimating function space, the quasi-score estimating function in that estimating function space would just be the score function. This fact can also be confirmed after we state the relation between the quasi-score estimating function and the score function.

We can also get the relation between the quasi-likelihood estimators and the least squares estimators. In any regression model, if the variances of the errors are constants, then, after determining a suitable estimating function space, it is easy to check that the quasi-likelihood estimators are the least squares estimators(see Wedderburn(1974), McCullagh (1983)). Thus the quasi-likelihood method has a very close relationship with the ML and LS methods. But from the definition of the quasi-score estimating function we should notice that the quasi-score estimating function is strongly dependent on the estimating function space we have chosen.

We know that, for a given estimating function space $\mathcal{G}_T$, if the score function U_T belongs to $\mathcal{G}_T$, then the quasi-score estimating function G_T^* in $\mathcal{G}_T$ satisfies $G_T^* = \alpha_T U_T$ for some $p \times p$ matrix α_T which only depends on θ and T and

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

is the Fisher Information matrix since $E(G_T^*)(G_T^*)' = \alpha_T EUTU_T'\alpha_T = -E\dot{G}_T^*\alpha_T'$. If the score function U_T does not belong to $\mathcal{G}_T$ and the parameters appear in the score function U_T, Godambe and Thompson (1985) showed that G^* is a quasi-score estimating function in $\mathcal{G}_T$ iff , for some α_T depending on θ and T, $E(\alpha_T U_T - G_T^*)'(\alpha_T U_T - G_T^*) \leq E(\alpha_T - G_T)'(\alpha_T - G_T)$ for all $G_T \in \mathcal{G}_T$. The above statements tell us the relations between the quasi-score estimating function and the score function and in particular that the quasi-score estimating function is a best estimating function among the estimating function space since it has the minimum dispersion distance from the generally unknown score function.

In the above we have explained the relations between the quasi-likelihood , ML and LS methods, and also between the quasi-score estimating function and the score function. Now we consider some applications of the quasi-likelihood method. Various examples can be found in Hutton and Nelson(1986), Sørensen(1990), Desmond(1991), Heyde and Lin(1992), Liang and Zeger(1986), Prentice(1988), Morton(1989), Liang(1990), McCullagh and Nelder(1989), the references therein and this thesis. Since the quasi-likelihood method is not likelihood based and it avoids some strong restrictions on the error part of the regression equation, i.e. the independence of the errors and knowledge of the distribution of the errors, it provides a way of straightforwardly solving some complicated problems of the generalized linear model and the analysis of variance model. Here we give two examples which are very close to the generalized linear model. We choose these examples to show how to use the view-point of Godambe and Heyde. This idea is very important because it provides a simple way to find the quasi-score estimating function for many complex models.

First let us consider a model for longitudinal data, in which repeat measurements made on the same subject are usually positively correlated. Suppose that $Y_n = (y_1, y_2, \cdots, y_n)$ consists of independent vector random variables, and that $y_i = (y_{i,1}, \cdots, y_{i,k})'$ are sequences of dependent random variables. Let $E_\theta y_i = \mu_i(\theta)$, $cov_\theta(y_i) = \sigma^2 V_i(\theta)$ and the σ-field $\mathcal{F}_i = \sigma\{y_j, j \leq i\}$. Now consider an estimating function space $\mathcal{G} = \{\sum_{i=1}^n \alpha_i(y_i - \mu_i(\theta)) | \quad \alpha_i \in \mathcal{F}_{i-1}\}$. Since $\{y_i\}$ are independent, $\{y_i - \mu_i(\theta)\}$ is a martingale difference sequence with respect to $\{\mathcal{F}_i\}$. Following the result in Godambe and Heyde(1987), the quasi-score estimating function in the estimating function space is

$$G_n^* = \sum_{i=1}^n E_\theta(\frac{\partial}{\partial\theta}(y_i - \mu_i(\theta))|\mathcal{F}_{i-1})' E_\theta((y_i - \mu_i(\theta))(y_i - \mu_i(\theta))'|\mathcal{F}_{i-1})^+ (y_i - \mu_i(\theta))$$

$$= \sigma^{-2} \sum_{i=1}^n (\frac{\partial \mu_i(\theta)}{\partial\theta})' V_i^+ (y_i - \mu_i(\theta)).$$

This is the generalized estimating equation given by Liang and Zeger(1986) for modeling longitudinal data(also see Desmond(1991)). If the nuisance correlation parameters are known, then we can get the estimators of the parameters from the quasi-score equation $G_n^* = 0$ and discuss the statistical properties of the quasi-likelihood estimator θ. But in practice the nuisance correlation parameters are usually unknown. For this case, Liang and Zeger(1986) provided an asymptotic method based on the quasi-score estimating equation. They provided an iterative procedure and got a sequence of estimators, denoted by $\hat{\beta}_{G(k)}$, which is consistent under some conditions.

From the above discussion, we can see that it is possible to apply the quasi-likelihood method to more complex models than the above one, for example, where $\{y_i - \mu_i(\theta)\}$ is a martingale difference sequence with respect to some σ-fields $\{\mathcal{F}_i\}$ and not just independent random variable.

It is well known that there is a relation between regression models and ANOVA models and some ANOVA question can be solved through the regression method. Since it is developed from solving generalized linear model problems, the quasi-likelihood method should provide us with a way to deal with some complex questions concerning ANOVA models. Let us consider a model given by Morton(1987). We demonstrate the ideas by considering just three nested strata labelled 2, 1, 0 with respective subscripts, $0 \le i \le I$, $0 \le j \le J$, $0 \le k \le K$ having arbitrary ranges. Stratum 0, the bottom stratum, has scaled Poisson errors; stratum 1 introduces errors Z_{ij} and the top stratum 2 has further errors Z_i. Conditioning on the Z's gives the model in stratum 0 for the data Y:

$$E(Y_{ijk}|Z_{ij}Z_i) = \mu_{ijk}Z_{ij}Z_i,$$

$$Var(Y_{ijk}|Z_{ij}Z_i) = \phi\mu_{ijk}Z_{ij}Z_i,$$

where ϕ is an unknown scale parameter. We assume further that the Y's are conditionally uncorrelated and that the Z's are mutually independent with

$$E(Z_{ij}) = 1,\ \ E(Z_i) = 1,\ \ Var(Z_{ij}) = \sigma_1^2,\ \ Var(Z_i) = \sigma_2^2.$$

Thus the marginal expectation of Y_{ijk} is μ_{ijk}. Morton has also derived the covariance matrix V for the data vector $\boldsymbol{Y} = \{Y_{ijk}\}$. Define $\psi_1 = \sigma_1^2(\sigma_2^2+1)/\phi$ and $\psi_2 = \sigma_2^2/\phi$. Then the elements of V are

$$var(Y_{ijk}) = \phi\mu_{ijk}\{1 + (\psi_1 + \psi_2)\mu_{ijk}\},$$

$$cov(Y_{ijk}, Y_{ijn}) = \phi(\psi_1 + \psi_2)\mu_{ijk}\mu_{ijn}, \qquad (k \ne n),$$

$$cov(Y_{ijk}, Y_{imn}) = \phi\psi_2\mu_{imn} \qquad (j \ne m),$$

$$cov(Y_{ijk}, Y_{lmn}) = 0, \qquad (i \neq l).$$

The same argument as mentioned before gives

$$G^* = (\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}})' \boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})$$

as a quasi-score estimating function in the estimating function space

$$\mathcal{G} = \{\sum_i \boldsymbol{\alpha}_i(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)\},$$

where $\{\boldsymbol{\alpha}_i\}$ are nonrandom matrices, $(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$ is the data vector $\{Y_{ijk} - \mu_{ijk}\}$ written in lexicographic order,

$$(\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_I) \begin{pmatrix} \frac{\partial \boldsymbol{\mu}_1}{\partial \boldsymbol{\theta}} \\ \vdots \\ \frac{\partial \boldsymbol{\mu}_I}{\partial \boldsymbol{\theta}} \end{pmatrix} = (\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_I)\boldsymbol{D}$$

and

$$E[(\sum \boldsymbol{\alpha}_i(\boldsymbol{Y}_i - \boldsymbol{\mu}_i))(\sum \boldsymbol{\alpha}_i(\boldsymbol{Y}_i - \boldsymbol{\mu}_i))']$$
$$= (\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_I)\boldsymbol{V}(\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_I)'$$

are nonsingular, where $\boldsymbol{V}$ is the covariance matrix for $\{Y_{ijk}\}$. Since $cov(Y_{ijk}, Y_{lmn}) = 0$, $(i \neq l)$, we have

$$G^* = \sum_i (\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}})' \boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i).$$

From the quasi-score equation $G^* = 0$, we can get an estimator of the parameters if $\{ \boldsymbol{V}_i \}$ are known or after $\{ \boldsymbol{V}_i\}$ are first estimated by other methods.

In Morton's paper, he supposed that $\mu_{ijk} = exp(\boldsymbol{X}'_{ijk}\boldsymbol{\theta})$, where $\boldsymbol{X}_{ijk}$ is a vector of regressors. Under these assumptions, there are no nuisance parameters in $\{\boldsymbol{V}_i\}$ except for ϕ. If in a practical situation, we can first use other methods to estimate ϕ, then we can solve to find the quasi-likelihood estimator from the quasi-score equation

$$G^* = \sum_i (\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\theta}})' \boldsymbol{V}_i^{-1}(\boldsymbol{\theta})(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta})) = 0.$$

Sometimes we can also use the combination of quasi-score estimating functions to establish a quasi-score equation. As we will discuss in Chapter 3 in this thesis, some quasi-score estimating functions do not provide enough information about the unknown parameter. Therefore we sometimes need to combine several quasi-score estimating functions to get a new quasi-score estimating function . This method was given by Heyde in 1987. The above example of Morton provides a good illustration for the combination method. In Morton's paper, he introduced the following notations:

$$Y_{ij\cdot} = \sum_k w_{ijk} Y_{ijk}, \quad \mu_{ij\cdot} = \sum_k w_{ijk}\mu_{ijk},$$
$$r_{ijk} = Y_{ijk} - (\mu_{ijk}/\mu_{ij\cdot})Y_{ij\cdot},$$
$$w_{ij} = 1/(1+\psi_1\mu_{ij\cdot}), \quad Y_{i\cdot\cdot} = \sum_j w_{ij} Y_{ij\cdot} \quad \mu_{i\cdot\cdot} = \sum_j w_{ij}\mu_{ij\cdot},$$
$$r_{ij} = Y_{ij\cdot} - (\mu_{ij\cdot}/\mu_{i\cdot\cdot})Y_{i\cdot\cdot}$$

and

$$r_i = Y_{i\cdot\cdot} - \mu_{i\cdot\cdot}.$$

Following those notations, we can construct three mutually orthogonal estimating function spaces

$$\mathcal{G}_1 = \{\sum_i A_i R_{1,i}\},$$
$$\mathcal{G}_2 = \{\sum_i B_i R_{2,i}\}$$

and

$$\mathcal{G}_3 = \{\sum_i C_i r_i\},$$

where $\{A_i\}$ and $\{B_i\}$ are nonrandom matrices, $\{C_i\}$ are nonrandom coefficients, $R_{1,i} =(r_{i11},\ldots,r_{iJK})'$, $R_{2,i} =(r_{i1},\ldots,r_{iJ})'$ and the elements $\{G\}$ in those estimating function spaces all satisfy the conditions that EGG' and $E\dot{G}$ are nonsingular. Then three orthogonal quasi-score estimating functions can be obtained from those estimating function spaces respectively and , after combining them, we achieve a quasi-score estimating function

$$\sum_{ijk}\frac{r_{ijk}}{\mu_{ijk}}\frac{\partial\mu_{ijk}}{\partial\theta} + \sum_{ij}\frac{r_{ij}}{\mu_{ij}(1+\psi_1\mu_{ij\cdot})}\frac{\partial\mu_{ij\cdot}}{\partial\theta} + \sum_i \frac{r_i}{\mu_{i\cdot\cdot}(1+\psi_2\mu_{i\cdot\cdot})}\frac{\partial\mu_{i\cdot\cdot}}{\partial\theta},$$

which is the quasi-score estimating function Morton used in his paper to obtain the quasi-likelihood estimators. Morton's derivation was quite different.

Sometimes the quasi-score equations may contain nuisance parameters, e.g. the longitudinal data example and an example in Lloyd-Smith(1990), or sometimes the forms of the quasi-score equations are complex, for example nonlinear with respect to the parameters. In those cases, it is may be very difficult to get the quasi-score estimators from the quasi-score estimating equations. This problem is also met with the ML and LS methods. For the quasi-likelihood method, there are several ways to obtain the quasi-likelihood estimators from a quasi-score estimating equation. This is just a question of finding the root of a (random) equation. Since many equations have no analytic solution, numerical methods become the usual way to obtain the quasi-likelihood estimators from quasi-score estimating equation. Many statisticians have used iterative methods, stochastic approximation and others numerical techniques(see Szabłowski(1988), Osborne(1992)). The close relationship between the quasi-likelihood method, nonlinear regression and generalized linear regression means that it is possible to borrow Fisher's method of scoring, Newton's method and other related methods(see Seber and Wild(1989), Bates and Watts(1988)), which are used for nonlinear regression and generalized linear regression, for the quasi-likelihood method. If the quasi-score equation contains only one parameter, it will sometimes be easy to get the quasi-likelihood estimator from the estimating equation. Otherwise, the procedure to get the quasi-likelihood estimators is always complex. The usual steps for a numerical method is to obtain a sequence of asymptotic roots of the equations, denoted by $\{\hat{\theta}_n\}$. Among the technical and theoretical problems met by a numerical method are how to find the conditions under which $\{\hat{\theta}_n\}$ converges and how to make the convergence fast.

Sometimes, when the quasi-score equations contain nuisance parameters, the asymptotic quasi-likelihood method is a useful method to obtain quasi-likelihood estimators(see, Heyde and Gay(1989), Lloyd-Smith(1990) and Chapter 5 in this thesis). Anyway, although there may be a problem to obtain quasi-likelihood estimators in some cases, the quasi-likelihood method can at least indicate the optimal estimating function in a given estimating function space. It is important for a statistician to have this choice.

Asymptotic inference for stochastic processes is an important subject. Many papers in this field can be found from Basawa and Prakasa Rao(1980), Basawa and Scott(1983) and the reference therein. However, much of this work is likelihood-based and it is possible to remove those assumptions based on likelihoods in some cases. As Basawa(1991) has mentioned,* the quasi-likelihood method is available for this purpose. Although the criterion for a quasi-score estimating function does not rely on asymptotic properties and the quasi-likelihood method is used for finite samples, increasing the sample size provides a sequence of quasi-likelihood estimators. In most applications of the quasi-likelihood method it turns out that these estimators indeed exhibit good sampling properties asymptotically. Chapters 3 and 4 in this thesis discuss this question.

3 Summary of thesis

As we have mentioned before, in a certain sense the quasi-score estimating function has minimum dispersion distance from the generally unknown score function and, associated with an estimating function space, the quasi-score estimating function is a best estimating function in that estimating function space in the above sense. There is also another judgement in favour of the quasi-score estimating function as a best estimating function in the associated estimating function space. Corresponding to the concepts of sufficiency and ancillary in classical statistics, McLeish and Small(1988) introduced the concepts of E-sufficient, locally E-sufficient, E-ancillary and locally E-ancillary and they showed that the best way, in a certain sense, to choose an estimating function from an estimating function space for inference is to choose an estimating function which is E-sufficient or locally E-sufficient. In Chapter 2, we explore the relationships between the quasi-score estimating function and E-sufficient or locally E-sufficient estimating functions and point out that, under mild regularity conditions, the quasi-score estimating function is locally E-sufficient. Several sufficient conditions for the regularity conditions are given and some examples of quasi-score estimating functions which are E-sufficient or locally E-sufficient are

* and others, e.g. Hutton and Nelson (1986), Heyde and Gay (1989), 1992
Sorenson (1990), have done.

presented.

The consistency of quasi-likelihood estimators is discussed in Chapter 3. It is obvious that the consistency of the quasi-likelihood estimators depends strongly on how much information we can get from the data or the process. In process inference problems, the usual way to establish an estimating function space associated with the inference problem we consider is to choose a martingale from the semimartingale representation of the given process. As we point out in Chapter 3, sometimes the quasi-likelihood estimators may not be consistent if the martingale from the given process does not contain information about the parameters which is unbounded as the sample size increases. How can we get more information from a given process or sample and when do the martingales contain all the information we can get from the sample? Those questions are discussed in Chapter 3. Furthermore, we also give several sufficient conditions for the consistency of the quasi-likelihood estimators. The discussions are based on two different approaches; one concerns the integral surface case and the other the asymptotic linear case.

Score estimating functions can always provide smaller asymptotic confidence zones for the true parameters than other estimating functions . We are concerned with the question of whether the quasi-score estimating function can maintain this property in the associated estimating function space. This subject is discussed in Chapter 4 where a natural generalization of the classical result for the score function is provided. Several examples will be given in Chapter 6 to support this result.

Sometimes it is not easy to get a quasi-likelihood estimator from a quasi-score equation, especially when the quasi-score estimating function contains nuisance parameters. In this case, we need a new estimating function to replace the quasi-score estimating function . Heyde and Gay(1989) introduced a new concept of asymptotic quasi-likelihood estimating functions to solve the problem in the large sample case. However their definition is not satisfactory in that it will not always provide minimum size asymptotic confidence zones for the true parameter. Therefore, in Chapter 5 a new definition of asymptotic quasi-score estimating function is given. Also, the concept " asymptotic locally E-sufficient " is introduced and the relationships between asymptotic quasi-score estimating functions and asymptotic locally E-sufficient estimating functions discussed.

Consider a model

$$X_t = F_t(\theta) + M_t(\theta), \qquad t \in \mathcal{T}$$

where $\mathcal{T}$ is an index set which may not be discrete. In the model, $F_t(\theta)$ usually serves as a prediction function, or fitted function, which contains parameters and previous information(if considering the inference of processes); $M_t(\theta)$ usually serves as prediction error or measurement error and , therefore, we call $\{M_t(\theta)\}$ the error part of the model. If $\{M_t\}$ is a quite general process, for example, $\{M_t\}$ is not a independent increment process and time is not discrete time, it is impossible to use the least squares or weighted least squares method to fit the above model. However, if M is a martingale and the quadratic variation process $\{< M >_t\}$ of $\{M_t\}$ is known, then the quasi-likelihood method enables us to fit the model. Thus, we can say that the quasi-likelihood method makes a very important contribution to the theory of inference. But if M is a semimartingale process, can we apply the quasi-likelihood method framework to this case? Furthermore, a key step to the use of the quasi-likelihood method is to write down the quasi-score estimating function from a given estimating function space. So, given an estimating function space, how can we write down the quasi-score estimating function in the space and what is the formula for the quasi-score estimating function ? All of the above questions are discussed in Chapter 7. For the sake of completeness we also provide several strong law of large numbers theorems in Chapter 8, which usually can be used as to establish the strong consistency of quasi-likelihood estimators in various contexts.

CHAPTER 2
THE RELATIONS BETWEEN QUASI-SCORE ESTIMATING FUNCTIONS AND LOCALLY E-SUFFICIENT ESTIMATING FUNCTIONS

1 Introduction

This chapter is concerned with the relations between quasi-score estimating functions and locally E-sufficient estimating functions. As we have mentioned in Chapter 1, Godambe and Thompson(1985) indicated that, under minor regularity conditions, for a given estimating function space Ψ, the quasi-score estimating function in Ψ is an estimating function which has a minimum distance from the score function compared with other estimating functions in Ψ. From this point of view, the quasi-score estimating function is an optimum estimating function to choose from Ψ. Here we shall discuss the same question from a different point of view.

In classical statistics, there is an important criterion for choosing statistics for parameter inference, which is the concept of "sufficiency" or the dual concept of "ancillarity". A sufficient statistic is a statistic such that the conditional distribution of the original observations given that statistic does not depend on any unknown constants. The concept of sufficiency was first defined in Fisher(1922). Fisher (1925) said that the sufficient statistic " is equivalent, for all subsequent purpose of estimation, to the original data from which it was derived." Therefore, good estimators should depend on the original observations only through a sufficient statistic. Later this principle was extended to say that any inference (estimators, tests, confidence intervals, etc.) should be based only on the sufficient statistic. This is often called the sufficiency principle and is accepted by most statisticians. An ancillary statistic is one whose distribution is insensitive to change in the parameter and under some conditions the complete sufficient statistics form the orthogonal complement to the space of ancillary statistics(see McLeish and Small(1988) pp15-16). McLeish and Small(1988) pointed out that, in the theory of estimating function or, as they called it, the theory of inference functions, changes in parameters are first evident through changes in the expectation of the estimating functions . From this perspective, the classical concepts " ancilla~~ry~~ity " and " sufficiency " motivated McLeish and Small to put forward new concepts of E-sufficiency, E-ancilla~~ry~~ity, local E-sufficiency and local E-ancilla~~ry~~ity for estimating function . Following their definitions which will be given later, the E-ancillary estimating function $\psi(\theta(P))$ is an estimating function such that $E_Q\psi(\theta(P)) = 0$ for all Q in $\mathcal{P}$; the locally E-ancillary estimating function $\psi(\theta(P))$ is an estimating function such that $E_P\psi(\theta(P_0)) = o(\theta(P) - \theta(P_0))$, as $\theta(P) \to \theta(P_0)$. Therefore in a certain sense, an E-ancillary estimating function or a locally E-ancillary estimating function is insensitive to change in the parameters. By extending the classical concepts, we say that an estimating function which belongs to the orthogonal ~~set~~ to that of E-ancillary or locally E-ancillary estimating functions is "sensitive" to the parameters. From this point of view, McLeish and Small argued that, in the theory of estimating functions , for the choice of an optimum estimating function from an estimating function space, it is better to choose one from the E-sufficient estimating function subset or the locally E-sufficient estimating function subset of Ψ if this is possible, because the estimating functions in those subsets are "sensitive" to parameters in the above sense. Therefore, discussing the relations between quasi-score estimating functions and E-sufficient estimating functions is an important subject. The discussion can show us whether the quasi-score estimating function is a good estimating function in McLeish and Small's framework and provide us with an easy way to write down the E-sufficient estimating function or the locally E-sufficient estimating function if the quasi-score estimating function is E-sufficient or locally E-sufficient.

In some particular examples, McLeish and Small(1988) showed that the E-sufficient estimating function and locally E-sufficient estimating function are also quasi-score estimating functions . However, they did not touch the questions of why these two different concepts of estimating function coincide in these examples and when they have this property. Here we shall discuss this subject in detail.

We shall show that whether a quasi-score estimating function is locally E-sufficient depends strongly on what kind of estimating function space we choose. We shall also show that, under regularity conditions and some additional conditions, the quasi-score estimating function can be locally E-sufficient. Several examples will be given to explain this fact.

2 Definitions and notations

Let $\mathcal{X}$ be a sample space and $\mathcal{P}$ be a class of probability measures P on $\mathcal{X}$. For each P, we denote by v_P the vector space of vector valued functions f defined on the sample space X such that $E_P\|f\|^2 < \infty$, where

$$\|f\|^2 = (f_1(x), \dots, f_n(x)) \begin{pmatrix} f_1(x) \\ \dots \\ f_n(x) \end{pmatrix} = f'(x)f(x).$$

If f and g belong to v_P, the inner product on v_P is defined by

$$< f, g >_P = E_P(f'(x)g(x)).$$

Let θ be a real valued function on the class of probability measures $\mathcal{P}$ and denote by $\Theta = \{\theta(P), P \in \mathcal{P}\} \subseteq R^n$ the parameter space. Notice that θ need not be a one to one function. If it is , we call the model a one-parameter model.

Now we consider a space Ψ in which every estimating function $\psi(\theta, x) = \psi(\theta)$ is a mapping from the parameter space and the sample space into R^n. Each element of Ψ is unbiased and square integrable, i.e.

$$E_P[\psi(\theta(P))] = 0,$$

and

$$\|\psi\|^2_{\theta(P)} = E_P[\psi'(\theta(P))\psi(\theta(P))] < \infty, \tag{1}$$

for all P and $\theta = \theta(P)$. The space is chosen to have constant covariant structure so that the inner product $E_P[\psi'(\theta(P))\phi(\theta(P))]$ depends on P only through $\theta(P)$, where $\psi, \phi \in \Psi$. As usual, we also require that Ψ be weak square closed, i.e. Ψ is closed in the topology induced by the norm (1) and Ψ is a Hilbert space.

Definition 1 *If an unbiased estimating function $\phi(\theta)$ is a weak square limit of functions $\{\psi\} \in \Psi$ which satisfy*

$$E_P[\psi(\theta)] = o[|\theta(P) - \theta|], \qquad \text{as } \theta(P) \to \theta, \tag{2}$$

we say that $\phi(\theta)$ is locally E-ancillary.

If differentiation and integration can be exchanged, (2) becomes

$$E_\theta(\frac{\partial \psi_i}{\partial \theta_j})_{n \times n} = 0.$$

Therefore, under suitable regularity conditions, Definition 1 can be replaced by Definition 2.

Definition 2 *If an unbiased estimating function $\phi(\theta)$ is a weak square limit of functions $\{\psi\} \in \Psi$ which satisfy*

$$E_\theta(\frac{\partial \psi_i}{\partial \theta_j})_{n \times n} = 0,$$

we say that $\phi(\theta)$ is locally E-ancillary.

We denote the subset of locally E-ancillary estimating functions by $\mathcal{A}_{loc}$.

Definition 3 *The subset $\mathcal{S}_{loc}$ is said to be complete locally E-sufficient* iff*, for* every *$\psi \in \mathcal{S}_{loc}$,* ~~iff~~

$$E_\theta(\psi(\theta)\phi(\theta)') = 0_{n \times n}$$

for all P, $\theta = \theta(P)$ and $\phi \in \mathcal{A}_{loc}$.

Following the definition, $\mathcal{S}_{loc} = \mathcal{A}_{loc}^{\perp}$. From the assumptions on Ψ, we have

$$\Psi = \mathcal{S}_{loc} \oplus \mathcal{A}_{loc}$$

(see Berberian(1961)).

Definition 4 *A subset L is said to be locally E-sufficient if $\phi \in \Psi$ such that $< \phi, \psi >_\theta = 0$ for all $\psi \in L$ and $\theta \in \Theta$ is locally E-ancillary.*

In the following, let $\{X_t, 0 \leq t \leq T\}$ be a process whose distribution depends on a parameter θ taking values in a subset $\Theta \in R^n$ and Ψ is a Hilbert space of unbiased functions. Assume that any pair elements of Ψ has constant covariant structure and Ψ satisfies the usual regularity conditions. Therefore we use Definition 2 to define the concept of local E-ancillarity.

Let $\mathcal{G}_1$ be a subset of Ψ of estimating functions G which satisfy the conditions that $E\dot{G}(\theta) = E(\frac{\partial G_i}{\partial \theta_j})_{n \times n}$ and $EG(\theta)G'(\theta)$ are nonsingular. In the following, without further explanation, the " $\cdot$ " will denote differentiation with respect to θ.

If P_θ is absolutely continuous with respect to some σ-finite measure λ and has density $p(\theta)$, we denote the score function by $U(\theta) = p^{-1}(\theta)\dot{p}(\theta)$ if it exists and in this case we also suppose that U is differentiable with respect to θ.

Definition 5 *G^* is a quasi-score estimating function in $\mathcal{G}_1$ if*

$$(E\dot{G})^{-1}(EGG')((E\dot{G})^{-1})' - (E\dot{G}^*)^{-1}(EG^*G^{*\prime})((E\dot{G}^*)^{-1})'$$

is nonnegative-definite, for all $G \in \mathcal{G}_1$ and $\theta \in \Theta$.

If the score function exists and it is a function of the parameter, there are two other equivalent definitions (see, Godambe and Heyde(1987)).

Definition 6 *G^* is a quasi-score estimating function in $\mathcal{G}_1$ if*

$$E((\alpha U - G^*)G') = E(G(\alpha U - G^*)') = 0,$$

for all $G \in \mathcal{G}_1$ and $\theta \in \Theta$, where α is a fixed matrix depending on θ.

Definition 7 *G^* is a quasi-score estimating function in $\mathcal{G}_1$ if*

$$E(\alpha U - G)(\alpha U - G)' - E(\alpha U - G^*)(\alpha U - G^*)'$$

is nonnegative-definite for all $G \in \mathcal{G}_1$ and $\theta \in \Theta$, where α is a fixed matrix depending on θ.

If $\mathcal{G}_1$ is a convex set, we also have the following theorem (Heyde(1988)).

__Theorem__ **1** *Suppose that $\mathcal{G}_1$ is a convex set. Then, $G^* \in \mathcal{G}_1$ is a quasi-score estimating function in $\mathcal{G}_1$ iff*

$$(E\dot{G})^{-1}(EGG^{*\prime}) = (E\dot{G}^*)^{-1}(EG^*G^{*\prime})$$

for all $G \in \mathcal{G}_1$.

3 Results

Since $\Psi = \mathcal{S}_{loc} \oplus \mathcal{A}_{loc}$, we first investigate the circumstances under which the quasi-score estimating function is in $\mathcal{S}_{loc}$ or $\mathcal{A}_{loc}$. Denote by G^* the quasi-score estimating function in $\mathcal{G}_1$ and write $G = 0$ if $G = 0$, a.s. .

__Theorem__ **2** *If $\mathcal{G}_1$ is a convex set and the quasi-score estimating function $G^* \in \mathcal{G}_1 \cap \mathcal{A}_{loc}$ (resp. $G^* \in \mathcal{G}_1 \cap \mathcal{S}_{loc}$), then $\mathcal{G}_1 \cap \mathcal{S}_{loc} = \emptyset$ (resp. $\mathcal{G}_1 \cap \mathcal{A}_{loc} = \emptyset$).*

Proof: Only the proof of the first part of the theorem is given here. The proof of the second part is similar.

If $\mathcal{G}_1 \cap \mathcal{S}_{loc} \neq \emptyset$, there is a $G \in \mathcal{G}_1 \cap \mathcal{S}_{loc}$ and, from Theorem 1,

$$(E\dot{G})^{-1}(EGG^{*\prime}) = (E\dot{G}^*)^{-1}(EG^*G^{*\prime}).$$

But $\mathcal{S}_{loc} = \mathcal{A}_{loc}^{\perp}$, therefore $EGG^{*\prime} = 0_{n\times n}$ and

$$EG^*G^{*\prime} = 0.$$

This forces $G^* = 0$ and gives a contradiction. So $\mathcal{G}_1 \cap \mathcal{S}_{loc} = \emptyset$.

Remark : Even if $\mathcal{G}_1$ is not a convex set, the second part of Theorem 2 still holds if the score function U exists. In fact, by Definition 4, G^* is a quasi-score estimating function iff, for some fixed matrix function α depending on θ,

$$E((\alpha U - G^*)G') = E(G(\alpha U - G^*)') = 0 \tag{3}$$

for all $G \in \mathcal{G}_1$ and $\theta \in \Theta$. If there is a $G_\circ \in \mathcal{G}_1 \cap \mathcal{A}_{loc}$, we obtain $\alpha = 0$ since $EUG_\circ{}' = E\dot{G}_\circ$ is nonsingular and $EG^*G_\circ{}' = 0$. However, (3) is true for all $G \in \mathcal{G}_1$, therefore we can choose $G = G^*$ and obtain $EG^*G^{*\prime} = 0$. This gives the contradiction $G^* = 0$. Therefore $\mathcal{G}_1 \cap \mathcal{A}_{loc} = \emptyset$.

From Theorem 2 we notice that, if there are no any other restrictions on $\mathcal{G}_1$ to reduce its size, it is possible that G^* is outside of $\mathcal{S}_{loc}$. But in general, under regularity conditions, we always have $\mathcal{G}_1 \cap \mathcal{A}_{loc} = \emptyset$. That means that G^* can only be in the complement set of $\mathcal{A}_{loc}$.

As we have mentioned before, we are very interested in the conditions under which $G^* \in \mathcal{S}_{loc}$. This is taken up in the following theorem.

Theorem 3 *Suppose that $\mathcal{G}_1$ is a convex set and $G^* \in \mathcal{G}_1 - \mathcal{G}_1 \cap \mathcal{A}_{loc}$. Let G^*_{ps} be the projection of G^* on $\mathcal{S}_{loc}$. If $G^*_{ps} \in \mathcal{G}_1$ and for any G with $E\dot{G} = 0$, we have $G^*_{ps} + G \in \mathcal{G}_1$, then $G^* \in \mathcal{S}_{loc}$.*

Proof: Assume that $G^* = G^*_{ps} + G^*_{pa}$, where $G^*_{pa} \in \mathcal{A}_{loc}$, and $G^*_{ps} \in \mathcal{S}_{loc}$.

Since $G^*_{pa} \in \mathcal{A}_{loc}$, there is a sequence $G_{pa,n} \in \mathcal{A}_{loc}$ such that $E\dot{G}_{pa,n} = 0$ and $E(G^*_{pa} - G_{pa,n})'(G^*_{pa} - G_{pa,n}) \to 0$, as $n \to \infty$.

Let $G_n = G^*_{ps} + G_{pa,n}$. By assumption we have $G_n \in \mathcal{G}_1$ and

$$(E\dot{G}_n)^{-1}(EG_nG^{*\prime}) = (E\dot{G}^*)^{-1}(EG^*G^{*\prime}),$$

i.e.

$$\begin{gathered}(E\dot{G}^*_{ps})^{-1}(EG^*_{ps}G^*_{ps}{}' + EG_{pa,n}G^*_{pa}{}') = \\ (E\dot{G}^*_{ps} + E\dot{G}^*_{pa})^{-1}(EG^*_{ps}G^*_{ps}{}' + EG^*_{pa}G^*_{pa}{}'). \end{gathered} \tag{4}$$

But

$$\begin{gathered}\|E(G_{pa,n}G^*_{pa}{}' - G^*_{pa}G^*_{pa}{}')\| = \|E((G_{pa,n} - G^*_{pa})G^*_{pa}{}')\| \\ \leq \|E((G_{pa,n} - G^*_{pa})\|\|E(G^*_{pa}{}')\| \to 0, \qquad \text{as } n \to \infty.\end{gathered}$$

Let $n \to \infty$ in (4); we obtain

$$(E\dot{G}^*_{ps})^{-1}(EG^*_{ps}G^*_{ps}{}' + EG^*_{pa}G^*_{pa}{}') = (E\dot{G}^*_{ps} + E\dot{G}^*_{pa})^{-1}(EG^*_{ps}G^*_{ps}{}' + EG^*_{pa}G^*_{pa}{}')$$

and therefore $E\dot{G}^*_{pa} = 0$. Using Theorem 1 again, then

$$(E\dot{G}^*_{ps})^{-1}(EG^*_{ps}G^{*\prime}) = (E\dot{G}^*)^{-1}(EG^*_{ps}G^*_{ps}{}' + EG^*_{pa}G^*_{pa}{}')$$

$$= (E\dot{G}^*_{ps})^{-1}(EG^*_{ps}G^*_{ps}{}' + EG^*_{pa}G^*_{pa}{}'),$$

i.e.

$$(E\dot{G}^*_{ps})^{-1}(EG^*_{ps}G^*_{ps}{}') = (E\dot{G}^*_{ps})^{-1}(EG^*_{ps}G^*_{ps}{}' + EG^*_{pa}G^*_{pa}{}').$$

Therefore $EG^*_{pa}G^*_{pa}{}' = 0$ and $G^*_{pa} = 0$. This gives

$$G^* \in \mathcal{G}_1 \cap \mathcal{S}_{loc}.$$

Theorem 4 *Suppose that $\mathcal{G}_1$ is a convex set. Let $G^* = G^*_{ps} + G^*_{pa}$, where $G^*_{ps} \in \mathcal{S}_{loc} \cap \mathcal{G}_1$ and $G^*_{pa} \in \mathcal{A}_{loc}$. If, for all θ, $E_\theta \dot{G}^* = E_\theta \dot{G}^*_{ps}$, then $G^* \in \mathcal{S}_{loc}$.*

The proof is same as the last part of the proof of Theorem 4.

In the above theorems, we do not consider whether U exits or not. However, if U exists, in 1-parameter case and under certain regular conditions, $U \in \mathcal{S}_{loc}$ and $\mathcal{S}_{loc} = \sigma(U)$ which is a subspace generated by U(see, McLeish and Small, 1988, p29). Following Definition 7, we should obtain $G^* \in \mathcal{S}_{loc}$ if $\mathcal{S}_{loc} \cap \mathcal{G}_1 \neq \emptyset$.

In order to obtain an important property for quasi-score estimating functions we need some new notations and definitions.

Let $\tilde{\mathcal{G}}_1 = \{G|\ E\dot{G}$ is nonsingular and $G \in \Psi\}$. Therefore $\tilde{\mathcal{G}}_1 \supseteq \mathcal{G}_1$.

Definition 8 *G^*_T is a sub-quasi-score estimating function in $\tilde{\mathcal{G}}_1$ if for all $G_T \in \tilde{\mathcal{G}}_1$, and all $\theta \in \Theta$,*

$$(E\dot{G}_T)^{-1}(EG_TG_T{}')((E\dot{G}_T)')^{-1} - (E\dot{G}^*_T)^{-1}(EG^*_TG^*_T{}')((E\dot{G}^*_T)')^{-1} \geq 0.$$

Remark : Following the above Definition, a sub-quasi-score estimating function G^*_T is also a quasi-score estimating function in $\mathcal{G}_1$ if $G^*_T \in \mathcal{G}_1$, but the converse is not necessarily true. However, if we restrict our consideration to the model

$$X_t = \int_0^t f_s d\lambda_s + m_s,$$

where $\{m_s\}$ is a martingale with respect to some σ-fields $\{\mathcal{F}_t\}$, $\{\lambda_s\}$ is an increasing function and we let $\Psi_T = \{\int_0^T \alpha_s(\theta)dm_s \mid \alpha_s(\theta) \in \mathcal{F}_s\}$, then $\mathcal{G}_1 \subseteq \tilde{\mathcal{G}}_1 \subseteq \Psi_T$ and following the proof in Godambe and Heyde(1987, p238), we can obtain that the quasi-score estimating function G^* is a sub-quasi-score estimating function in $\tilde{\mathcal{G}}_1$.

If we restrict Ψ such that the subset $\mathcal{A}_{loc}$ of Ψ satisfies :

Assumption 1 :

$$G \in \mathcal{A}_{loc} \quad iff \quad E_\theta \dot{G} = 0 \quad \text{for all } \theta,$$

then we have the following theorem.

Theorem 5 *Let G^* be a sub-quasi-score estimating function on $\tilde{\mathcal{G}}_1 \subseteq \Psi$. Then, under Assumption 1, $G^* \in \mathcal{S}_{loc}$.*

Proof: Represent G^* by $G^* = G^*_{ps} + G^*_{pa}$, where $G^*_{ps} \in \mathcal{S}_{loc}$ and $G^*_{pa} \in \mathcal{A}_{loc}$.

Since $G \in \mathcal{A}_{loc}$ iff $E_\theta \dot{G} = 0$, for all $\theta \in \Theta$, and $G^* \in \tilde{\mathcal{G}}_1$, we obtain that $E_\theta \dot{G}^*_{ps} = E_\theta \dot{G}^*$ is nonsingular for all $\theta \in \Theta$. Using Definition 8,

$$(E\dot{G}^*_{ps})^{-1}(EG^*_{ps}G^{*\prime}_{ps})(E\dot{G}^*_{ps})^{-1} - (E\dot{G}^*)^{-1}(EG^*G^*)(E\dot{G}^*)^{-1},$$

is nonnegative-definite, i.e.

$$(E\dot{G}^*_{ps})^{-1}(EG^*_{ps}G^{*\prime}_{ps})(E\dot{G}^*_{ps})^{-1} - (E\dot{G}^*_{ps})^{-1}(EG^*_{ps}G^*_{ps} + EG^*_{pa}G^*_{pa})(E\dot{G}^*_{ps})^{-1},$$

is nonnegative-definite. Therefore

$$\begin{aligned} EG^*_{ps}G^{*\prime}_{ps} - (EG^*_{ps}G^*_{ps} + EG^*_{pa}G^*_{pa}) \\ = -EG^*_{pa}G^{*\prime}_{pa} \end{aligned}$$

is nonnegative-definite. Then we obtain $EG^*_{pa}G^{*\prime}_{pa} = 0$ and $G^* \in \mathcal{S}_{loc}$.

Before an example is given for Theorem 5, we need the following additional definitions.

Definition 9 *An unbiased estimating function $\psi \in \Phi$ is said to be E-ancillary if ψ can be written as the weak square limit of functions ϕ such that*

$$E_Q\phi(\theta(P)) = 0$$

for all $Q \in \mathcal{P}$.

We let $\mathcal{A}$ be the set of E-ancillary estimating functions in Φ. By construction it is a closed linear subspace of Φ.

From Definition 9, we have $\mathcal{A}_{loc} \supseteq \mathcal{A}$. Let $\underline{A} = \{\psi | \ \psi \in \Psi$ and $E_\theta(\dot{\psi}) = 0$ for all $\theta \in \Theta\}$. Then if $\underline{A} \subseteq \mathcal{A}$ it will force $\mathcal{A}_{loc} = \mathcal{A}$.

Definition 10 *A subset $\mathcal{S}$ of Ψ is said to be a complete E-sufficient subset if (A) and (B) below are equivalent for all $\phi \in \Psi$:*

(A) $< \psi, \phi >_\theta = 0$ for all $\theta \in \Theta$, $\psi \in \mathcal{S}$,

(B) $\phi \in \mathcal{A}$.

Following Definition 10, $\mathcal{S} \supseteq \mathcal{S}_{loc}$. Therefore if $\mathcal{A} = \mathcal{A}_{loc}$ then it will induce $\mathcal{S} = \mathcal{S}_{loc}$.

Example 1 Consider a model

$$X_t = \lambda(\theta) \int_0^t a_s d < M >_s + M_t$$

for a non-random function $\lambda(\theta)$, such that $\lambda(\eta) \neq \lambda(\theta)$ for $\theta \neq \eta$ and $\dot{\lambda}(\theta) \neq 0$. Here $\{a_t\}$ is a predictable process which does not depend on the parameter θ and $\{M_t\}$ is a square integrable martingale . We assume that the predictable process $\{a_t\}$ is observable and denote by $\{< M >_t\}$ the predictable variation process which we also assume observable and not dependent on θ.

Let Ψ be a space of all linear estimating function of the form

$$\psi(\theta) = \int_0^T H_t dM_t(\theta).$$

Here $\{H_t\}$ is a predictable process which is free from the parameter and is such that the resulting estimating function is a square integrable martingale.

Since

$$0 = E_\theta \psi(\theta) = E_\theta \int_0^T H_t dM_t(\theta) = E_\theta \int_0^T H_t dX_t - \lambda(\theta) E_\theta \int_0^T H_t a_t d < M >_t,$$

then

$$E_\theta \psi(\eta) = E_\theta \int_0^T H_t dX_t - \lambda(\eta) E_\theta \int_0^T H_t a_t d < M >_t$$

$$= (\lambda(\theta) - \lambda(\eta))E_\theta \int_0^T H_t a_t d < M >_t = 0,$$

for all θ iff

$$E_\theta \int_0^T H_t a_t d < M >_t = 0, \qquad \text{for all } \theta.$$

Furthermore, since $E_\theta \dot{\psi}(\theta) = -\dot{\lambda}(\theta) E_\theta \int_0^T H_t a_t d < M >_t$ and $\dot{\lambda}(\theta) \neq 0$, $E_\theta \psi(\eta) = 0$ for all θ iff $E_\theta \dot{\psi}(\theta) = 0$. Therefore $\underline{A} \subseteq \mathcal{A}$ and induces $\mathcal{A}_{loc} = \mathcal{A}$.

Suppose that $\psi(\theta) = \int_0^T H_t dM_t(\theta)$ is a weak square limit of functions ψ_n such that

$$E_\eta \psi_n(\theta) = 0, \qquad \text{for all } \eta \text{ and } \theta.$$

We obtain

$$|E_\eta \psi(\theta)| = |E_\eta(\psi(\theta) - \psi_n(\theta))| \leq [E_\eta(\psi(\theta) - \psi_n(\theta))^2]^{1/2}$$

$$= [E_\eta(\int_0^T (H_t - H_{t,n}) dM_t(\theta))^2]^{1/2}$$

$$= [E_\eta \int_0^T (H_t - H_{t,n})^2 d < M >_t]^{1/2}$$

$$= [E_\eta(\psi(\eta) - \psi_n(\eta))^2]^{1/2} \to 0, \qquad \text{as } n \to \infty.$$

Therefore $E_\eta(\psi(\theta)) = 0$ for all η. This means that $\psi \in \mathcal{A}$ iff $E_\eta \psi(\theta) = 0$ for all θ. It also means that $\psi \in \mathcal{A}_{loc}$ iff $E_\theta \dot{\psi}(\theta) = 0$ for all θ. Following Theorem 5, the quasi-score estimating function

$$G^* = \dot{\lambda}(\theta) \int_0^t a_s dM_s \in \mathcal{S}_{loc} = \mathcal{S}.$$

Consider a stochastic process $\{X_t\}$ which satisfies a stochastic differential equation of the form

$$dX_t = a_t(\theta) d < M(\theta) >_t + dM_t(\theta) \tag{5}$$

where $\{M_t\}$ is a square integrable martingale with predictable variation process $< M >_t$ observable up to the value of the parameter θ. Also assume that the predictable process $a_t(\theta)$ is observable up to the parameter value and there exists a real-valued predictable process $f(t; \eta, \theta)$ such that

$$\int_0^s a_t(\eta) d < M(\eta) >_t = \int_0^s f(t; \eta, \theta) d < M(\theta) >_t,$$

for all $0 < s < T$ and all η sufficiently close to θ. Notice that $f(t;\theta,\theta) = a_t(\theta)$. Now restricting our consideration to a space of estimating functions of the form $\psi(\theta) = \int_0^T H_s(\theta)dM_s(\theta)$, we obtain the following proposition.

Proposition 1 *Under the conditions assumed above for the model (5), we have*

$$\mathcal{A}_{loc} = \{\phi : E_\theta\dot{\phi} = 0, \quad \textit{for all } \theta \in \Theta\}.$$

Proof: First we observe that for any function ψ and parameter η and θ, we have

$$E_\eta\psi(\theta) = E_\eta\{\int_0^T H_t(\theta)[dM_t(\eta) + a_t(\eta)d < M(\eta) >_t - a_t(\theta)d < M(\theta) >_t]\}$$

$$= E_\eta\{\int_0^T H_t(\theta)[f(t;\eta,\theta) - f(t;\theta,\theta)]d < M(\theta) >_t\}.$$

Dividing by $\eta - \theta$ and letting $\eta \to \theta$, we obtain

$$E_\theta\dot{\psi}(\theta) = \frac{\partial}{\partial\eta}E_\eta\psi(\theta)|_{\eta=\theta} = E_\theta\int_0^T H_t(\theta)[\frac{\partial}{\partial\theta}f(t;\theta,\theta)]d < M(\theta) >_t .$$

For any $\psi_o \in \mathcal{A}_{loc}$, if $\psi_o(\theta) = \int_0^T H_{t,0}(\theta)dM_t(\theta)$ is a weak square limit of functions $\{\psi_n = \int_0^T H_{t,n}dM_t(\theta)\}$ and

$$E_\theta\dot{\psi}_n = E_\theta\int_0^T H_t(\theta)[\frac{\partial}{\partial\theta}f(t;\theta,\theta)d < M(\theta) >_t] = 0,$$

for all ψ_n. Then,

$$|E_\theta\dot{\psi}_o(\theta)| = |E_\theta(\dot{\psi}(\theta) - \dot{\psi}_n(\theta))|$$

$$= |E_\theta\int_0^T(H_{t,0}(\theta) - H_{t,n}(\theta))[\frac{\partial}{\partial\theta}f(t;\theta,\theta)]d < M(\theta) >_t |$$

$$\leq \sqrt{E_\theta\int_0^T(H_{t,0}(\theta) - H_{t,n}(\theta))^2 d < M(\theta) >_t}\sqrt{E_\theta\int_0^T(\frac{\partial}{\partial\theta}f(t;\theta,\theta))^2 d < M(\theta) >_t}$$

$$= \|\psi_0 - \psi_n\|_2\sqrt{E_\theta\int_0^T(\frac{\partial}{\partial\theta}f(t;\theta,\theta))^2 d < M(\theta) >_t} \to 0$$

$$\text{as } n \to \infty,$$

which yields $E_\theta\dot{\psi}_o = 0$, for all $\theta \in \Theta$. The above inequality comes from Yan(1981). Therefore

$$\mathcal{A}_{loc} = \{\psi : E_\theta\dot{\psi}(\theta) = 0, \qquad \text{for all } \theta, \psi \in \Psi\}.$$

Now we are going to give another example.

Example 2: (diffusion process) Consider the model

$$dX_t = a_t(\theta)dt + \sigma_t(\theta)dW_t,$$

where $\{a_t(\theta)\}$ and $\{\sigma_t(\theta)\}$ are predictable processes, both observable up to the value of the parameter θ and W is a Wiener process. The above model can be rewritten as

$$dX_t = \frac{a_t(\theta)}{\sigma_t^2(\theta)}d < \int_0^{\cdot} \sigma_s(\theta)dW_s(\theta) >_t + d(\int_0^t \sigma_s(\theta)dW_s).$$

Let $f(t;\eta,\theta) = a_t(\eta)/\sigma_t^2(\theta)$; then we have

$$\int_0^s \frac{a_t(\eta)}{\sigma_t^2(\eta)}d < \int_0^{\cdot} \sigma_s(\eta)dW_s(\eta) >_t = \int_0^s a_t(\eta)dt$$

$$= \int_0^s \frac{a_t(\eta)}{\sigma_t^2(\theta)}d < \int_0^{\cdot} \sigma_s(\theta)dW_s(\theta) >_t$$

$$= \int_0^s f(t;\eta.\theta)d < \int_0^{\cdot} \sigma_s(\theta)dW_s(\theta) >_t .$$

Let the estimating function space Ψ be $\{\int_0^T H_t(\theta)[dX_t - a_t(\theta)dt] | \quad \{H_t\}$ is a predictable process $\}$ and $\mathcal{G}_1 \in \Psi$ be such that each element G satisfies the conditions that EGG' and $E\dot{G}$ are nonsingular. Due to Theorem 5, Proposition 1 and the remark under Definition 8, the quasi-score estimating function

$$G^* = \int_0^t \dot{a}_t(\theta)\frac{1}{\sigma_t}dW_t = \int_0^t \dot{a}_t(\theta)\frac{1}{\sigma_t^2}[dX_t - a_t(\theta)dt]$$

is locally E-sufficient, i.e. $G^* \in \mathcal{S}_{loc}$.

Remark: In Example 2, if we denote by $\sigma(\theta)$ a linear subspace generated by G^*. Then, for any $\psi(\theta) = \int_0^T H_t(\theta)[dX_t - a_t(\theta)dt] \in \Psi$, we have that

$$< \psi, \phi >_\theta = 0$$

for all $\phi \in \sigma(G^*)$ and $\theta \in \Theta$ iff

$$< \psi, G^* >_\theta = 0, \qquad \text{for all } \theta \in \Theta.$$

Therefore, if

$$< \psi, G^* >_\theta = 0, \qquad \text{for all } \theta \in \Theta,$$

we can obtain

$$0 = E\int_0^T H_t(\theta)\frac{\dot{a}_t(\theta)}{\sigma_t^2(\theta)}d < \int_0^{\cdot}\sigma_s(\theta)dW_s >_t$$

$$= E\int_0^T H_t(\theta)[\frac{\partial}{\partial\eta}f(t;\eta,\theta)]_{\eta=\theta}d < \int_0^{\cdot}\sigma_s(\theta)dW_s >_t$$

$$= E\dot{\psi}(\theta),$$

i.e. $\psi \in \mathcal{A}_{loc}$. That means $\sigma(G^*)$ is a locally E-sufficient set. But $\mathcal{S}_{loc}$ is a product set and $G^* \in \mathcal{S}_{loc}$ which forces $\mathcal{S}_{loc} = \sigma(G^*)$.

In the view of Theorem 5, Example 1 and Example 2, we find that Assumption 1 is an important condition. How to check this condition is still a question. Even though Proposition 1 tells us that, in the case of model (5) , Assumption 1 is always true and model (5) covers many process estimation problems, an answer for the above question is still required. This is provided in the following proposition.

Proposition 2 *Let Ψ be a subset of estimating functions and $\Theta \subset R$ be a parameter space. Suppose that Ψ is a Hilbert space, $E_\theta\psi = 0$ and $E_\theta\psi^2 < \infty$ for all $\psi \in \Psi$ and $\theta \in \Theta$. Let $f(x;\theta)$ be the likelihood function, $x \in R^m$ for some $m > 0$. If*

$$\int_{-\infty}^{\infty}\frac{\dot{f}^2(x;\theta)}{f(x;\theta)}dx < \infty,$$

i.e.

$$E_\theta U^2 < \infty,$$

where U is the score function, and

$$\int_{-\infty}^{\infty}\frac{\partial(\psi(x;\theta)f(x;\theta))}{\partial\theta}dx = \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}\psi(x;\theta)f(x;\theta)dx$$

then

$$\mathcal{A}_{loc} = \{\psi : E_\theta\dot{\psi} = 0, \text{ for all } \theta\}.$$

Proof: Since

$$E_\theta\dot{\psi}(x;\theta) = \int_{-\infty}^{\infty}\dot{\psi}(x;\theta)f(x;\theta)dx$$

$$= \int_{-\infty}^{\infty}\frac{\partial(\psi(x;\theta)f(x;\theta))}{\partial\theta}dx - \int_{-\infty}^{\infty}\psi(x;\theta)\dot{f}(x;\theta)dx$$

$$= \frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} \psi(x;\theta) f(x;\theta) dx - \int_{-\infty}^{\infty} \psi(x;\theta) \dot{f}(x;\theta) dx$$

$$= - \int_{-\infty}^{\infty} \psi(x;\theta) \dot{f}(x;\theta) dx,$$

if ψ is the weak square limit of ψ_n with $E_\theta \dot{\psi}_n = 0$ for all $\theta \in \Theta$, we have

$$|E_\theta(\dot{\psi}_n - \dot{\psi})| = | \int_{-\infty}^{\infty} (\psi_n(x,\theta) - \psi_n(x,\theta)) \dot{f}(x;\theta) dx|$$

$$= | \int_{-\infty}^{\infty} (\psi_n(x,\theta) - \psi_n(x,\theta)) \frac{\dot{f}(x;\theta)}{f(x;\theta)} f(x;\theta) dx|$$

$$\leq \sqrt{\int_{-\infty}^{\infty} (\psi_n(x,\theta) - \psi_n(x,\theta))^2 f(x;\theta) dx} \sqrt{\int_{-\infty}^{\infty} (\frac{\dot{f}(x;\theta)}{f(x;\theta)})^2 f(x;\theta) dx} \ .$$

Let $n \to \infty$; we obtain

$$|E_\theta \dot{\psi}| = |E_\theta(\dot{\psi}_n - \dot{\psi})| \to 0$$

for all $\theta \in \Theta$. Therefore $E_\theta \dot{\psi} = 0$ for all $\theta \in \Theta$, i.e. $\mathcal{A}_{loc} = \{\psi : E_\theta \dot{\psi} = 0$, for all $\theta \in \Theta\}$.

For the higher dimensional parameter space case, we also have a similar Lemma.

Lemma 1 *Let Ψ be a set of estimating functions with $E_\theta \psi = 0$ and $E_\theta \psi \psi' < \infty$ for all $\psi \in \Psi$ and $\theta \in \Theta \subseteq R^n$, where Θ is a parameter space. Let $f(x,\theta)$ be the likelihood function with $x \in R^m$ for some m. If*

$$\int_{R^m} \frac{(\frac{\partial f}{\partial \theta_i})}{f(x;\theta)} dx < \infty, \qquad i = 1, 2, \ldots, n,$$

and

$$\int_{R^m} \frac{\partial}{\partial \theta} (\psi(x;\theta) f(x;\theta)) dx = \frac{\partial}{\partial \theta} \int_{R^m} \psi(x;\theta) f(x;\theta) dx$$

then

$$\mathcal{A}_{loc} = \{\psi : E_\theta \dot{\psi} = 0, \quad \textit{for all } \theta \in \Theta\}.$$

The proof is as same as that in Lemma 1. Taking $m = 1$ for convenience, we only need to notice that

$$E_\theta \dot{\psi}(x;\theta_1, \ldots, \theta_n) = E_\theta(\frac{\partial \psi_i}{\partial \theta_j})_{n \times n}$$

$$= \left(\int_{-\infty}^{\infty} \frac{\partial \psi_i}{\partial \theta_j} f(x;\theta_1,\ldots,\theta_n)dx\right)_{n\times n}$$

$$= -\left(\int_{-\infty}^{\infty} \psi_i \frac{\frac{\partial}{\partial \theta_j} f(x;\theta)}{f(x;\theta)} f(x;\theta)dx\right)_{n\times n}.$$

If ψ is a weak square limit of ψ_n, then

$$\left|\int_{-\infty}^{\infty} (\psi_{n,i} - \psi_i) \frac{\frac{\partial}{\partial \theta_j} f(x;\theta)}{f(x;\theta)} f(x;\theta)dx\right|$$

$$\leq \sqrt{\int_{-\infty}^{\infty} (\psi_{n,i} - \psi_i)^2 f(x;\theta)dx} \sqrt{\int_{-\infty}^{\infty} \left(\frac{\frac{\partial}{\partial \theta_j} f(x;\theta)}{f(x;\theta)}\right)^2 f(x;\theta)dx}$$

$$\to 0, \qquad \text{as } n \to \infty.$$

CHAPTER 3
THE CONSISTENCY OF QUASI-LIKELIHOOD ESTIMATORS

1 Introduction

In many practical cases it is easy to write down the quasi-score estimating function for a given process and to obtain the quasi-likelihood estimators. However,as with other methods the quasi-likelihood method also faces an important question of whether the quasi-likelihood estimators are consistent. Although the concept of " quasi-likelihood " was posed by Wedderburn in 1974, it is only in recent years that much detailed attention has been given to the properties of quasi-likelihood method. Few papers discuss the consistency of the quasi-likelihood estimators.

Let $(X_t, \mathcal{F}_t)$ be a process and Θ be a parameter space. Given X, there is an estimating function space Ψ_T associated with X and the population size T. Under some necessary regularity conditions, let $\mathcal{G}_T$ be a subspace of Ψ_T such that every element G_T of $\mathcal{G}_T$ satisfies the conditions that $E_\theta \dot{G}_T$ and $E_\theta G_T G_T'$ are nonsingular for all $\theta \in \Theta$, where the prime denotes transpose and the dot refers to differentiation with respect to the components of θ. Denote by G_T^* the quasi-score estimating function in $\mathcal{G}_T$ if G_T^* exists. Denote by $\hat{\theta}_T$ the quasi-likelihood estimator if

$$G_T^*(\hat{\theta}_T) = 0.$$

In this chapter we always denote the true parameter by θ_o.

In Hutton and Nelson(1986) a sufficient condition for the consistency of quasi-likelihood estimators was given. They assumed that:

(1) there is a function $M_T(\theta)$ in R such that $\dot{M}_T(\theta) = G^*(\theta)$, for all $\theta \in \Theta$, where $G_T^*(\theta)$ is a quasi-score estimating function in $\mathcal{G}_T$ as well as a martingale,

(2) $\lambda_{min}(< G^*(\theta) >_T) \to \infty$, as $T \to \infty$, where $\lambda_{\min}(< G^*(\theta) >_T$ denotes the minimum eigenvalue of $< G^*(\theta) >_T$,

(3) there is an increasing nonnegative function $h(\cdot)$ satisfying $\int_0^\infty h^{-2}\, dx < \infty$,

such that

$$\limsup_{T\to\infty} \frac{h(\lambda_{\max}(< G^*(\theta) >_T))}{\lambda_{\min}(< G^*(\theta) >_T)} < \infty,$$

where $\lambda_{\max}(< G^*(\theta) >_T)$ denotes the maximum eigenvalue of $< G^*(\theta) >_T$,

(4)

$$\limsup_{T\to\infty}\left(\frac{\sup_{\|\theta-\theta_\circ\|=\delta} R_T(\theta)}{(\theta-\theta_\circ)' < G^*(\theta) >_T (\theta-\theta_\circ)}\right) < 1/2$$

for all small $\delta > 0$, where $\theta_\circ$ is the true parameter and $R_T(\theta)$ is given by

$$M_T(\theta) = M_T(\theta_\circ) + (\theta-\theta_\circ)' G_T^*(\theta_\circ) - 1/2(\theta-\theta_\circ)' < G^*(\theta_\circ) >_T (\theta-\theta_\circ) + R_T(\theta).$$

Then they showed that there is a sequence of quasi-likelihood estimators $\hat{\theta}_n$ which converges a.s. to the true parameter $\theta_\circ$.

The main idea of Hutton and Nelson to get the above sufficient condition for the consistency of quasi-likelihood estimators comes from the ML method. This idea is very natural. As we have mentioned before, the quasi-likelihood method is developed from the ML and LS methods and of course many ideas and techniques for proving consistency of estimators in these contexts can be borrowed for the quasi-likelihood method. However the quasi-likelihood method also has some specific features of its own and for the sake of completeness, a discussion of the consistency of quasi-likelihood estimators is necessary.

We discuss the consistency of the quasi-likelihood estimators in two important cases. The first one is where the quasi-score estimating function has an integral surface. Hutton and Nelson's result belongs to this case. We can see that Hutton and Nelson's conditions are not very general, especially in Assumption 3. They require that the speeds with which $\lambda_{max}(< G^*(\theta) >_T)$ and $\lambda_{min}(< G^*(\theta) >_T)$ tend to infinity should be controlled by an increasing nonnegative function $h(\cdot)$ which satisfies $\int_0^\infty h^{-2}dx < \infty$. Clearly this is rather restrictive.

Following the approach of Hutton and Nelson, we here give a more general result than Hutton and Nelson's by using an idea of Kaufmann(1989). The result appears in Section 3. We can see from there that the restriction in Hutton and Nelson's Condition 3 is loosened by substituting a positive definite matrix A_T for h. In some sense, this means that the speeds with which $\lambda_{max}(< G^*(\theta) >_T)$ and $\lambda_{min}(< G^*(\theta) >)$ tend to infinity can just be controlled individually.

The other important case we consider is the asymptotically linear case. In many situations, the quasi-score estimating function can be represented as a martingale. For example, consider a model

$$X_t = \int_0^t f_s(\theta)ds + m_t,$$

where $\{m_t\}$ is a Brownian motion with respect to $\sigma-$fields $\{\mathcal{F}_t\}$, $f_t(\theta) \in \mathcal{F}_{t-}$ and $\theta \in \Theta$ where Θ is a parameter space. It is easy to see that the quasi-score estimating function in $\mathcal{G}_T \subseteq \Psi_T = \{\int_0^t H_s(\theta)dm_s | H_s(\theta) \in \mathcal{F}_{s-}\}$ is

$$G_T^*(\theta) = \int_0^t (\dot{f}_s(\theta))' dm_s$$

and quasi-likelihood estimator θ_T^* satisfies

$$(\theta_T^* - \theta_0) = -(\dot{G}_T^*(\theta))^{-1} G_T^*(\theta),$$

if $f_s(\theta)$ is a linear function of θ, where θ_0 is the true parameter. Therefore the question about the strong consistency of the quasi-likelihood estimators can be transformed to another question about when the strong law of large numbers for martingales holds. Here we use the strong (or weak) law of large numbers for martingales as a tool for investigating the consistency of quasi-likelihood estimators when the quasi-score estimating function is asymptotically linear. Furthermore the property of asymptotic normality of martingales can also be utilized to obtain the consistency of the quasi-likelihood estimators.

Before we give the sufficient conditions for the consistency of quasi-likelihood estimators, we give two closely related examples, one of which shows that the quasi-likelihood method does not ensure the consistency of the estimators.

2 Examples

In the one parameter case with independent observations, the ML method always provides consistent ML estimators (see, Nguyen and Rogers, 1989). But in other cases, the consistency of ML estimators holds only under some sufficient conditions. That means that the ML method does not ensure the consistency of ML estimators. Not unexpectedly the quasi-likelihood method does not improve on the ML method in this respect. It should not be surprising that quasi-likelihood estimators may not converge to the true parameter θ_o. In a sense, a quasi-likelihood estimator just depends on one sample path of a realization of a given process and it is possible that this may not provide unboundedly increasing information about the parameter θ_o. In this case it may result in the quasi-likelihood estimators not converging to the true parameter θ_o. We give two examples coming from similar models, one of which was introduced by Winnicki(1988). The quasi-likelihood method is applied to both of these. However, in one case the quasi-likelihood estimators converge almost surely to the true parameter while in the other case this does not happen.

Example 1. Suppose that X_0 and $\{I_i, i = 1, 2, \ldots\}$ are independent nonnegative integer valued r.v's , the $\{I_i\}$ being independent and identically distributed. Let

$$X_i = mX_{i-1} + I_i, \qquad i = 1, 2, \ldots, \tag{1}$$

where $EI_i = \lambda$, the variance $D(I_i) = b^2$ and $m > 1$ is a known parameter.

It is easy to construct a martingale $\{M_n = \sum_{i=0}^{n}(X_{i+1} - X_i m - \lambda) = \sum_{i=1}^{n}(I_i - \lambda)\}$ with respect to the natural σ-fields $\{\mathcal{F}_n = \sigma(X_j, j \leq n+1)\}$ from (1). Then there is a quasi-score estimating function G_n^* in $\mathcal{G}_n \subseteq \Psi_n = \{\sum \alpha_i(X_{i+1} - X_i m - \lambda) | \quad \alpha_i \in \mathcal{F}_{i-1}\}$, namely

$$G_n^* = \sum_{i=0}^{n} \frac{X_{i+1} - mX_i - \lambda}{b^2} = \sum_{i=1}^{n} \frac{I_i - \lambda}{b^2}.$$

Let $G_n^* = 0$; then we obtain a sequence of quasi-likelihood estimators $\{\hat{\lambda}_n = \sum_{i=1}^{n} I_i/n\}$ which converges almost surely to the true parameter λ_o, as $n \to \infty$.

But in Example 2 , we shall find a different behaviour.

Example 2. Let $\{X_n\}$ be a sequence of random variables such that

$$X_n = \sum_{i=1}^{X_{n-1}} Y_{n,i} + I_n, \qquad n = 1, 2, \ldots,$$

where X_0 is a nonnegative integer valued r.v. and $\{Y_{n,i}, i = 1, 2, \ldots, n = 1, 2, \ldots\}$

and $\{I_n, n = 1, 2, \ldots\}$ are independent families of i.i.d. nonnegative integer valued r.v.'s. Suppose that $m = EY_{n,i} \geq 1$, $\lambda = EI_n$, $\sigma^2 = D(Y_{n,i}) = D(I_n)$, $Y_{n,i} \geq 2$ and $X_0 \geq 1$. Let

$$M_n = \sum_{i=0}^{n}(X_{i+1} - mX_i - \lambda).$$

Then $\{M_n\}$ is a martingale with respect to $\{\mathcal{F}_n = \sigma(X_0, \ldots, X_{n+1})\}$ and the quasi-score estimating function G_n^* in $\mathcal{G}_n \subseteq \Psi_n = \{\sum \alpha_i M_i | \quad \alpha_i \in \mathcal{F}_{i-1}\}$ is

$$G_n^* = \sum_{i=0}^{n} \begin{pmatrix} X_i \\ 1 \end{pmatrix} \frac{X_{i+1} - mX_i - \lambda}{\sigma^2(X_i + 1)}.$$

Let $G_n^* = 0$; we obtain the quasi-likelihood estimators

$$\hat{m}_n = \frac{\sum_0^n X_i X_{i+1}/(1+X_i) \sum_0^n 1/(1+X_i) - \sum_0^n X_{i+1}/(1+X_i) \sum_0^n X_i/(1+X_i)}{\sum_0^n X_i^2/(1+X_i) \sum_0^n 1/(1+X_i) - (\sum_0^n X_i/(1+X_i))^2},$$

$$\hat{\lambda}_n = \frac{\sum_0^n X_i^2/(1+X_i) \sum_0^n X_{i+1}/(1+X_i) - \sum_0^n X_i/(1+X_i) \sum_0^n X_i X_{i+1}/(1+X_i)}{\sum_0^n X_i^2/(1+X_i) \sum_0^n 1/(1+X_i) - (\sum_0^n X_i/(1+X_i))^2}.$$

We are going to prove that $\hat{m}_n$ and $\hat{\lambda}_n$ are convergent. First, $\hat{m}_n$ can be rewritten as

$$\begin{aligned} \hat{m}_n = & \frac{\sum_0^n 1/(1+X_i) \sum_0^n X_i(X_{i+1} - m_0 X_i - \lambda_0)/(1+X_i)}{\sum_0^n X_i^2/(1+X_i) \sum_0^n 1/(1+X_i) - (\sum_0^n X_i/(1+X_i))^2} \\ & - \frac{\sum_0^n X_i/(1+X_i) \sum_0^n (X_{i+1} - m_0 X_i - \lambda_0)/(1+X_i)}{\sum_0^n X_i^2/(1+X_i) \sum_0^n 1/(1+X_i) - (\sum_0^n X_i/(1+X_i))^2} + m_0 \\ & = T_{n1} + T_{n2} + m_0, \end{aligned}$$

where m_0 and λ_0 are the true parameters. Following the assumptions,

$$\begin{aligned} T_{n1} + T_{n2} \sim & \frac{\sum_0^n 1/(1+X_i) \sum_0^n X_i(X_{i+1} - m_0 X_i - \lambda_0)/(1+X_i)}{\sum_0^n X_i^2/(1+X_i) \sum_0^n 1/(1+X_i)} \\ & - \frac{\sum_0^n X_i/(1+X_i) \sum_0^n (X_{i+1} - m_0 X_i - \lambda_0)/(1+X_i)}{\sum_0^n X_i^2/(1+X_i) \sum_0^n 1/(1+X_i)} \\ & = L_1 + L_2. \end{aligned}$$

Since $\sum_0^n X_i(X_{i+1} - m_0 X_i - \lambda_0)/(1+X_i)$ is a martingale with respect to $\{\mathcal{F}_n\}$ and

$$\sum_0^\infty \frac{E((X_n X_{n+1}/(1+X_n))^2 | \mathcal{F}_{n-1})}{(\sum_0^n X_i^2/(1+X_i))^2} < \infty,$$

then

$$L_1 \xrightarrow{a.s.} 0, \qquad \text{as } n \to \infty,$$

by using Theorem 2.18 of Hall and Heyde(1980). Similarly, we can also obtain

$$L_2 \xrightarrow{a.s.} 0, \qquad \text{as } n \to \infty.$$

Therefore $\hat{m}_n \xrightarrow{a.s.} m_o$, as $n \to \infty$.

Now $\hat{\lambda}_n$ can be rewritten as

$$\begin{aligned}
\hat{\lambda}_n &= \lambda_o + \frac{\sum_0^n X_i^2/(1+X_i)\sum_0^n(X_{i+1}-m_oX_i-\lambda_o)/(1+X_i)}{\sum_0^n X_i^2/(1+X_i)\sum_0^n 1/(1+X_i)-(\sum_0^n X_i/(1+X_i))^2} \\
&\quad - \frac{\sum_0^n X_i/(1+X_i)\sum_0^n X_i(X_{i+1}-m_oX_i-\lambda_o)/(1+X_i)}{\sum_0^n X_i^2/(1+X_i)\sum_0^n 1/(1+X_i)-(\sum_0^n X_i/(1+X_i))^2} \\
&\sim \lambda_o + \frac{\sum_0^n(X_{i+1}-m_oX_i-\lambda_o)/(1+X_i)}{\sum_0^n 1/(1+X_i)} \\
&\quad - \frac{\sum_0^n X_i/(1+X_i)\sum_0^n X_i(X_{i+1}-m_oX_i-\lambda_o)/(1+X_i)}{\sum_0^n X_i^2/(1+X_i)\sum_0^n 1/(1+X_i)}.
\end{aligned}$$

It is easy to show that

$$\hat{\lambda}_n \xrightarrow{a.s.} \lambda_o + \lim_{n\to\infty}\frac{\sum_0^n(X_{i+1}-m_oX_i-\lambda_o)/(1+X_i)}{\sum_0^n 1/(1+X_i)};$$

that is, $\hat{\lambda}_n$ converges to a random limit but not the true parameter λ_o, even though $\hat{m}_n \xrightarrow{a.s.} m_o$, as $n \to \infty$.

Example 2 essentially tells us that the martingale $\{M_n\}$ does not provide enough information for the estimation of λ_o. This fact can also be seen from the information matrix $(E\dot{G}_n^*)^{-1}(EG_n^*G_n^{*\prime})((E\dot{G}_n^*)')^{-1}$. Since

$$G_n^* = \sum_{i=0}^n \begin{pmatrix} X_i \\ 1 \end{pmatrix} \frac{X_{i+1}-mX_i\lambda}{X_i+1},$$

we have

$$\begin{aligned}
&(E(\dot{G}_n^*)')(EG_n^*G_n^{*\prime})^{-1}(E(\dot{G}_n^*)) \\
&= E\sum_{i=0}^n \begin{pmatrix} \frac{X_i^2}{X_i+1} & \frac{X_i}{X_i+1} \\ \frac{X_i}{X_i+1} & \frac{1}{X_i+1} \end{pmatrix}
\end{aligned}$$

which has maximum eigenvalue

$$\lambda_{max}(n) = 1/2[E(\sum_{i=0}^n(X_i^2+1)/(X_i+1))+\sqrt{(E\sum_{i=0}^n(X_i-1))^2+4(E\sum_{i=0}^n X_i/(X_i+1))^2}\ ]$$

and minimum eigenvalue

$$\lambda_{min}(n) = 1/2[E(\sum_{i=0}^{n}(X_i^2+1)/(X_i+1)) - \sqrt{(E\sum_{i=0}^{n}(X_i-1))^2 + 4(E\sum_{i=0}^{n}X_i/(X_i+1))^2}\ ]$$

respectively. After some algebra, we obtain

$$\lim_{n\to\infty} \lambda_{max}(n) = \infty$$

and

$$0 \le \lambda_{min}(n) \le E\sum_{i=0}^{n}(1+X_i)^{-1} < \infty,$$

Thus, λ_{max} tends to ∞ but λ_{min} is bounded. In Chapter 4, we shall see that, under some suitable conditions, $(E(\dot{G}_n^*))^{-1}(EG_n^*G_n^{*\prime})(E(\dot{G}_n^*)')^{-1}$ can determine the asymptotic confidence intervals of the true parameters. Therefore if one of the eigenvalues of $(E(\dot{G}_n^*))^{-1}(EG_n^*G_n^{*\prime})(E(\dot{G}_n^*)')^{-1}$ does not tend to 0 or tends to 0 very slowly, the quasi-likelihood estimators obtained from the quasi-likelihood equation $G_n^*(\theta) = 0$ may be poor estimators. The reason is that in this case some components of the quasi-likelihood estimator may be far from the corresponding components of the true parameter.

Given a process we can construct a martingale from it and get an estimating function space associated with this martingale. Then we can choose a quasi-score estimating function from this estimating function space if the quasi-score estimating function exists. Finally, a quasi-likelihood estimator can be obtained from the quasi-score estimating function . If the quasi-likelihood estimators are consistent, we say that the martingale provides enough information about the parameter or the martingale carries enough information from the given process. Since the martingale $\{M_n\}$ is constructed as a free choice, we can of course construct another different martingale from same process. Therefore we can get different estimating function spaces. From these different estimating function spaces, different quasi-score estimating functions are respectively determined. Then we can obtain different sequences of quasi-likelihood estimators. As we have indicated before, these sequences of quasi-likelihood estimators may determine which martingales provide enough information about the parameters. From this point of view, the above two examples motivate us to ask whether there is any other martingale associated with the given process which provides more information about the parameters of the process than $\{M_t\}$. Or we can ask whether, for a given process, we can determine a martingale associated with this process which carries the most information about the parameters from the process?

Before we examine these questions, we need some notations and some basic knowledge about semimartingales(see, Chapter III in Jacod and Shiryaev(1987).

A stochastic basis $(\Omega, \mathcal{F}, F, P)$ supports a semimartingale $X = (X^i)_{i \leq d}$. Let X^c be the continuous martingale part of X. Define a random measure μ on $R^d \times [0, \infty)$ by

$$\mu(dx, dt) = \sum_s I_{\{\triangle x_s \neq 0\}} \varepsilon_{(s, \triangle x_s)}(dt, dx),$$

where $\triangle x_s = x_s - x_{s-}$ and ε_a is the Dirac measure at a. Further, let ν be the predictable compensator of μ.

Let $\mathcal{A}^+_{loc}$ be a set of all real value processes A that are cádlág (right-continuous and admit left-hand limits) adapted, with $A_0 = 0$ and where each path is non-decreasing. Let $G_{loc}(\mu)$ be a set of all measurable functions W on $\Omega \times R^d \times [0, \infty)$ such that $\tilde{W}_t(\omega) = W(\omega, t, \triangle x_t) I_{(\triangle x \neq 0)} - E[W(t, \triangle x_t) I_{(\triangle x \neq)}(t) | \mathcal{F}_{t-}]$ satisfies $[\sum_{s \leq \cdot} (\tilde{W}_s)^2]^{1/2} \in \mathcal{A}^+_{loc}$. Let $L^2_{loc}(X^c)$ be a set of all predictable processes H such that the increasing process $(H \cdot c \cdot H) \cdot A$ is locally integrable, where $c, A \in R^d \times R^d$, $< X^{c(i)}, X^{c(j)} >= c_{ij} \cdot A$ and $X^c = (X^{c(i)}, \cdots, X^{c(d)})$.

<u>Definition</u>: We say that a local martingale M has the representation property relative to X if it has the form

$$M = M_0 + H \cdot X^c + W * (\mu - \nu),$$

where $H = (H^i)_{i \leq d}$ belongs to $L^2_{loc}(X^c)$ and $W \in G_{loc}(\mu)$.

Jacod and Shiryaev(1987, p172) pointed out that:

<u>Lemma III 4.24</u> : Every local martingale M has a decomposition

$$M = H \cdot X^c + W * (\mu - \nu) + N, \qquad (*)$$

where $H \in L^2_{loc}(X^c)$, $W \in G_{loc}(\mu)$, N is a local martingale and

$$< N^c, (C^c)^i >= 0, \qquad \forall i \leq d, \qquad M^p_\mu(\triangle N|\tilde{\mathcal{P}}) = 0,$$

where $\tilde{\mathcal{P}} = \mathcal{P} \times \mathcal{B}$, $\mathcal{P}$ is generated by all adapted processes that have continuous paths and $\mathcal{B}$ is the Borel σ-field in R^d. Moreover, this decomposition is unique, up to indistinguishability.

Jacod and Shiryaev(1987, III 4d) showed that, if X is of some specific form which includes cases such as X is a Wiener process or Poisson process, or more generally, a (homogeneous) exponential family of stochastic processes or independent increments processes, then every local martingale M has a decomposition

$$M = H \cdot X^c + W * (\mu - \nu). \qquad (**)$$

Now we consider the question raised above. Firstly, suppose that the given process X possesses the specific properties required in Jacod and Shiryaev(1987, III 4d). In this case (**) holds. Following the idea of Heyde(1987), taking any one of the local martingales of the form $H \cdot X^c$ and any one of the local martingales of the form $W * (\mu - \nu)$ we can establish an estimating function space $\Psi_t = \{\int_0^t \alpha_s(\theta) dX^c + \int_0^t \int_{R^d} \beta(\theta; x, s)(\mu - \nu(\theta))(dx, ds)\}$, where $\{\alpha_s\}$ and $\{\beta_s\}$ are predictable processes. It is obvious that this estimating function space contains all estimating function spaces based on martingales which are constructed from the given process X. From Ψ_t we can determine a quasi-score estimating function G^*_t (see Sørensen(1990)) and quasi-likelihood estimator θ^*_t. Assume that $\tilde{\Psi}_t$ is another estimating function space based on another martingale $\tilde{M}$ constructed from X, that is, $\tilde{\Psi}_t = \{\int_0^t \alpha_s d\tilde{M}_s | \{\alpha_t\}$ is a predictable process$\}$. Then we have $\tilde{\Psi}_t \subseteq \Psi_t$. Assume that $\tilde{G}^*_t$ is the quasi-score estimating function in $\tilde{\Psi}_t$ and $\tilde{\theta}^*_t$ is the quasi-likelihood estimator based on $\tilde{G}^*_t$. If the asymptotic confidence zones for θ_o based on $\tilde{G}^*_t$ and G^*_t are determined by $(E\dot{\tilde{G}}^*_t)^{-1}(E\tilde{G}^*_t\tilde{G}^{*\prime}_t)(E\dot{\tilde{G}}^{*\prime}_t)^{-1}$ and $(E\dot{G}^*_t)^{-1}(EG^*_tG^{*\prime}_t)(E\dot{G}^{*\prime}_t)^{-1}$ respectively, following the discussion in Chapter 4, the asymptotic confidence zone for θ_o based on G^*_t is smaller than that based on $\tilde{G}^*_t$. Therefore if $\{\theta^*_t\}$ is not consistent, $\{\tilde{\theta}^*_t\}$ will not be either. In other words, in this situation, there exist martingales the combination of which provides us with maximum information from the given process. However in practice it is often very difficult to write down them.

If X does not have some of the specific requirements mentioned in Jacod and Shiryaev(1987, III 4d), then not all local martingales M have the representation (**). That means that the local martingale N in (*) does not usually disappear. In this case the answer for the maximum information question is not clear.

Nevertheless, the technique discussed in Heyde(1987) and Heyde(1989) provides us with an important method for gathering more information from a given process. The procedure is that, after we get a martingale from the given process, and find that it does not provide us with enough information about the unknown parameter, i.e. the estimators are not consistent or it is difficult to get the quasi-likelihood estimator from the quasi-score equation, we need to find another martingale which is not in the space generated by the first martingale(preferable a martingale which is orthogonal to the first martingale) and then combine them to get more information from the given process.

3 Sufficient conditions for the consistency of quasi-likelihood estimators : integral surface case

Like the ML estimator which comes from the score equation $\partial(\log L(\theta))/\partial\theta = 0$, where L is the likelihood function, the quasi-likelihood estimator comes from the quasi-score equation $G_T^*(\theta) = 0$, where $G_T^*(\theta)$ is the quasi-score estimating function. The maximum likelihood estimator corresponds to maximizing the probability of the observed samples. Therefore, under the maximum likelihood principle, the ML estimator seems easy to accept. But in general the quasi-likelihood estimator does not show us any similar sensible interpretation except the meaning that under some regularity conditions the quasi-likelihood estimator is the root of an estimating function in the estimating function space, which has a shorter distance from

$\partial(logL)/\partial(\theta)$ than other estimating functions in the estimating function space(see, Godambe and Thompson, 1985). However we can still employ the idea of maximum likelihood to deal with the quasi-likelihood estimator question.

Here we always assume that the quasi-score estimating function G_T^* is a square integrable martingale.

Definition 1 *Let $M_T(\theta)$ be a function of θ in R. If*

$$\dot{M}_T(\theta) = \left(\frac{\partial M_T(\theta)}{\partial \theta_i}\right)_{k\times 1} = G_T^*(\theta),$$

we call $M_T(\theta)$ the integral surface of $G_T^*(\theta)$.

Assumption 1. For each $T \geq 0$, $G_T^*(\theta)$ has an integral surface $M_T(\theta)$.

For a process of positive definite matrices A_T (possibly random), let

$$N_T(\delta) = \{\theta| \quad \|A_T(\theta - \theta_\circ)\| \leq \delta \|A_T^{-1} G_T^*(\theta_\circ)\|\}.$$

Assumption 2.

$$\sup_{\theta \in N_T(\delta)} \|\theta - \theta_\circ\| \xrightarrow{a.s.} 0, \qquad \text{as } T \to \infty.$$

Assumption 3. There exists a T_0 such that, if $T \geq T_0$,

$$\dot{G}_T^*(\theta) = -< G^*(\theta) >_T + R_T(\theta)$$

and, for all $\theta \in N_T(\delta)$ and $\|\tilde{\theta} - \theta_\circ\| \leq \|\theta - \theta_\circ\|$,

$$\|[(\theta - \theta_\circ)' < G^*(\tilde{\theta}) >_T (\theta - \theta_\circ)]^{-1}[(\theta - \theta_\circ)' R_T(\tilde{\theta})(\theta - \theta_\circ)]\| \leq (1 - \alpha),$$

where $\alpha < 1$.

Assumption 4. There exists $T_\circ \geq 0$ such that, if $T \geq T_\circ$

$$P(< G^*(\theta) >_T - cA_T' A_T \geq 0, \qquad \text{for all } \theta \in N_T(\delta)) = 1,$$

where $\delta c = 2/\alpha$.

Theorem 1 *Under Assumptions 1-4, there exists a sequence of quasi-likelihood estimators $\hat{\theta}_n$ such that*

$$\hat{\theta}_n \xrightarrow{a.s.} \theta_\circ, \qquad \textit{as } n \to \infty.$$

Proof: Without loss of generality, we may suppose that Assumptions 3 and 4 hold for all T.

For $\omega \in \Omega$, if $G_T(\theta_\circ, \omega) = 0$, let $\hat{\theta}_T(\omega) = \theta_\circ$; if $G^*_T(\theta_\circ, \omega) \neq 0$, since $N_T(\delta)$ is a nonempty compact convex subset of Θ, for any boundary point θ of $N_T(\delta)$,

$$M_T(\theta) - M_T(\theta_\circ) = (\theta - \theta_\circ)' G^*_T(\theta_\circ) + 1/2(\theta - \theta_\circ)' \dot{G}^*_T(\tilde{\theta})(\theta - \theta_\circ)$$

$$= (\theta - \theta_\circ)' G^*_T(\theta_\circ) - 1/2(\theta - \theta_\circ)' < G^*(\tilde{\theta}) >_T (\theta - \theta_\circ)$$

$$+1/2(\theta - \theta_\circ)' R_T(\tilde{\theta})(\theta - \theta_\circ),$$

where $\tilde{\theta}$ lies in between θ and $\theta_\circ$. Set $V_T = A_T(\theta - \theta_\circ)$; then

$$M_T(\theta) - M_T(\theta_\circ) = V_T' A_T^{-1} G^*_T(\theta_\circ) - \frac{\alpha}{2} V_T' A_T^{-1} < G^*(\tilde{\theta}) >_T A_T^{-1} V_T$$

$$-\frac{1-\alpha}{2} V_T' A_T^{-1} < G^*(\tilde{\theta}) >_T A_T^{-1} V_T (1 - \frac{1}{1-\alpha}((\theta - \theta_\circ)' < G^*(\tilde{\theta}) >_n (\theta - \theta_\circ))^{-1}$$

$$\times((\theta - \theta_\circ)' R_n (\theta - \theta_\circ)))$$

$$\leq \|V_T\| \|A_T^{-1} G^*_T(\theta_\circ)\| - \frac{\alpha}{2} \|V_T\|^2 \lambda_{min}(A_T^{-1} < G^*(\tilde{\theta}) >_T A_T^{-1})$$

$$\leq \|V_T\| (\|A_T^{-1} G^*_T(\theta_\circ)\| - \frac{\alpha\delta c}{2} \|A_T^{-1} G^*_T(\theta_\circ)\|) = 0.$$

Therefore there exists $\hat{\theta}_T$ in $N_T(\delta)$ such that

$$\dot{M}_T(\hat{\theta}_T) = G^*_T(\hat{\theta}_T) = 0.$$

By virtue of Assumption 2, we have

$$\hat{\theta}_T \xrightarrow{a.s.} \theta_\circ, \qquad \text{as } T \to \infty.$$

As applications of Theorem 1, three examples are given below. Two of them, Example 1 and Example 2, are borrowed from Hutton and Nelson(1989).

Example 1. Suppose that $\{X_k, \mathcal{F}_k\}$ is a real valued discrete time stochastic process with

$$E(X_k|\mathcal{F}_{k-1}) = \theta(1+\theta)X_{k-1},$$

$$D(X_k|\mathcal{F}_{k-1}) = (1+2\theta)X_{k-1},$$

where $X_k \geq 1$, $k = 0, 1, \ldots$ and $\theta \geq 0$. Then $\tilde{X}_n = \sum_1^n (X_k - E(X_k|\mathcal{F}_{k-1}))$ is a martingale with respect to $\{\mathcal{F}_n\}$ and the quasi-score estimating function in $\mathcal{G}_n \subseteq \Psi_n = \{\sum_{k=1}^n \alpha_k(X_k - E(X_k|\mathcal{F}_{k-1}))|\{\alpha_k\}$ is a predictable process$\}$ is

$$G_n^*(\theta) = \sum_1^n (X_k - \theta(1+\theta)X_{k-1}).$$

It is easy to check that $M_n(\theta) = \sum_1^n(\theta X_k - \theta^2(1/2+1/3\theta)X_{k-1})$ is the integral surface of $G_n^*(\theta)$. In Assumptions 2-4, choosing $A_n = < G^*(\theta_\circ) >_n^{1/2} = (\sum_1^n(1+2\theta_\circ)X_{k-1})^{1/2}$, we have

$$\sup_{\theta \in N_n(\delta)}|\theta - \theta_\circ| \leq \sup_{\theta \in N_n(\delta)} \delta \frac{G_n^*(\theta_\circ)}{< G^*(\theta_\circ >_n} \xrightarrow{a.s.} 0, \qquad \text{as } n \to \infty,$$

since

$$\sum_{n=1}^{\infty} \frac{E((G_n^* - G_{n-1}^*)^2|\mathcal{F}_{n-1})}{(< G^*(\theta_\circ) >_n)^2} = \sum_{n=1}^{\infty} \frac{(1+2\theta_\circ)X_{n-1}}{(\sum_{k=1}^n(1+2\theta_\circ)X_{k-1})^2} < \infty.$$

Assumptions 3 and 4 can be checked by some straightforward calculations. Following Theorem 1 there exists a sequence of quasi-likelihood estimators $\hat{\theta}_n$ which satisfy $\hat{\theta}_n \xrightarrow{a.s.} \theta$, as $n \to \infty$.

Example 2. Let $\{m_t, \mathcal{F}_t\}$ be a locally square-integrable martingale with $< m >_T = T$. Consider the model

$$dy_t = (\theta_1\theta_2 + \theta_1\sqrt{ct^{c-1}})dt + dm_t,$$

where $\theta_1, \theta_2 \geq 0$, $1/2 \leq c \leq 1$. There exists a quasi-score estimating function $G_T^*(\theta)$ in $\mathcal{G}_T \subseteq \Psi_T = \{\int_0^T \alpha_s dm_s, \alpha_s \in \mathcal{F}_{s-}\}$ and

$$G_T^*(\theta) = \int_0^T \begin{pmatrix} \theta_2 + \sqrt{ct^{c-1}} \\ \theta_1 \end{pmatrix} dm_t.$$

It can easy be shown directly that G_T^* has an integral surface.

In Assumptions 3-4, we can choose $A_T = \sqrt{\theta_{o,1}^2 T^c/(\theta_{o,1}^2+\theta_{o,2}^2)}\ I$, in which $\theta_{o,1}$ and $\theta_{o,2}$ are the true parameters and I is the unit matrix. Since $1/2 \le c \le 1$, by virtue of Lepinglé's(1977) law of large numbers,

$$\sup_{\theta\in N_n(\delta)}\|\theta-\theta_o\| \le \sup_{\theta\in N_n(\delta)}\delta\|\frac{G_T^*}{(\theta_{0,1}^2/(\theta_{0,1}^2+\theta_{0,2}^2))T^c}\| \overset{a.s.}{\rightarrow} 0,$$

$$\text{as } n\rightarrow\infty.$$

Furthermore, since

$$\lambda_{max}(< G^* >_T) = (\theta_1^2+\theta_2^2)T(1+o(1))$$

and

$$\lambda_{min}(< G^* >_T) = (\theta_1^2/(\theta_1^2+\theta_2^2)T^c(1+o(1)),$$

we have

$$\|[(\theta-\theta_o)' < G^*(\tilde\theta) >_T (\theta-\theta_o)]^{-1}[(\theta-\theta_o)'R_T(\theta-\theta_o)]\|$$
$$\le |\frac{m_T(\tilde\theta)}{\tilde\theta_1^2/(\tilde\theta_1^2+\tilde\theta_2^2)T^c(1+o(1))}| \overset{a.s.}{\rightarrow} 0, \qquad \text{as } T\rightarrow\infty,$$

where $\tilde\theta \in N_T(\delta)$, and

$$< G^*(\theta) >_T -\theta_1^2/(\theta_1+\theta_2^2)T^cI \ge 0.$$

Following Theorem 1, there exists a sequence of quasi-likelihood estimators $\hat\theta_T$ such that $\hat\theta_T \overset{a.s.}{\rightarrow} \theta_o$, as $T\rightarrow\infty$.

In the above two examples, the quasi-score estimating functions satisfy Assumptions 1-4 of this section as well as Hutton and Nelson's Assumptions as listed in Section 1 of this Chapter. In the following example the quasi-score estimating function satisfies Assumptions 1-4 (and hence Theorem 1) but not Hutton and Nelson's Assumptions.

Example 3. Consider the model

$$dy_t = (\theta_1 t + \theta_2 t^{-1/4})dt + m_t,$$

where θ_1 and θ_2 are unknown parameters and $\{m_t, \mathcal{F}_t\}$ is a Wiener process. Let $\mathcal{G}_t = \{G_t = \int_0^t f_s(\theta)dm_s | f_s(\theta) \in \mathcal{F}_{s-}\}$ be an estimating function space, in which every element G_t satisfies the conditions that $E\dot{G}_t$ and EG_tG_t' are nonsingular. Then

$$G_t^* = \int_0^t \begin{pmatrix} s \\ s^{-1/4} \end{pmatrix} dm_s$$

is a quasi-score estimating function in $\mathcal{G}_t$.

In this example, we want to use Theorem 1 to show that there exists a sequence of quasi-likelihood estimators which converges strongly to the true parameter θ_o. We show that the quasi-score estimating function G^* satisfies Assumptions 1-4 but not Hutton and Nelson's Assumptions.

Firstly, let us evaluate the eigenvalues of $< G^* >_t$. Because $\{m_t\}$ is a Wiener process, we have

$$< G^* >_t = \begin{pmatrix} (1/3)t^3 & (4/7)t^{7/4} \\ (4/7)t^{7/4} & 2t^{1/2} \end{pmatrix}.$$

From the above matrix, we obtain the maximum eigenvalue and the minmum eigenvalue of $< G^* >_t$ following

$$\lambda_{max}(t) = 1/2[(2t^{1/2} + (1/3)t^3) + \sqrt{((1/3)t^3 - 2t^{1/2})^2 + (4/3)t^{7/2} - (12/441)t^{7/2}}]$$

$$\geq (1/3)t^3 > t, \qquad \text{for } t \text{ large enough,}$$

and

$$\lambda_{min}(t) = 1/2[(2t^{1/2} + (1/3)t^3) - \sqrt{((1/3)t^3 - 2t^{1/2})^2 + (4/3)t^{7/2} - (12/441)t^{7/2}}]$$

$$\leq 2t^{1/2}, \qquad \text{for } t \text{ large enough.}$$

In meanwhile, we also can get that $\lambda_{min}(t) \sim t^{1/2}O(1)$, as $t \to \infty$.

In fact G^* cannot satisfy Hutton and Nelson's Assumptions. If G^* does, there is an increasing positive function $h(x)$ such that

$$\int_0^\infty \frac{dx}{h^2(x)} < \infty$$

and

$$\overline{\lim_{t\to\infty}} \frac{h(\lambda_{max}(t))}{\lambda_{min}(t)} < \infty.$$

Therefore, there are $T_\circ > 0$ and $K > 0$ such that, for $t > T_\circ$,

$$h(\lambda_{max}(t)) \leq K\lambda_{min}(t) \leq 2Kt^{1/2}.$$

This leads to

$$\frac{1}{4K^2 t} \leq \frac{1}{K^2\lambda_{min}^2(t)} \leq \frac{1}{h^2(\lambda_{max}(t))} \leq \frac{1}{h^2(t)}$$

and

$$\infty = \int_0^\infty \frac{1}{4K^2 t} dt \leq \int_0^\infty \frac{dt}{h^2(t)} < \infty$$

which is a contradiction.

Now we prove that G^* satisfies Assumptions 1-4. Assumptions 1 and 3 hold obviously, because G^* has an integral surface

$$M_t(\theta) = \int_0^t [(\theta_1 s + \theta_2 s^{-1/4}) dy_s - 1/2(\theta_1 s + \theta_2 s^{-1/4})^2 ds]$$

and

$$\dot{G}_t^* = - < G^* >_t .$$

In the remaining argument we prove that G^* satisfies Assumption 2 and 4. We choose

$$A_t = \begin{pmatrix} 1/3\,t^3 & 0 \\ 0 & \lambda_{min}(t) \end{pmatrix}^{1/2} .$$

Let $N_t(\delta) = \{\theta | \|A_t(\theta - \theta_\circ\| \leq \delta \|A_t^{-1} G_t^*(\theta_\circ)\|\}$ for a δ which satisfies that $\delta c = 2$, where $0 < c \leq 1 - \sqrt{24/49} < 1$. Then, for any $\theta \in N_t(\delta)$, we have

$$1/3\,t^3(\theta_1 - \theta_{\circ,1})^2 + \lambda_{min}(t)(\theta_2 - \theta_{\circ,2})^2$$

$$\leq \delta^2[3/t^3(\int_0^t s dm_s)^2 + 1/\lambda(t)(\int_0^t s^{-1/4} dm_s)^2],$$

i.e.

$$\frac{t^3}{3\lambda_{min}^{1/2}(t)}(\theta_1 - \theta_{\circ,1})^2 + \lambda_{min}^{1/2}(t)(\theta_2 - \theta_{\circ,2})^2$$

$$\leq \delta^2[\frac{3}{t^3\lambda_{min}^{1/2}(t)}(\int_0^t s dm_s)^2 + \frac{1}{\lambda_{min}^{3/2}}(\int_0^t s^{-1/4} dm_s)^2].$$

Let $h_1(x) = x^{49/24}$ and $h_2(x) = x^{3/4}$. By Lepinglé's(1977) law of large numbers, we obtain

$$\frac{3}{t^{3/2}\lambda_{min}^{1/4}(t)} \int_0^t s dm_s$$
$$= \frac{3^{49/24} t^{1/8}}{\lambda_{min}^{1/4}(t)} \cdot \frac{\int_0^t s dm_s}{h_1(< \int_0^{\cdot} s dm_s >_t)} \stackrel{a.s.}{\rightarrow} 0, \quad \text{as } t \rightarrow \infty,$$

and

$$\frac{1}{\lambda_{min}^{3/4}} \int_0^t s^{-1/4} dm_s$$
$$= \frac{2^{3/4} t^{3/8}}{\lambda_{min}^{3/4}(t)} \cdot \frac{\int_0^t s^{-1/4} dm_s}{h_2(< \int_0^{\cdot} s^{-1/4} dm_s >_t)} \stackrel{a.s.}{\rightarrow} 0, \quad \text{as } t \rightarrow \infty.$$

Therefore

$$\sup_{\theta \in N_t(\delta)} \|\theta - \theta_\circ\| \stackrel{a.s.}{\rightarrow} 0, \quad \text{as } t \rightarrow \infty$$

and G^* satisfies Assumption 2.

Because $0 < c \leq 1 - \sqrt{24/49} < 1$, after some calculation, we find that

$$< G^* >_t - c A_t' A_t \geq 0,$$

i.e. G^* satisfies Assumption 4 and Theorem 1 applies.

Remark : If in the proof of Theorem 1, we replace Assumption 2 by

Assumption 2' With a process of positive definite matrices A_T (possibly random), let

$$N_T(\delta) = \{\theta \| \|A_T(\theta - \theta_\circ)\| \leq \delta \|A_T^{-1} G_T^*(\theta_\circ)\| \}$$

we have

$$\sup_{\theta \in N_T(\delta)} \|\theta - \theta_\circ\| \stackrel{p}{\rightarrow} 0, \qquad \text{as } T \rightarrow \infty.$$

That is, weak consistency holds for the quasi-likelihood estimators.

Remark : We should notice that not all quasi-score estimating functions have integral surfaces. Here is an example.

Example 4. Let $\{m_t, \mathcal{F}_t\}$ be a locally square-integrable martingale with $< m >_t = t$. Consider the model

$$dy_t = (\theta_1 \theta_2) dt + \theta_1 dm_t,$$

where $\theta_1,\ \theta_2 > 0$. There exists a quasi-score estimating function $G_T^*(\theta)$ in $\mathcal{G}_T \subseteq \Psi_T = \{\int_0^T \alpha_s dm_s |\ \alpha_s \in \mathcal{F}_{s-}\}$ and

$$G_T^* = \int_0^T \begin{pmatrix} \theta_2/\theta_1^2(dy_t - \theta_1\theta_2 dt) \\ \theta_1^{-1}(dy_t - \theta_1\theta_2 dt) \end{pmatrix}$$

$$= \begin{pmatrix} \theta_2/\theta_1^2 y_T - \theta_2^2/\theta_1 T \\ \theta_1^{-1} y_T - \theta_2^2 T \end{pmatrix}.$$

If G_T^* has an integral surface, say $\phi(\theta_1, \theta_2)$, then we should have

$$\frac{\partial \phi}{\partial \theta_1} = \theta_2/\theta_1^2 y_T - \theta_2^2/\theta_1 T.$$

This forces

$$\phi = -\theta_2/\theta_1 y_t - (\theta_2^2 log\theta_1)t + C(\theta_2)$$

and

$$\frac{\partial \phi}{\partial \theta_2} = -\theta_1^{-1} y_t - 2(\theta_2 log\theta_1)t + \frac{\partial}{\partial \theta_2} C(\theta_2)$$

$$= \theta_1^{-1} y_t - \theta_2 t,$$

where $C(\theta_2)$ is free from θ_1. Therefore we get

$$\frac{\partial}{\partial \theta_2} C(\theta_2) = 2(\theta_2^2 log\theta_1)t - \theta_2 t$$

which is a contradiction. Hence, it is impossible for G_T^* to have an integral surface. However, if in the above model, the quadratic process of the martingale part is free from the parameters, then the quasi-score estimating function usually has integral surface. For example, if the model is

$$dy_t = f_t(\theta_1, \theta_2)dt + dm_t,$$

where $< m >_t = \int_0^t a_s ds$, then the quasi-score estimating function is

$$G_t = \int_0^t \begin{pmatrix} \partial f_s/\partial \theta_1 \\ \partial f_s/\partial \theta_2 \end{pmatrix} a_s^+ dm_s$$

$$= \int_0^t \begin{pmatrix} \partial f_s/\partial \theta_1 \\ \partial f_s/\partial \theta_2 \end{pmatrix} a_s^+ (dy_s - f_s(\theta_1, \theta_2)ds).$$

Let $\phi_t(\theta_1, \theta_2)$ satisfy

$$\partial\phi_t/\partial\theta_1 = \int_0^t \partial f_s/\partial\theta_1 a_s^+(dy_s - f_s(\theta_1, \theta_2)ds).$$

Therefore, under regularity conditions,

$$\phi_t(\theta_1, \theta_2) = \int_0^t (f_s a_s^+ dy_s - 1/2 f_s^2 ds) + C(\theta_2).$$

Let

$$\partial\phi_t/\partial\theta_2 = \int_0^t \partial f_s/\partial\theta_2(a_s^+ dy_s - f_s(\theta_1, \theta_2)ds).$$

We get that $C(\theta_2) = 0$ and

$$\phi_t(\theta_1, \theta_2) = \int_0^t (f_s a_s^+ dy_s - 1/2 f_s^2 ds)$$

is the integral surface of G_t.

It is of course important to know when the quasi-score estimating function has an integral surface. This question for the generalized linear model has been discussed by McCullagh and Nelder(1989,p333). In this case they pointed out that the quasi-score estimating function G^* has integral surface iff the matrix of $\dot{G}^*$ is symmetric. However, there are very many cases in which, for $r \neq s$, $(\partial G_r^*(\theta))/(\partial\theta_s) \neq (\partial G_s^*(\theta))/(\partial\theta_r)$.

4 Sufficient conditions for the consistency of quasi-likelihood estimators : asymptotic linear case

Theorem 1 gives the existence of a quasi-likelihood estimator and sufficient conditions for consistency when G_T^* has an integral surface. Since not all G_T^* have an integral surface, we certainly need other methods to deal with the consistency of quasi-likelihood estimators when the quasi-score estimating function does not have an integral surface. In this section, we discuss this question in the asymptotic linear case.

Many estimating functions which have integral surface belong to the asymptotic linear category, but the converse is not true. An example will be given below. Therefore the asymptotic linear case is different from the integral surface case. Because the ~~skills~~ techniques used to prove the consistency of quasi-likelihood estimators can be quite different from case to case, we can only partially show the ideas here. We just state or prove some theorems and corollaries here. Many other results can be achieved by combining the results in Chapter 8 with those of this section.

Definition 2 *Suppose that $G_T(\theta)$ is a vector in R^k, where $\theta \in \Theta$ and Θ is a bounded subset of R^k. If $G_T(\theta)$ can be represented as*

$$G_T(\theta) = G_T(\theta_\circ) + \dot{G}_T(\theta_\circ)(\theta - \theta_\circ) + R_T(\theta, \theta_\circ)$$

where $(\dot{G}_T(\theta_\circ))^{-1} R_T \xrightarrow{a.s.} 0$, as $T \to \infty$, for all $\theta \in \Theta$, we say that $\{G_T\}$ is asymptotic linear .

Here is an example. Let us consider a model

$$dy_t = ((t+\theta_1)^3 + (t+\theta_2)^3)dt + \sqrt{(t+\theta_1)(t+\theta_2)}dW_t,$$

where θ_1 and θ_2 are parameters and W is a Wiener process. Following the formula in Godambe and Heyde(1980), the quasi-score estimating function in $\mathcal{G}_t \subseteq \Psi = \{\int_0^t \alpha_t(\theta_1, \theta_2) dW_t | \ \{\alpha_t\}$ is a predictable process$\}$ is

$$G_t^* = 3 \begin{pmatrix} \int_0^t (t+\theta_1)(t+\theta_2)^{-1} dy_t - \int_0^t (t+\theta_1)^4 (t+\theta_2)^{-1} dt - \int_0^t (t+\theta_1)(t+\theta_2)^2 dt \\ \int_0^t (t+\theta_2)(t+\theta_1)^{-1} dy_t - \int_0^t (t+\theta_2)^4 (t+\theta_1)^{-1} dt - \int_0^t (t+\theta_2)(t+\theta_1)^2 dt \end{pmatrix}.$$

After some algebra we obtain

$$\dot{G}_t^* = 3 \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix},$$

where

$$a_{1,1} = \int_0^t (t+\theta_2)^{-1} dy_t - 4\int_0^t (t+\theta_1)^3 (t+\theta_2)^{-1} dt - \int_0^t (t+\theta_2)^2 dt,$$

$$a_{1,2} = -\int_0^t (t+\theta_1)(t+\theta_2)^{-2} dy_t + \int_0^t (t+\theta_1)^4 (t+\theta_2)^{-2} dt - 2\int_0^t (t+\theta_1)(t+\theta_2)^3 dt,$$

$$a_{2,1} = -\int_0^t (t+\theta_2)(t+\theta_1)^{-2}dy_t + \int_0^t (t+\theta_2)^4(t+\theta_1)^{-2}dt - 2\int_0^t (t+\theta_2)(t+\theta_1)^3 dt$$

and

$$a_{2,2} = \int_0^t (t+\theta_1)^{-1}dy_t - 4\int_0^t (t+\theta_2)^3(t+\theta_1)^{-1}dt - \int_0^t (t+\theta_1)^2 dt,$$

while

$$2/3R_t = v_1(\theta_1 - \theta_{1,o})^2 + v_2(\theta_2 - \theta_{2,o})^2 + v_3(\theta_1 - \theta_{1,o})(\theta_2 - \theta_{2,o}).$$

where

$$v_1 = \begin{pmatrix} -12\int_0^t (t+\theta_1)^2(t+\theta_2)^{-1}dt \\ 2\int_0^t (t+\theta_1)^{-3}(t+\theta_2)dy_t - 2\int_0^t (t+\theta_1)^{-3}(t+\theta_2)^4 - 6\int_0^t (t+\theta_1)^2(t+\theta_2)dt \end{pmatrix},$$

$$v_2 = \begin{pmatrix} 2\int_0^t (t+\theta_2)^{-3}(t+\theta_1)dy_t - 2\int_0^t (t+\theta_2)^{-3}(t+\theta_1)^4 - 6\int_0^t (t+\theta_2)^2(t+\theta_1)dt \\ -12\int_0^t (t+\theta_2)^2(t+\theta_1)^{-1}dt \end{pmatrix}$$

and

$$v_3 = \begin{pmatrix} -\int_0^t (t+\theta_2)^{-2}dy_t + 4\int_0^t (t+\theta_1)^3(t+\theta_2)^{-2}dt - 2\int_0^t (t+\theta_2)^3 dt \\ -\int_0^t (t+\theta_1)^{-2}dy_t + 4\int_0^t (t+\theta_2)^3(t+\theta_1)^{-2}dt - 2\int_0^t (t+\theta_1)^3 dt \end{pmatrix}.$$

Applying a basic property of martingales, we get $y_t = O(t^4)$. Therefore, it is easy to get

$$(\dot{G}_t^*)^{-1}R_t \overset{a.s.}{\to} 0,$$

that is , G_t^* is asymptotic linear. But it is obvious that G_t^* does not have an integral surface.

Assumption 5. $\{G_T\}$ is a square integrable local martingale with respect to some σ-fields $\{\mathcal{F}_T\}$ and

$$< G(\theta) >_T^{-1} \dot{G}_T(\theta) \overset{a.s.}{\to} 1, \qquad \text{as } T \to \infty.$$

Assumption 6. There is sequence $\{c_T\}$ such that

$$(| < G(\theta) >_T | - |G_{T-1}(\theta)G_{T-1}(\theta)'|)/| < G(\theta) >_T | \leq c_T, a.s.,$$

for T large enough , where $E\sum_{T=1}^{\infty} c_T < \infty$.

In the following, we still denote the quasi-score estimating function by G_T^*.

Theorem 2 *Suppose that $\{G_T^*\}$ is asymptotic linear . Under Assumptions5 and 6, if $\lambda_{min}(< G^*(\theta) >_n) \xrightarrow{a.s.} \infty$, as $T \to \infty$, and there is a positive increasing function $h(x)$ such that*

$$\int_0^\infty \frac{dx}{xh(x)} < \infty,$$

where $h(|\dot{G}_T^*|) = o(\lambda_{min}(\dot{G}_T^*))$, then the quasi-likelihood estimator $\hat{\theta}_T \xrightarrow{a.s.} \theta_\circ$.

Proof: Let $\hat{\theta}_T$ be the quasi-likelihood estimator. Following the assumptions,

$$0 = G_T^*(\hat{\theta}_T) = G_T^*(\theta_\circ) + \dot{G}_T^*(\theta_\circ)(\hat{\theta}_T - \theta_\circ) + R_T(\theta)$$

$$= G_T^*(\theta_\circ) + < G^*(\theta_\circ) >_T (\hat{\theta}_T - \theta_\circ) + o(< G^*(\theta_\circ) >_T).$$

From Theorem 5 in Chapter 8, we obtain

$$\hat{\theta}_T \xrightarrow{a.s.} \theta_\circ, \qquad \text{as } T \to \infty.$$

The above theorem is based on the strong law of large numbers for martingales. Sometimes the asymptotic normality property of martingales can also be used to provide sufficient conditions for the consistency of quasi-likelihood estimators.

Theorem 3 *Under Assumption 5 , if*

(1) $\{G_T^*\}$ is asymptotic linear ,

(2) $< G^*(\theta) >_T^{-1/2} G_T^*(\theta) \xrightarrow{d} N(0, I)$, as $T \to \infty$,

(3) $\lambda_{min}(< G^*(\theta) >_T) \xrightarrow{a.s.} \infty$, as $T \to \infty$,

(4) $\hat{\theta}_T$ is a.s. convergent,

then $\hat{\theta}_T - \theta_\circ \xrightarrow{a.s.} 0$, as $n \to \infty$.

Proof: Let $\hat{\theta}_T$ be the quasi-likelihood estimator. Following the assumptions,

$$0 = G_T^*(\hat{\theta}_T) = G_T^*(\theta_\circ) + \dot{G}_T^*(\theta_\circ)(\hat{\theta}_T - \theta_\circ) + R_T$$

$$= G_T^*(\theta_\circ) + < G^*(\theta_\circ) >_T (\hat{\theta}_T - \theta_\circ) + o(< G^*(\theta_\circ) >_T),$$

and

$$(\hat{\theta}_T - \theta_\circ) = - < G^*(\theta_\circ) >_T^{-1/2} < G^*(\theta_\circ) >_T^{-1/2} G_T^*(\theta_\circ) + o(1).$$

Because of $\lambda_{min}(< G^*(\theta_\circ) >_T) \xrightarrow{a.s.} \infty$, every element of $< G^*(\theta_\circ) >_T^{-1/2}$ converges almost surely to 0. Then, by Condition 2 and some basic probability properties,

$$\hat{\theta}_T - \theta_\circ \xrightarrow{a.s.} 0, \qquad \text{as } n \to \infty.$$

Remark : 1. In practice, if Conditions 1 - 3 hold and the sequence of realizations of $\hat{\theta}_T$ converges, we may say that $\{\hat{\theta}_T\}$ converges to the true parameter.

2. Condition 3 is always true when Condition 2 holds.

3. Condition 2 can be weakened to "$< G^* >_T^{-1/2} G_T^* \xrightarrow{d} \mathcal{L}(X)$" for some X, where X is a *a.s.* finite r.v. and $\mathcal{L}(X)$ denotes the law of X.

<u>**Corollary**</u> **1** *Under Assumptions 5 and 6, if*

(1) G_T^* is asymptotic linear ,

(2) $< G^*(\theta) >_T^{-1/2} G_T^*(\theta) \xrightarrow{d} N(0, I)$, as $T \to \infty$,

(3) $< G^*(\theta) >_T$ is monotone, i.e. $\| < G^* >_T x\| \geq \| < G^* >_S x\|$ for all $x \in R^k$, when $T \leq S$,

(4) $\lambda_{min}(< G^*(\theta) >_T) \xrightarrow{a.s.} \infty$, as $T \to \infty$,

then $\hat{\theta}_T - \theta_\circ \xrightarrow{a.s.} 0$, as $T \to \infty$.

Proof: If $< G^*(\theta) >_T$ is monotonic, then

$$< G^*(\theta) >_T^{-1} G_T^*(\theta)$$

converges almost surely(see Chapter 8); that is

$$(\hat{\theta}_T - \theta_\circ) = -(\dot{G}_T^*(\theta_\circ))^{-1} G_T^*(\theta_\circ) + o(1)$$

converges almost surely. Following Theorem 2, $\hat{\theta}_T - \theta_\circ \to 0$, as $T \to \infty$.

As an application, we consider an AR model.

<u>**Theorem**</u> **4** *Suppose that $\{X_t\}$ is a stationary process and*

(1) $EX_0^2 < \infty$,

(2) $\{X_t\}$ *is an ergodic sequence,*

(3) $X_t = \alpha + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + \beta_k X_{t-k} + \varepsilon_k$,
where $E(\varepsilon_t|\mathcal{F}_{t-1}) = 0$, $E(\varepsilon_t^2|\mathcal{F}_{t-1}) = c$ *constant and* $\mathcal{F}_t = \sigma(X_1, \ldots, X_t)$,

(4) Let

$$A_T = \begin{pmatrix} t & \sum_1^t X_{s-1} & \ldots & \sum_1^t X_{s-k} \\ \sum_1^t X_{s-1} & \sum_1^t X_{s-1}^2 & \ldots & \sum_1^t X_{s-1}X_{s-k} \\ \vdots & \vdots & \vdots & \vdots \\ \sum_1^t X_{s-k} & \sum_1^t X_{s-k}X_{s-1} & \ldots & \sum_1^t X_{s-1}^2 \end{pmatrix},$$

and suppose that if there is a $\delta > 0$ *such that*

$$\lambda_{min} E A_T = O(|EA_T|^\delta).$$

Then the quasi-likelihood estimators

$$(\hat{\alpha}_t, \hat{\beta}_{t_1}, \ldots, \hat{\beta}_{t_k}) \xrightarrow{a.s.} (\alpha, \beta_1, \ldots, \beta_k).$$

Proof: By the assumptions,

$$\sum_1^t \varepsilon_t = \sum_1^t (X_s - \alpha - \beta_1 X_{s-1} - \ldots - \beta_k X_{t-k})$$

is a martingale with respect to $\{\mathcal{F}_t\}$. Let $\mathcal{G}_t$ be a subset of estimating function space $\{\sum_1^t f_t\varepsilon_t | \{f_t\}$ is a predictable process with respect to $\{\mathcal{F}_t\}\}$. Then the quasi-score estimating function is

$$G_t^* = -\frac{1}{c}\sum_1^t \begin{pmatrix} 1 \\ X_{s-1} \\ \vdots \\ X_{s-k} \end{pmatrix} (X_s - \alpha - \beta_1 X_{s-1} - \ldots - \beta_k X_{s-k}),$$

which is a linear function of $(\alpha, \beta_1, \ldots, \beta_k)'$.

We find that

$$< G^* >_t = \frac{1}{c}\sum_1^t \begin{pmatrix} 1 \\ X_{s-1} \\ \vdots \\ X_{s-k} \end{pmatrix} (1, X_{s-1}, \ldots, X_{s-k})$$

$$= \frac{1}{c}\begin{pmatrix} t & \sum_1^t X_{s-1} & \dots & \sum_1^t X_{s-k} \\ \sum_1^t X_{s-1} & \sum_1^t X_{s-1}^2 & \dots & \sum_1^t X_{s-1}X_{s-k} \\ \vdots & \vdots & \vdots & \vdots \\ \sum_1^t X_{s-k} & \sum_1^t X_{s-k}X_{s-1} & \dots & \sum_1^t X_{s-1}^2 \end{pmatrix}$$

$$= \dot{G}_t^*.$$

Following Conditions 1 - 2 and Theorem 4.3.1 of Révész(1968),

$$(E < G^* >_t)^{-1} < G^* >_t \xrightarrow{a.s.} I, \qquad \text{as } t \to \infty.$$

Applying Theorem 3 in Chapter 8,

$$< G^* >_t^{-1} G_t^* \xrightarrow{a.s.} 0, \qquad \text{as } t \to \infty.$$

Therefore,

$$\hat{\theta}_t - \theta_\circ \xrightarrow{a.s.} 0, \qquad \text{as } t \to \infty,$$

where

$$\hat{\theta}_t = \begin{pmatrix} \hat{\alpha}_t \\ \hat{\beta}_{t,1} \\ \vdots \\ \hat{\beta}_{t,k} \end{pmatrix}$$

$$= \begin{pmatrix} t & \sum_1^t X_{s-1} & \dots & \sum_1^t X_{s-k} \\ \sum_1^t X_{s-1} & \sum_1^t X_{s-1}^2 & \dots & \sum_1^t X_{s-1}X_{s-k} \\ \vdots & \vdots & \vdots & \vdots \\ \sum_1^t X_{s-k} & \sum_1^t X_{s-k}X_{s-1} & \dots & \sum_1^t X_{s-1}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_1^t X_s \\ \sum_1^t X_{s-1}X_s \\ \vdots \\ \sum_1^t X_{s-k}X_s \end{pmatrix}.$$

CHAPTER 4
ASYMPTOTIC CONFIDENCE INTERVALS

1 Introduction

In the previous chapters, several sufficient conditions for the consistency of quasi-likelihood estimators have been discussed. However we still need to ask some further questions. The first one is what kind of confidence zones we can recommend for the quasi-likelihood estimators. The second one is whether the quasi-score estimating function provides smaller asymptotic confidence intervals for the true parameters than other estimating functions .

In the following, our framework is that of a family of experiments defined on a statistical space $(\Omega, \mathcal{F}, \{P_\theta\})$ where $\theta \in \Theta$, an open subset of p-dimensional Euclidean space. The experiments are indexed by $t \in [0, \infty)$ and $\{\mathcal{F}_t\}$ denotes a standard filtration generated from these experiments. The framework covers discrete as well as continuous time and our object is to obtain asymptotic confidence statements about θ of maximum precision. The setting may be parametric or nonparametric; θ could be, for example, the mean of a stationary process.

A discussion concerning confidence zones of minimum size can be most naturally formulated in terms of properties of estimating functions ; see Heyde (1989) and references therein. To this end, and with little loss in generality, we shall confine attention to the class $\mathcal{G}$ of zero mean, square integrable, $\mathcal{F}_t$-measurable, semimartingale estimating functions $G_t(\theta)$ which are vectors of dimension p such that

$$\dot{G}_t(\theta) = (\frac{\partial G_{t,i}}{\partial \theta_j}),$$

$E\dot{G}_t(\theta)$ and $EG_t(\theta)G_t(\theta)'$ are nonsingular for each $t > 0$, the prime denoting transpose.

An effective choice of $G_t(\theta)$ is vital and in the case where the likelihood function exists, is known, and is tractable, the score function provides the benchmark and should ordinarily be used for G_t. It is well known, for example, that maximum likelihood (ML) estimation is associated with minimum size asymptotic confidence zones under suitable regularity conditions (e.g. Hall and Heyde (1980, Chapter 6)). We shall, however, take a more general approach, via a discussion of quasi-likelihood, for which properties similar to that of the ML estimator hold within what may be a more restricted setting.

In this chapter we shall point out what kind of confidence intervals we usually choose for the quasi-likelihood estimators and also the fact that the quasi-score estimating function always gives smaller asymptotic confidence intervals for the true parameter than other estimating functions under mild conditions.

2 Formulation

Asymptotic confidence statements about the " parameter" θ are based on the consideration of the asymptotic properties of a suitably chosen estimating function $G_T(\theta)$ as $T \to \infty$. For this purpose we shall assume, as is typically the case in regular problems of genuine physical relevance, that $G_T(\theta)$ has a limit distribution under some appropriate normalization.

Suppose that Ψ_T is a space of estimating functions , G_T is an estimating function in $\mathcal{G} \subseteq \Psi_T$ and θ_T^* is an estimator which satisfies $G_T(\theta_T^*) = 0$. If $\theta_t^* \xrightarrow{a.s.} \theta_\circ$, where $\theta_\circ$ is the true parameter, then for large T we use Taylor expansion to obtain

$$0 = G_T(\theta_T^*) = G_T(\theta_\circ) + \dot{G}_T(\theta_{1,T})(\theta_T^* - \theta_\circ) \tag{1}$$

where $\|\theta_\circ - \theta_{1,T}\| \leq \|\theta_\circ - \theta_T^*\|$, the norm denoting the positive square root of the sum of squares of elements. Then, assuming that, if there is a sequence of matrices $\{B_T\}$ such that

$$B_T G_T(\theta_\circ) \xrightarrow{d} X,$$

where X does not depend on the choice of G_T, then

$$B_T \dot{G}_T(\theta_\circ)(\theta_T^* - \theta_\circ) \xrightarrow{d} X, \tag{2}$$

by $\theta_t^* \xrightarrow{a.s.} \theta_\circ$ and $\dot{G}_T(\theta_{1,T})(\dot{G}_T(\theta_\circ))^{-1} \xrightarrow{p} I$, as $T \to \infty$. Theoretically, asymptotic confidence intervals for $\theta_\circ$ can be determined from (2) if the distribution of X is known. However, confidence zones based on (2) are mostly difficult to formulate unless X is normally distributed and it is desirable, wherever possible, to renormalize to obtain asymptotic normality. One important reason is that this is convenient in practice and it is easy to determine the boundary of the confidence interval. If, by using different norms, the estimators have an asymptotic normal distribution or an asymptotic mixed normal distribution respectively, we still prefer to use the norm which leads to the asymptotic normal distribution. This follows because the confidence interval based on asymptotic normality is smaller on average than that based on asymptotic mixed normality(see Heyde(1992)). That is why we always choose B_T such that

$$B_T G_T(\theta_\circ) \xrightarrow{d} N(0, I) \tag{3}$$

or

$$B_T \dot{G}_T(\theta_\circ)(\theta_T^* - \theta_\circ) \xrightarrow{d} N(0, I). \tag{4}$$

Since in practice Ψ_T is always chosen as a martingale space if the quasi-likelihood method is used, there are many sufficient conditions which can be used to check whether (3) holds. Some of these sufficient conditions can be found in Hall and Heyde (1980) and the bibliography therein. Although those results are mainly for the one dimensional case, it is easy to extend them to the higher dimensional case by using the Cramér-Wold Device which can be written as follows:

Suppose that the k-dimensional random vectors $X_n = (X_{n,1}, \ldots, X_{n,k})$ and $X = (X_1, \ldots, X_k)$ satisfy

$$\sum_{j=1}^{k} t_j X_{n,j} \xrightarrow{d} \sum_{j=1}^{k} t_j X_j$$

for each point $t = (t_1, \ldots, t_k)$ of R^k. Then the characteristic functions $f_n(s) = Eexp(is \sum_{j=1}^{k} t_j X_{n,j})$ of these one dimensional random variables converge to $f(s) = Eexp(is \sum_{j=1}^{k} t_j X_j)$ for each real s. Taking $s = 1$, we see that

$$Eexp(itX_n) \rightarrow Eexp(itX).$$

Since t is arbitrary, $X_n \xrightarrow{d} X$ follows by the continuity theorem for characteristic functions.

3 Central Limit Theorems

For the sake of completeness we now give two central limit theorems for multivariate martingales. We use them explicitly in Chapter 6.

The following theorem just extends Theorem 3.2 in Hall and Heyde(1980) to the higher dimensional case using the Cramér-Wold Device.

Theorem 1 *Suppose that $S_{n,k_n} = \sum_{k=1}^{k_n} X_{n,k}$ is a martingale in R^d, $d \geq 1$. If*

$$
\begin{array}{llll}
(1) & \max_k |X_{n,k}^{(j)}| \xrightarrow{p} 0, & j = 1, \ldots, d, & \text{as } n \to \infty \\
(2) & \sum_{k=1}^{k_n} X_{n,k} X_{n,k}' \xrightarrow{p} I, & & \text{as } n \to \infty \\
(3) & E \max_k (|X_{n.k}^{(i)} X_{n,k}^{(j)}|) & \text{is bounded in } n, &
\end{array}
$$

where $i, j = 1, \ldots, d$, then

$$S_{n,k_n} \xrightarrow{d} Z, \qquad \text{as } n \to \infty,$$

where $Eexp(it'Z) = exp(-\frac{1}{2}t't)$, $t \in R^d$.

Proof: Using the Cramér-Wold Device, we only need to prove, for each point $t = (t_1, \ldots, t_d) \in R^d$,

$$\sum_{j=1}^{d} t_j S_{n,k_n}^{(j)} \xrightarrow{d} \sum_{j=1}^{d} t_j Z_j, \qquad \text{as } n \to \infty.$$

But

$$\sum_{j=1}^{d} t_j S_{n,k_j}^{(j)} = \sum_{j=1}^{d} \sum_{k=1}^{k_n} X_{n,k}^{(j)} = \sum_{k=1}^{k_n} (\sum_{j=1}^{d} t_j X_{n,k}^{(j)})$$

and $\{\sum_{j=1}^{d} t_j X_{n,k}^{(j)}\}_{1 \leq k \leq k_n}$ is a martingale difference array. Using Theorem 3.2 in Hall and Heyde (1980), we only need to prove

$$\max_k |\sum_{j=1}^{d} t_j x_{n,k}^{j}| \xrightarrow{d} 0, \qquad \text{as } n \to \infty,$$

$$\sum_{k=1}^{k_n} (\sum_{j=1}^{d} t_j x_{n,k}^{(j)})^2 \xrightarrow{d} \sum_{j=1}^{d} t_j^2, \qquad \text{as } n \to \infty,$$

and

$$E(\max_k(\sum_{j=1}^d t_j X_{n,k}^{(j)})^2)$$

is bounded.

By the assumptions, we have

$$\max_k |\sum_{j=1}^d t_j X_{n,k}^{(j)}| \le \max_k \sum_{j=1}^d |t_j||Xn, k^{(j)}| \le \sum_{j=1}^d |t_j| \max_k |X_{n,k}^{(j)}| \xrightarrow{p} 0,$$

as $n \to \infty$, and,

$$\sum_{k=1}^{k_n} \begin{pmatrix} t_1 X_{n,k}^{(1)} \\ \vdots \\ t_d X_{n,k}^{(d)} \end{pmatrix} \begin{pmatrix} t_1 X_{n,k}^{(1)} & \dots & t_d X_{n,k}^{(d)} \end{pmatrix}$$

$$= \sum_{k=1}^{k_n} \begin{pmatrix} t_1^2 X_{n,k}^{(1)2} & \dots & t_1 t_d X_{n,k}^{(1)} X_{n,k}^{(d)} \\ \vdots & \vdots & \vdots \\ t_1 t_d X_{n,k}^{(1)} X_{n,k}^{(d)} & \dots & t_d^2 X_{n,k}^{(d)2} \end{pmatrix} \xrightarrow{p} \begin{pmatrix} t_1^2 & 0 & \dots & 0 \\ 0 & t_2^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & t_d^2 \end{pmatrix},$$

as $n \to \infty$. Then,

$$\sum_{k=1}^{k_n}(\sum_{j=1}^d t_j X_{n,k}^{(j)})^2 \xrightarrow{p} \sum_{j=1}^d t_j^2, \qquad \text{as } n \to \infty,$$

and

$$\begin{aligned} E(\max_k(\sum_{j=1}^d t_j X_{n,k}^{(j)})^2) &\le E(\sum_{i,j} \max_k |t_i t_j X_{n,k}^{(i)} X_{n,k}^{(j)}|) \\ &\le E(\sum_{i,j} |t_i t_j| \max_k |X_{n,k}^{(i)} X_{n,k}^{(j)}|) \end{aligned}$$

is bounded. Therefore

$$t' S_{n,k} \xrightarrow{p} t' Z, \qquad \text{as } n \to \infty,$$

i.e.

$$S_{n,k} \xrightarrow{p} Z, \qquad \text{as } n \to \infty,$$

where $Eexp(itz) = exp(-\frac{1}{2}t't)$.

A process X is considered as a mapping from $\Omega \times R_t$ into R^m. Each mapping $t \rightsquigarrow X_t(\omega)$, for a fixed $\omega \in \Omega$, is called a path, or a trajectory, of the process X. A process X is called cád (resp. cág, resp. cádlág), iff all its paths are right- continuous (resp. are left-continuous, resp. right-continuous and admit left- hand limits). Let $\{M_t\}$ be a square integrable martingale. The quadratic variation $[M]_t$ is defined by

$$[M]_t = < M^{cm} >_t + \sum_{0<s\leq t} (\triangle M_t)(\triangle M_t)',$$

where M^{cm} is the unique continuous martingale part of M, $< M^{cm} >$ is a predictable integrable process which satisfies that $(M^{cm})^2 - < M^{cm} >$ is a uniformly integrable martingale with initial value zero and

$$\triangle M_s = M_s(\theta) - M_{s-}(\theta);$$

see e.g. Rogers and Williams(1987, p391) for a discussion.

We shall now give a central limit theorem which extends Theorem 1.1 in Hutton and Nelson(1984) to the higher dimensional case. They also gave a rather different extension of that theorem in their Corollary 2.3. Let $\mathcal{N}(\mathcal{N}^+)$ denote the set of nonnegative (positive) definite symmetric matrices. If $A \in \mathcal{N}$, let $A^{1/2}$ be the unique, nonnegative definite square root of A and let $\lambda_{min}(A)$ denote the minimum eigenvalue of A. For any matrix $A = (a_{i,j})$, let

$$\|A\|^2 = \sum_{i,j} a_{i,j}^2.$$

Hutton and Nelson(1984)(Corollary 2.3) gave the following result:

Let $M = (M_t, \mathcal{F}_t, 0 \leq t < \infty)$ be a $p \times 1$ square integrable martingale with quadratic variation matrix $[M]_t$. Set $B_T = E(M_T M_T')$ and assume that $\lambda_{min}(B_T) \to \infty$, as $T \to \infty$. Suppose there is a function K_t increasing to infinity as t tends to infinity such that

(1) $K_T^{-1/2} \sup_{t\leq T} \|\triangle M_t\| \xrightarrow{p} 0$, as $t \to \infty$,

(2) $K_T^{-1}[M]_T \xrightarrow{p} \eta^2$, as $T \to \infty$, where η^2 is an element(possibly random) of $\mathcal{N}$,

(3) $K_T^{-1} B_T \to \sum$, as $T \to \infty$, where $\sum$ is an element of $\mathcal{N}^+$.

Then

$$K_T^{-1/2} M_T \xrightarrow{d} Z^* \text{ (stably)}$$

where Z^* has the characteristic function $E(exp(-1/2\lambda'\eta^2\lambda))$ and $\lambda = (\lambda_1, \ldots, \lambda_p)'$ is a column vector.

We can see that, in Hutton and Nelson's result, they require that the covariance matrix B_T be controlled by an increasing function K_t. In a sense, this is like requiring that $\lambda_{max}(B_T) = O(\lambda_{min}(B_T))$, which is a strong condition. Sørensen(1988) later extended the Hutton and Nelson's (1984) result by substituting a vector function $K_t = (K_{1,t}, \ldots, K_{n,t})$ for the increasing function K_t. He used a diagonal matrix as a normalizer for B_T. This is a significant restriction on the covariance matrix B_T. But, we would wish to find conditions just based on the covariance matrix B_T alone. Furthermore, in both Hutton and Nelson's and Sørensen's results, they only got $K_t^{-1/2} M_t \xrightarrow{d} Z$. However when we use the quasi-likelihood method to estimate parameters, we often meet a question such as under which conditions we have

$$B_t^{-1/2} M_t \xrightarrow{d} Z$$

or

$$< M >_t^{-1/2} M_t \xrightarrow{d} Z$$

or

$$[M]_t^{-1/2} M_t \xrightarrow{d} Z.$$

Therefore we need other forms of the central limit theorem which focus on these issues. Such a result is given below.

Theorem 2 *Let a cág process $M = (M_{(1)}, M_{(2)}, \ldots, M_{(m)})$ be an m-dimensional square integrable $\{\mathcal{F}_t\}$-measurable martingale with quadratic variation matrix $[M]_t$ and set $B_t = E(M_t M_t')$. If*

(1) $\lambda_{min}(B_t) \to \infty$, as $t \to \infty$,

(2) $\sup_{s \le t} x' B_t^{-1/2} \Delta M_s \Delta M_s' B_t^{-1/2} x \xrightarrow{p} 0$, as $t \to \infty$, for all $x \in R^m$,

(3) $B_t^{-1/2} [M]_t B_t^{-1/2} \xrightarrow{p} A$,

where A is a $m \times m$ random positive definite matrix,

then

$$B_t^{-1/2} M_t \xrightarrow{d} Z,$$

where Z has the characteristic function $E(exp(-1/2\lambda' A\lambda))$.

Proof: By the Cramér-Wold Device, we only need show that, for all $x \in R^m$, $x'x \neq 0$,

$$x' B_t^{-1/2} M_t \xrightarrow{d} x' Z, \qquad \text{as } t \to \infty. \tag{5}$$

Since $\lambda_{min}(B_t) \to \infty$, as $t \to \infty$. We only need to prove (5) for any sequence $\{T_n\}$ which satisfies

$$|x' B_{T_{n+1}}^{-1/2} B_{T_n}^{1/2} (xx')^{-1} x| \leq 2^{-2}. \tag{6}$$

For every T_n fixed, let $M_{T_n,0} = M_0 = 0$,

$$\begin{aligned} T_{n,k+1} &= \inf\{t > T_{n,k} | \quad |x' B_{T_n}^{-1/2}(M_t - M_{T_{n,k}})| > 2^{-n}\} \\ &= T_n, \qquad \text{if no such } t \leq T_n \text{ exists,} \end{aligned}$$

and

$$X_{n,k} = x' B_{T_n}^{-1/2}(M_{T_n,k} - M_{T_n,k-1}).$$

By the definition of $\{M_t\}$ and $T_{n,k}$, we have

$$|X_{n,k}| \leq 2^{-n}.$$

Therefore

$$\sup_k E|\sum X_{n,k}|^2 = \sup_k E|\sum x' B_{T_n}^{-1/2}(M_{T_n,k} - M_{T_n,k-1})|^2$$

$$= \sup_k E(x' B_{T_n}^{-1/2} M_{T_n} M'_{T_n} B_{T_n}^{-1/2} x) = xx' < \infty;$$

$$\sup_k |X_{n,k}| \leq 2^{-n} \to 0, \qquad \text{as } n \to \infty;$$

$$E(\sup_k X_{n,k}^2) \leq 2^{-2n}, \qquad \text{is bounded in } n.$$

By the definition of T_n, we have

$$|x' B_{T_{n+1}}^{-1/2}(M_{T_n,1} - M_0)|$$

$$= |x' B_{T_{n+1}}^{-1/2} B_{T_n}^{1/2} (xx')^{-1} x x' B_{T_n}^{-1/2}(M_{T_n,1} - M_0)|$$

$$\leq 2^{-n-2} < 2^{-n-1}.$$

This means $T_{n,1} < T_{n+1,1}$. Now assume that $T_{n,k-1} \leq T_{n+1,k-1}$ is true. We are going to prove $T_{n,k} \leq T_{n+1,k}$. If $T_{n,k} \leq T_{n+1,k-1}$, then it is obvious that $T_{n,k} \leq T_{n+1,k}$; if $T_{n,k} > T_{n+1,k-1}$, by the definition of $T_{n,k}$ and applying (6), we have

$$|x'B_{T_{n+1}}^{-1/2}(M_{T_{n,k}} - M_{T_{n+1,k-1}})|$$

$$\leq |x'B_{T_{n+1}}^{-1/2}(M_{T_{n,k}} - M_{T_{n,k-1}})| + |x'B_{T_{n+1}}^{-1/2}(M_{T_{n+1,k-1}} - M_{T_{n,k-1}})|$$

$$\leq 2^{-n-2} + 2^{-n-2} = 2^{-n-1},$$

i.e. $T_{n,k} \leq T_{n+1,k}$. By induction, we get $T_{n,k} \leq T_{n+1,k}$ for all k and therefore $\mathcal{F}_{n,k} \subseteq \mathcal{F}_{n+1,k}$. Using the integral expression,

$$M_{T_n}M'_{T_n} = \int_0^{T_n} M_{s-}dM'_s + \int_0^{T_n} dM_sM'_s + [M, M']_{T_n}$$

(see, Liptser and Shiryaev (1986), p118) and the fact that

$$\sum X_{n,k}X'_{n,k}$$

$$= x'B_{T_n}^{-1/2}\sum(M_{T_{n,k}}M'_{T_{n,k}} - M_{T_{n,k}}M_{T_{n,k-1}} - M_{T_{n,k-1}}M'_{T_{n,k}}$$
$$+M_{T_{n,k-1}}M'_{T_{n,k-1}})B_{T_n}^{-1/2}x$$
$$= x'B_{T_n}^{-1/2}(M_{T_n}M'_{T_n} - \sum(M_{T_{n,k}} - M_{T_{n,k-1}})M'_{T_{n,k-1}}$$
$$-\sum M_{T_{n,k-1}}(M'_{T_{n,k}} - M'_{T_{n,k-1}}))B_{T_n}^{-1/2}x,$$

we have

$$x'B_{T_n}^{-1/2}[M]_{T_n}B_{T_n}^{-1/2}x - \sum X_{n,k}X'_{n,k}$$
$$= 2x'B_{T_n}^{-1/2}(\sum(M_{T_{n,k}} - M_{T_{n,k-1}})M'_{T_{n,k-1}} - \int_0^{T_n} dM_sM'_s)B_{T_n}^{-1/2}x$$
$$= x'B_{T_n}^{-1/2}(\sum(\int_{T_{n,k-1}}^{T_{n,k}} dM_s(M'_{T_{n,k-1}} - M'_{s-}))B_{T_n}^{-1/2}x$$
$$= x'B_{T_n}^{-1/2}\int_0^{T_n} dM_sL_{T_n}^{(s)}B_{T_n}^{-1/2}x,$$

where $L_{T_n}^{(s)} = \sum(M'_{T_{n,k-1}} - M'_s)I_{(T_{n,k-1},T_{n,k})}$. Since

$$P(|x'B_{T_n}^{-1/2}\int_0^{T_n} dM_sL_{T_n}^{(s)}B_{T_n}^{-1/2}x| > \varepsilon)$$

$$\leq 1/\varepsilon^2 E(tr(\int_0^{T_n} L_{T_n}^{(s)}B_{T_n}^{-1/2}xx'B_{T_n}^{-1/2}dM_s)^2)$$

$$\leq 1/\varepsilon^2 tr(E\int_0^{T_n} L_{T_n}^{(s)} B_{T_n}^{-1/2} xx' B_{T_n}^{-1/2} d[M]_s B_{T_n}^{-1/2} xx' B_{T_n}^{-1/2} L_{T_n}^{(s)\prime})$$

$$\leq 1/\varepsilon^2 2^{-2n} tr(E(x' B_{T_n}^{-1/2}[M]_{T_n} B_{T_n}^{-1/2} x))$$

$$\leq K2^{-2n}/\varepsilon^2 \to 0, \qquad \text{as } n \to \infty,$$

for some constant K, this forces

$$x' B_{T_n}^{-1/2}[M]_{T_n} B_{T_n}^{-1/2} x - \sum X_{n,k} X'_{n,k} \xrightarrow{p} 0.$$

Then, by assumption, we have

$$\sum X_{n,k} X'_{n,k} \xrightarrow{p} x'Ax, \qquad \text{as } n \to \infty.$$

Applying Lemma 2.2 in Hutton and Nelson (1984),

$$\sum X_{n,k} \xrightarrow{p} Z^*, \qquad \text{as } n \to \infty,$$

where $Eexp(i\lambda Z^*) = Eexp(-1/2\lambda x' A x \lambda)$. Therefore

$$\sum X_{n,k} \xrightarrow{p} x'z, \qquad \text{as } n \to \infty,$$

where $Eexp(i\vec{\lambda} z) = Eexp(-1/2\vec{\lambda} A \vec{\lambda})$, and (5) holds.

4 Asymptotic Confidence Intervals

Following the discussion of (1) in Section 2 of this Chapter, if

$$(EG_T(\theta_0)G_T(\theta_0)')^{-1/2}\dot{G}_T(\theta_{1,T})(E\dot{G}_T(\theta_{1,T}))^{-1}E(G_T(\theta_0)G_T(\theta_0)')^{1/2} \xrightarrow{p} YI_p$$

for some random variable Y (> 0 a.s.), we have, as $T \to \infty$,

$$(EG_T(\theta_\circ)G_T(\theta_\circ)')^{-1/2}(E\dot{G}_T(\theta_{1,T}))(\theta_T^* - \theta_\circ) \xrightarrow{d} X, \tag{7}$$

say, not depending on the choice of G_t. In this case the size of confidence zones for $\theta_\circ$ is then governed by the scaling " information "

$$\mathcal{E}(G_t(\theta)) = (E\dot{G}_t(\theta))'(E(G_t(\theta)G_t(\theta))')^{-1}(E\dot{G}_t(\theta))$$

and we prefer estimating function $G_{1,t}$ to $G_{2,t}$ if $\mathcal{E}(G_{1,t}) \geq \mathcal{E}(G_{2,t})$ for each $t \geq t_0$ in the partial order of nonnegative definite matrices as indicated in Chapter 1.

For $G_t(\theta) \in \mathcal{G}$, the result

$$[G_t]^{-1/2} G_t \xrightarrow{d} N_p(0, I_p), \tag{8}$$

N_p denoting the p-variate normal, seems to encapsulate the most general form of asymptotic normality result. Norming by a random process such as $[G]_t$ is essential in what is termed the non-ergodic case(for which $(E[G]_t)^{-1}[G]_t \overset{p}{\not\to}$ constant as $t \to \infty$). A simple example of this furnished by the pure birth process N_t with intensity θN_{t-}, where we take for G_t the estimating function and

$$G_t = \theta^{-1}(N_t - 1) - \int_0^t N_{s-} ds,$$

$$[G]_t = \theta^{-2}(N_t - 1),$$

while

$$[G]_t^{-1/2} G_t \xrightarrow{d} N(0, I), \quad (EG_t^2)^{-1/2} G_t \xrightarrow{d} W^{1/2} N(0,1)$$

and

$$(E[G]_t)^{-1}[G]_t \xrightarrow{a.s.} W$$

where W has a gamma distribution with form parameter N_0 and shape parameter N_0 and $W^{1/2}N(0,1)$ is distributed as the product of independent $W^{1/2}$ and $N(0,1)$ variables.

In the general case, to obtain confidence zones for θ_0 from (2), we use the Taylor expansion (1) and then, under appropriate continuity conditions for $\dot{G}_t$,

$$\dot{G}_T(\theta_{1,T})(\dot{G}_T(\theta_0))^{-1} \xrightarrow{p} I_p$$

and,

$$[G(\theta)]_T^{-1/2} \dot{G}_T(\theta_T)(\theta_T^* - \theta_0) \xrightarrow{d} N_p(0, I_p) \tag{9}$$

as $T \to \infty$, if (8) holds.

For the construction of confidence intervals we actually need this convergence to be uniform in compact intervals of θ_0; we shall use $\xrightarrow{u.d.}$ to denote such convergence and henceforth suppose that (9) holds in this mode. We shall also write

$$\bar{\mathcal{E}}(G_t(\theta_0)) = (-\dot{G}_t(\theta_0))'[G(\theta)]_t^{-1}(-\dot{G}_t(\theta_0)). \tag{10}$$

Which of the asymptotically equivalent forms should be used for asymptotic confidence zones based on (9) is unclear in general despite the fact that many special investigations have been conducted. For example, in the case of likelihoods rather than quasi-likelihoods there is evidence to support the use of the observed information $-\dot{G}_t(\theta)$ rather than the expected information $EG_t(\theta)G_t(\theta)'$(e.g. Efron and Hinkley(1978)). Barndorff-Nielsen and Sørensen(1989) have used a number of examples to argue the advantages of (10) which, as we have seen, comes naturally from (8). On the other hand, Heyde(1987) proposed the form

$$(-\bar{G}_t(\theta))' < G(\theta) >_t^{-1} (-\bar{G}_t(\theta))$$

where $\bar{G}_t(\theta)$ is a matrix of predictable processes such that $\dot{G}_t(\theta) - \bar{G}_t(\theta)$ is a martingale. This is a direct extension of the classical Fisher information and is closely related to $\mathcal{E}(G_t(\theta))$.

In (1) let G_T be a quasi-score estimating function in $\mathcal{G}$, denoted by G_T^*. Here we consider those case in which (4) hold for some B_T. If (4) holds, then the asymptotic confidence intervals for $\theta_\circ$ can be determined by quasi-score estimating function in the following two ways:

(1)If $B_t\dot{G}_T^*$ are nonrandom, then

$$\theta_T^* - \theta_\circ \overset{d}{\approx} N(0, (\dot{G}_T^*)^{-1}(B_T'B_T)^{-1}(\dot{G}_T^{*\prime})^{-1})$$

for large T and the $(1-\alpha)\times 100\%$ asymptotic confidence intervals for $c'\theta_\circ$, c being any vector in R^p if θ_T^*, $\theta_\circ \in R^p$, are

$$c'\theta_T^* \pm z_{\alpha/2}c'(\dot{G}_T^*)^{-1}(B_T'B_T)^{-1}(\dot{G}_T^{*\prime})^{-1}c,$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ point of the standard normal distribution.

(2) If $B_t\dot{G}_T^*$ are random, then

$$(\theta_T^* - \theta_\circ)'\dot{G}_T^{*\prime}B_T'B_T\dot{G}_T^*(\theta_T^* - \theta_\circ) \overset{d}{\approx} \chi_p^2$$

for large T, if $\theta_T^*, \theta_\circ \in R^p$. Therefore the $(1-\alpha) \times 100\%$ asymptotic confidence intervals for $\theta_\circ$ can be determined from

$$\begin{aligned} &P((\theta_T^* - \theta_\circ)'\dot{G}_T^{*\prime} B_T' B_T \dot{G}_T^*(\theta_T^* - \theta_\circ) \leq \chi^2_{p,\alpha}) \\ &= P(\max_{c \in R_p} \frac{|\,(c'(\theta_T^* - \theta_\circ))\,|^2}{c'(\dot{G}_T^{*\prime} B_T' B_T \dot{G}_T^*)^{-1} c} \leq \chi^2_{p,\alpha}) \\ &\approx 1 - \alpha, \end{aligned} \tag{11}$$

where $\chi^2_{p,\alpha}$ is the upper α point of the chi-squared distribution with p degrees of freedom. The result (11) can be written again as follows:

$$\begin{aligned} &P((\theta_T^* - \theta_\circ)'\dot{G}_T^{*\prime} B_T' B_T \dot{G}_T^*(\theta_T^* - \theta_\circ) \leq \chi^2_{p,\alpha}) \\ &= P(c'\theta_\circ \in c'\theta_T^* \pm \chi_{p,\alpha}[c'(\dot{G}_T^{*\prime} B_T' B_T \dot{G}_T^*)^{-1} c]^{1/2}, \quad \text{for all } c \in R_p) \\ &= P((\theta_T^* - \theta_\circ)(\theta_T^* - \theta_\circ)' - \chi^2_{p,\alpha}(\dot{G}_T^{*\prime} B_T' B_T \dot{G}_T^*)^{-1} \quad \text{is negative definite}) \\ &\approx 1 - \alpha. \end{aligned}$$

An important question to ask is whether the quasi-score estimating function can provide a smaller asymptotic confidence zones for $\theta_\circ$ than any other estimating functions (assuming always that the quasi-likelihood estimators θ_T^* converges almost surely to the true parameter $\theta_\circ$). For the general case, it is too hard to answer this question. Firstly, since a quasi-score estimating function is associated with an estimating function space, it is difficult to compare a quasi-score estimating function in one estimating function space with an estimating function in another estimating function space. Secondly, even for two estimating functions, $G_{n,1}$ and $G_{n,2}$, from the same estimating function space, and assuming that the Taylor expansions

$$0 = G_{n,1}(\theta_{n,1}) = G_{n,1}(\theta_\circ) + \dot{G}_{n,1}(\tilde{\theta}_{n,1})(\theta_{n,1} - \theta_\circ)$$

and

$$0 = G_{n,2}(\theta_{n,2}) = G_{n,2}(\theta_\circ) + \dot{G}_{n,2}(\tilde{\theta}_{n,2})(\theta_{n,2} - \theta_\circ)$$

hold, if there are norms, say $B_{n,1}$ and $B_{n,2}$, such that

$$B_{n,1} G_{n,1}(\theta_\circ) \xrightarrow{d} X_1$$

and

$$B_{n,2} G_{n,2}(\theta_\circ) \xrightarrow{d} X_2,$$

where X_1 and X_2 belong to different kinds of distribution laws, it is still difficult to compare the confidence intervals for $\theta_\circ$. Therefore, we just focus our attention on a fixed estimating functions space Ψ and assume that the estimators have the asymptotic normality property. Before discussing the above question, we need the following theorem:

Theorem 3 *Suppose that* $\mu = \{\{B_n\}|\ B_n \quad \text{is nonsingular}, n = 1, 2, \ldots.\}$ *is a subset of sequences of matrices and* $\theta_\circ$ *is the true parameter. If*

$$\begin{array}{ll} (1) & B_n(\theta_n - \theta_\circ) \xrightarrow{d} N(0, I), \\ (2) & B_n^*(\theta_n^* - \theta_\circ) \xrightarrow{d} N(0, I), \\ (3) & B_n^{-1}(B_n')^{-1} - (B_n^*)^{-1}(B_n^{*\prime})^{-1} \geq 0 \quad \textit{for all } n, \end{array}$$

then the asymptotic confidence zones determined by B_n^* *are smaller than those determined by* B_n.

Proof: If B_n and B_n^* are non-random, then for large n,

$$\theta_n - \theta_\circ \overset{d}{\approx} N(0, B_n^{-1}(B_n')^{-1}),$$

$$\theta_n^* - \theta_\circ \overset{d}{\approx} N(0, (B_n^*)^{-1}(B_n^{*\prime})^{-1},$$

and the theorem is true. For the general case, if

$$B_n(\theta_n - \theta_\circ) \xrightarrow{d} N(0, I)$$

and

$$B_n^*(\theta_n^* - \theta_\circ) \xrightarrow{d} N(0, I)$$

then the $(1 - \alpha) \times 100\%$ asymptotic confidence intervals for $\theta_\circ$ associated with B_n and B_n^* are determined by

$$P((\theta_n - \theta_\circ)(\theta_n - \theta_\circ)' - \chi_\alpha^2(B_n'B_n)^{-1} \quad \text{is negative definite})$$

and

$$P((\theta_n^* - \theta_\circ)(\theta_n^* - \theta_\circ)' - \chi_\alpha^2(B_n^{*\prime}B_n^*)^{-1} \quad \text{is negative definite})$$

respectively. But $(B_n^{*\prime}B_n^*)^{-1} \leq (B_n'B_n)^{-1}$. Thus the theorem is true.

Let $\mathcal{G}_T$ be a subset of estimating function space Ψ_T and G_T^* be the quasi-score estimating function in $\mathcal{G}_T$. Suppose that $G_T \in \mathcal{G}_T$, θ_T and θ_T^* satisfy the conditions $G_T(\theta_T) = 0$ and $G_T^*(\theta_T^*) = 0$ respectively and $\theta_\circ$ is the true parameter. Using the notations in Godambe and Heyde(1987), if some suitable conditions are satisfied and

$$\langle G\rangle_T^{-1/2}\bar{G}_T(\theta_T - \theta_\circ) \xrightarrow{d} N(0, I), \tag{12}$$

$$\langle G^*\rangle_T^{-1/2}\bar{G}_T^*(\theta_T^* - \theta_\circ) \xrightarrow{d} N(0, I), \tag{13}$$

then, following Theorem 3,

$$(\bar{G}_T)^{-1}\langle Q\rangle_T((\bar{G}_T)^{-1})' - (\bar{G}_T^*)^{-1}\langle G^*\rangle_T((\bar{G}_T^*)^{-1})' \geq 0$$

implies that the asymptotic confidence zones determined by $\langle G^*\rangle_T^{-1/2}\bar{G}_T^*$ are smaller than those determined by $\langle G\rangle_T^{-1/2}\bar{G}_T$.

Let $\Psi_T = \{\int_0^T \alpha_s(\theta)dm_s(\theta)\}$ be an estimating function space , where $\{\alpha_s(\theta)\}$ is a predictable process which is almost surely continuously differentiable with respect to the components of θ, and

$$m_T(\theta) = X_T - \int_0^T f_t(\theta)d\lambda_t, \tag{14}$$

where $\{\lambda_t\}$ is a real, monotonically increasing, right continuous process with $\lambda_0 = 0$, while $\{m_T(\theta), \mathcal{F}_T\}$ is a cádlág, square integrable $r \times 1$ vector martingale with quadratic characteristic $\langle m(\theta)\rangle_T$ given, for all T, by

$$\langle m(\theta)\rangle_T = \int_0^T a_t(\theta)d\lambda_t.$$

The processes $\{\lambda_t\}$, $\{a_t(\theta)\}$ and $\{f_t(\theta)\}$ are predictable and the elements of $f_t(\theta)$ are almost surely continuously differentiable with respect to the element of θ. For the above model Godambe and Heyde(1987) indicated that the quasi-score estimating function G_T^* in $\mathcal{G}_T \subseteq \Psi_T$ is both O_A and O_F optimal(for definitions see Godambe and Heyde(1987)). Therefore, in this case, if (12) and (13) are true, we should choose a quasi-score estimating function G_T^* to estimate the true parameter, because it can provide smaller asymptotic confidence intervals than others. In essence, Theorem 3 shows that we can compare the asymptotic confidence intervals determined by

B_n and B_n^* respectively if $(B_n'B_n)^{-1}$ and $(B_n^{*\prime}B_n^*)^{-1}$ can be ordered. Naturally this statement will lead to another question. If there are sequences of matrices $\{\tilde{A}_n\}$, $\{A_n\}$ and $\{B_n\}$ satisfying

$$\tilde{A}_n(\theta_n - \theta_\circ) \xrightarrow{d} N(0, I),$$

$$A_n(\theta_n - \theta_\circ) \xrightarrow{d} N(0, I),$$

$$B_n(\theta_n^* - \theta_\circ) \xrightarrow{d} N(0, I),$$

and $(A_n'A_n)^{-1} \geq (B_n'B_n)^{-1}$, can we also get $(\tilde{A}_n'\tilde{A}_n)^{-1} \geq (B_n'B_n)^{-1}$ or O_A optimality. If the answer is "yes", it means that, for every sequence of parameters $\{\theta_n\}$, if there is one kind of norm $\{A_n\}$ such that

$$A_n(\theta_n - \theta_\circ) \xrightarrow{d} N(0, I)$$

and

$$(A_n{}'A_n)^{-1} \geq (B_n{}'B_n)^{-1},$$

it is impossible to find another kind of norm $\{\tilde{A}_n\}$ such that

$$\tilde{A}_n(\theta_n - \theta_\circ) \xrightarrow{d} N(0, I)$$

and

$$(B_n{}'B_n)^{-1} > (\tilde{A}_n'\tilde{A}_n)^{-1}.$$

In this case we can claim that the quasi-score estimating function is the best estimating function from the estimating function space regardless of what kind of norm we have chosen. Unfortunately it seems impossible to give a complete answer for this question. Nevertheless the following theorem will give a partial answer.

Theorem 4 *Suppose that $\{\theta_n\}$ is a sequence of estimators, and $\{B_n\}$ and $\{A_n\}$ are nonsingular matrices (possible random). If*

$$B_n(\theta_n - \theta_\circ) \xrightarrow{d} N(0, I),$$

$$A_n(\theta_n - \theta_\circ) \xrightarrow{d} N(0, I),$$

and

$$A_n B_n^{-1} \xrightarrow{a.s.} C,$$

as $n \to \infty$, *where* C *is a nonsingular non-random matrix, then*

$$C'C = I.$$

Proof: Denote $B_n(\theta_n - \theta_\circ)$ and $A_n B_n^{-1}$ by X_n and C_n respectively. Following the assumptions,

$$X_n \xrightarrow{d} N(0, I),$$

$$C_n X_n \xrightarrow{d} N(0, I),$$

and

$$C_n \xrightarrow{a.s.} C.$$

We denote by $O_x(\varepsilon)$ a subset of R^p which is a ball with centre at x and radius ε. Suppose that E is a subset of R^p, $E + \varepsilon = \{\bigcup_{x \in E} O_x(\varepsilon)\}$. Let E_1 and E_2 be subsets of R^p, $d(E_1, E_2) = \min\{\varepsilon |\ E_1 + \varepsilon \supset E_2 \textit{ and } E_2 + \varepsilon \supset E_1\}$.

For any sets $A = ((-\infty, a_1], \cdots, (-\infty, a_p]) \in R^p$, we have

$$P(C_n X_n \in A) = P(X_n \in C_n^{-1} A)$$

$$= P(X_n \in C_n^{-1} A) - P(X_n \in C^{-1} A) + P(X_n \in C^{-1} A).$$

Since $C_n \xrightarrow{a.s.} C$, as $n \to \infty$, then $d(C_n^{-1} A, C^{-1} A) \xrightarrow{a.s.} 0$. Using Egoroff's theorem, for any $\delta > 0$, there exists $D(\delta)$ such that $P(\Omega - D(\delta)) < \delta$ and $d(C_n^{-1} A, C^{-1} A) \xrightarrow{u} 0$ on $D(\delta)$, where $\xrightarrow{u}$ denotes uniform convergence.

Therefore, for any $\varepsilon > 0$, there exists $N(\delta, \varepsilon) > 0$ such that, if $n \geq N$,

$$C_{\cdot}^{-1} A + \varepsilon \supseteq C_n^{-1} A \supseteq C^{-1} A - \varepsilon$$

where $C^{-1} A - \varepsilon$ is a subset of $C^{-1} A$ such that $(C^{-1} A - \varepsilon) + \varepsilon \subseteq C^{-1} A$. Thus

$$P(C_n X_n \in A) - \delta - P(X_n \in (C^{-1} A + \varepsilon) \backslash (C^{-1} A - \varepsilon))$$

$$\leq P(X_n \in C^{-1} A) \leq P(C_n X_n \in A) + \delta + P(X_n \in (C^{-1} A + \varepsilon) \backslash (C^{-1} A - \varepsilon)).$$

Since $X_n \xrightarrow{d} N(0,1)$, first let $n \to \infty$, then $\varepsilon \to 0$ and $\delta \to 0$. We get

$$\lim_{n\to\infty} P(X_n \in C^{-1}A) = \lim_{n\to\infty} P(CX_n \in A) = \lim_{n\to\infty} P(C_nX_n \in A).$$

But $C_nX_n \xrightarrow{d} N(0,I)$, which also implies $CX_n \xrightarrow{d} N(0,I)$. Combining

$$X_n \xrightarrow{d} N(0,I)$$

and

$$CX_n \xrightarrow{d} N(0,I).$$

we get

$$C'C = I.$$

Remark : Because of the existence of C^{-1}, $C'C = I$ implies that $CC' = I$,

Now we continue our explanation. Suppose that (12) and (13) are true and, in the meanwhile there are other matrices $\{B_T\}$ such that

$$B_T(\theta_T - \theta_\circ) \xrightarrow{d} N(0,I)$$

and

$$< G >_T^{-1/2} \bar{G}_T B_T^{-1} \xrightarrow{a.s.} C, \qquad \text{as } T \to \infty,$$

where C is a non-random matrix. Due to Theorem 4, we obtain $CC' = I$. Then

$$B_T^{-1}B_T^{-1\prime} - \bar{G}_T^{*-1} < G^* >_T (\bar{G}_T^{*-1})'$$

$$= (< G >_T^{-1/2} \bar{G}_T)^{-1}(< G >_T^{-1/2} \bar{G}_T)B_T^{-1}B_T^{-1\prime}(< G >_T^{-1/2} \bar{G}_T)'((< G >_T^{-1/2} \bar{G}_T)')^{-1}$$
$$-\bar{G}_T^{*-1} < G^* >_T (\bar{G}_T^{*-1})'$$
$$= (< G >_T^{-1/2} \bar{G}_T)^{-1}[(< G >_T^{-1/2} \bar{G}_T)B_T^{-1}B_T^{-1\prime}(< G >_T^{-1/2} \bar{G}_T)' - I]$$
$$\times((< G >_T^{-1/2} \bar{G}_T)')^{-1} + \bar{G}_T^{-1} < G >_T (\bar{G}_T^{-1})' - \bar{G}_T^{*-1} < G^* >_T (\bar{G}_T^{*-1})'$$

is asymptotically negative-definite if $\|(< G >_T^{-1/2} \bar{G}_T)^{-1}\|$ is bounded for large enough T. Therefore, in the asymptotic sense, $< G^* >_T^{-1/2} \bar{G}_T^{*-1}$ also provides smaller asymptotic confidence intervals than those provided by $\{B_T\}$.

CHAPTER 5
ASYMPTOTIC QUASI-SCORE ESTIMATING FUNCTIONS

1 Introduction

Given a process X, if there is a parameter $\theta \in \Theta \subseteq R^n$ associated with X, which needs to be estimated, this can usually be done using the quasi-likelihood method. Let Ψ_T be an estimating function space associated with the sample $\{X_t, t \leq T\}$, Θ be a parameter space and $\mathcal{G}_T$ be a subset of Ψ_T. The element G_T of $\mathcal{G}_T$ satisfies the conditions that $E_\theta G_T G_T{}'$ and $E_\theta \dot{G}_T$ are nonsingular. If there exists a $G_T^* \in \mathcal{G}_T$ such that

$$(E_\theta \dot{G}_T)^{-1} E_\theta G_T G_T{}' (E_\theta \dot{G}_T')^{-1} - (E_\theta \dot{G}_T^*)^{-1} E_\theta G_T^* G_T^*{}' (E_\theta \dot{G}_T^*{}')^{-1}$$

is nonnegative for all $G_T \in \mathcal{G}_T$ and all $\theta \in \Theta$, then G_T^* is called the quasi-score estimating function in $\mathcal{G}_T$ and the root of $G_T^*(\theta) = 0$ is called the quasi-likelihood estimator. However, in practice it is possible that a quasi-score estimating function in $\mathcal{G}_T$ does not exist or sometimes it is difficult to obtain the quasi-likelihood estimator from the quasi-score estimating equation $G_T^*(\theta) = 0$ since the representation of $G_T^*(\theta)$ may be complicated, especially when the problem involves some nuisance parameters. For instance, an example which we shall provide below gives us a quasi-score estimating function

$$G_T^* = \sum_{k-1}^{T} X_{k-1}(\sigma^2 X_{k-1} + \eta^2 X_{k-1}^2)^{-1}(X_k - \theta X_{k-1}),$$

where σ^2 and η^2 are nuisance parameters. From the form of G_T^*, it ~~It~~ is not possible to get a ~~the~~ quasi-likelihood estimator of θ ~~from the quasi-score estimating equation when~~ which depends only on $\{X_k : k \leq T\}$ if σ^2 and η^2 are unknown. Therefore we need a new estimating function to substitute for the quasi-score estimating function which retains the same basic properties. For this purpose, Heyde and Gay(1989) introduced a new concept of asymptotic quasi-score estimating functions. We shall further discuss this important concept here.

In the following discussion, all expectations are functions of $\theta \in \Theta$, i.e. the expectation should be written as "$E_\theta(\cdot)$". But for convenience, we always write it as "$E(\cdot)$". Without further explanations, we postulate that all equations and inequalities below hold for all $\theta \in \Theta$ if those equations or inequalities are associated with θ. For a matrix $A = (a_{ij})$, we denote by $\|A\|$ the Frobenius norm of A (sometimes known as the Euclidean norm)

$$\|A\| = (\sum_i \sum_j a_{ij}^2)^{1/2}.$$

Let $\{A_n\}$, $\{B_n\}$ be sequences of symmetric positive definite (p.d.) matrices. If there is a sequence $\{D_n\}$ of matrices such that $A_n - B_n + D_n$ is nonnegative definite for each n and $\|D_n\| \to 0$ as $n \to \infty$, we shall say that $A_n - B_n$ is asymptotically nonnegative definite(annd).

Heyde and Gay(1989) gave the following definition of an asymptotic quasi-score estimating function :

__Definition__ **1** *: Suppose that $G_T^* \in \mathcal{G}_T \subseteq \Psi$ for each $T \geq 0$. If there is a sequence of positive functions $\{\alpha_T, T \geq 0\}$ such that*

$$\alpha_T{}^{-1}\{(E\dot{G}_T^*)'(EG_T^*G_T^{*\prime})^{-1}(E\dot{G}_T^*) - (E\dot{G}_T)'(EG_TG_T{}')^{-1}(E\dot{G}_T)\} \tag{1}$$

is asymptotically nonnegative definite for all $G_T \in \mathcal{G}_T$ and all $\theta \in \Theta$ as $T \to \infty$ and

$$\underline{\lim}_{T\to\infty} \alpha_T{}^{-1}\|(E\dot{G}_T^*)'(EG_T^*G_T^{*\prime})^{-1}(E\dot{G}_T^*)\| > 0.$$

we shall say that $\{G_T^*\}$ is an asymptotic quasi-score sequence of estimating functions within $\mathcal{G} = \{\mathcal{G}_T\}$. A solution θ_T^* of $G_T^*(\theta) = 0$ will be called an asymptotic quasi-likelihood (AQL) estimator.

It is obvious that this concept provides a new idea for developing the quasi-likelihood method. However the above definition of an asymptotic quasi-score sequence of estimating functions is not entirely satisfactory. This subject will be discussed in Section 2. We shall give a new definition of a quasi-score sequence of estimating functions in Section 3 and discuss the relations between the sequence of quasi-score estimating functions and the sequence of asymptotic quasi-score estimating functions as well as the relations between the definition and the new definition of asymptotic quasi-score estimating functions . Following the discussion in Chapter 2, we also give a definition of asymptotic E-sufficiency in Section 4 and show that, under certain regularity conditions, the asymptotic quasi-score estimating function will be asymptotically E-sufficient.

2 Deliberations

In Chapter 4 , we have pointed out that, in a certain sense, the quasi-score estimating function provides smaller confidence zones for an unknown parameter than other estimating functions in an estimating function space. Therefore, in the view of its method of construction, we also expect that the asymptotic quasi-score estimating function also has this important property. In Chapter 4, we see that the matrix $(E\dot{G}_T)^{-1}(EG_TG_T')(E\dot{G}_T')^{-1}$ may be employed to determine the dimensions of the confidence zone. Following the discussion there, in order to obtain smaller confidence zones than those generated by other estimating functions , we need to choose G_T^* such that

$$(E\dot{G}_T)^{-1}(EG_TG_T')(E\dot{G}_T')^{-1} - (E\dot{G}_T^*)^{-1}(EG_T^*G_T^{*\prime})(E\dot{G}_T^{*\prime})^{-1} > 0,$$

for all $G_T \in \mathcal{G}_T$. Therefore, if under Definition 1, the asymptotic quasi-score estimating function is to provide smaller confidence zones than other estimating functions the following property should hold : there is a sequence of positive functions $\{\beta_T, T \geq 0\}$ such that

$$\beta_T{}^{-1}\{(E\dot{G}_T)^{-1}(EG_TG_T')(E\dot{G}_T')^{-1} - (E\dot{G}_T^*)^{-1}(EG_T^*G_T^{*\prime})(E\dot{G}_T^{*\prime})^{-1}\} \qquad (2)$$

is asymptotically nonnegative definite for all $G_T \in \mathcal{G}_T$, as $T \to \infty$, and

$$\varliminf_{T\to\infty} \beta_T{}^{-1}\|(E\dot{G}_T^*)^{-1}(EG_T^*G_T^{*\prime})(E\dot{G}_T^{*\prime})^{-1}\| > 0.$$

The last inequality gives us some information about the rate at which

$$\|(E\dot{G}_T^*)^{-1}(EG_T^*G_T^{*\prime})(E\dot{G}_T^{*\prime})^{-1}\|$$

tends to 0. However (1) and (2) are not equivalent and an example is given below to show this. Equivalence may break down when the eigenvalues of $(E\dot{G}_T^*)^{-1}$ $(EG_T^*G_T^{*\prime})$ $(E\dot{G}_T^{*\prime})^{-1}$ go to zero at different rates. That means that, under Definition 1, the asymptotic quasi-score estimating function may not provide smaller confidence zones than others in some cases. In order to discuss the relations between (1) and (2), we say, for convenience, that the asymptotic quasi-score estimating function under Definition 1 is a 1-asymptotic quasi-score estimating function . We give the definition of a 2-asymptotic quasi-score estimating function below:

<u>**Definition**</u> **2** *Suppose that $G_T^* \in \mathcal{G}_T \subseteq \Psi$ for each $T \geq 0$. If there is a sequence of positive functions $\{\beta_T, T \geq 0\}$ such that*

$$\beta_T{}^{-1}\{(E\dot{G}_T)^{-1}(EG_TG_T{}')(E\dot{G}_T')^{-1} - (E\dot{G}_T^*)^{-1}(EG_T^*G_T^{*\prime})(E\dot{G}_T^{*\prime})^{-1}\}$$

is asymptotically nonnegative definite for all $G_T \in \mathcal{G}_T$, as $T \to \infty$, and

$$\varliminf_{T\to\infty} \beta_T{}^{-1}\|(E\dot{G}_T^*)^{-1}(EG_T^*G_T^{*\prime})(E\dot{G}_T^{*\prime})^{-1}\| > 0,$$

we shall say that $\{G_T^*\}$ is a 2-asymptotic quasi-score sequence of estimating functions within $\{\mathcal{G}_T\}$.

We denote by $G_T^{\triangle}$ the quasi-score estimating function within $\mathcal{G}_T$ for each T, if it exists.

<u>**Theorem**</u> **1** *Suppose that, for every T, $G_T^{\triangle}$ exists within $\mathcal{G}_T$ and*

$$\frac{\lambda_{max}((E\dot{G}_T^*)'(EG_T^*G_T^{*\prime})^{-1}(E\dot{G}_T^*))}{\lambda_{min}((E\dot{G}_T^*)'(EG_T^*G_T^{*\prime})^{-1}(E\dot{G}_T^*))} = O(1).$$

Then, $\{G_T^*\}$ is a 1-asymptotic quasi-score sequence of estimating functions within $\{\mathcal{G}_T\}$ and

$$0 < \varliminf_{T\to\infty} \alpha_T^{-1}\|(E\dot{G}_T^*)'(EG_T^*G_T^{*\prime})^{-1}(E\dot{G}_T^*)\|$$
$$\leq \varlimsup_{T\to\infty} \alpha_T^{-1}\|(E\dot{G}_T^*)'(EG_T^*G_T^{*\prime})^{-1}(E\dot{G}_T^*)\| < \infty$$

iff $\{G_T^*\}$ is a 2-asymptotic quasi-score sequence of estimating functions within $\{\mathcal{G}_T\}$ and

$$0 < \varliminf_{T\to\infty} \alpha_T\|(E\dot{G}_T^*)^{-1}(EG_T^*G_T^{*\prime})(E\dot{G}_T^{*\prime})^{-1}\|$$

$$\leq \overline{\lim_{T\to\infty}} \alpha_T \|(E\dot{G}_T^*)^{-1}(EG_T^*G_T^{*\prime})(E\dot{G}_T^{*\prime})^{-1}\| < \infty$$

where α_T is the positive function in Definition 1.

Proof: We need only prove the necessity part. The proof of the other part is similar.

Let

$$D_T = -\alpha_T^{-1}\{(E\dot{G}_T^*)'(EG_T^*G_T^{*\prime})^{-1}(E\dot{G}_T^*) - (E\dot{G}_T^\triangle)'(EG_T^\triangle G_T^{\triangle\prime})^{-1}(E\dot{G}_T^\triangle)\}.$$

Following the definition of $\{G_T^*\}$, there is a sequence of matrices $\{\tilde{D}_T\}$ such that

$$\alpha_T^{-1}\{(E\dot{G}_T^*)'(EG_T^*G_T^{*\prime})^{-1}(E\dot{G}_T^*) - (E\dot{G}_T^\triangle)'(EG_T^\triangle G_T^{\triangle\prime})^{-1}(E\dot{G}_T^\triangle)\} + \tilde{D}_T \geq 0$$

and $\|\tilde{D}_T\| \to 0$ as $T \to \infty$. Therefore, from the definition of $G_T^\triangle$, and for any $x \in R^n$ fixed,

$$0 \geq \overline{\lim_{T\to\infty}} x'\{\alpha_T^{-1}[(E\dot{G}_T^*)'(EG_T^*G_T^{*\prime})^{-1}(E\dot{G}_T^*) - (E\dot{G}_T^\triangle)'(EG_T^\triangle G_T^{\triangle\prime})^{-1}(E\dot{G}_T^\triangle)]\}x$$

$$\geq \underline{\lim_{T\to\infty}}\{x'\{\alpha_T^{-1}[(E\dot{G}_T^*)'(EG_T^*G_T^{*\prime})^{-1}(E\dot{G}_T^*) - (E\dot{G}_T^\triangle)'(EG_T^\triangle G_T^{\triangle\prime})^{-1}(E\dot{G}_T^\triangle)] + \tilde{D}_T\}x$$
$$-x'\tilde{D}_T x\}$$

$$\geq \underline{\lim_{T\to\infty}} x'\{\alpha_T^{-1}[(E\dot{G}_T^*)'(EG_T^*G_T^{*\prime})^{-1}(E\dot{G}_T^*) - (E\dot{G}_T^\triangle)'(EG_T^\triangle G_T^{\triangle\prime})^{-1}(E\dot{G}_T^\triangle)] + \tilde{D}_T\}x$$
$$-\overline{\lim_{T\to\infty}}|x'\tilde{D}_T x|.$$

But

$$0 \leq \overline{\lim_{T\to\infty}}|x'\tilde{D}_T x| = 0$$

and

$$x'\{\alpha_T^{-1}[(E\dot{G}_T^*)'(EG_T^*G_T^{*\prime})^{-1}(E\dot{G}_T^*) - (E\dot{G}_T^\triangle)'(EG_T^\triangle G_T^{\triangle\prime})^{-1}(E\dot{G}_T^\triangle)] + \tilde{D}_T\}x \geq 0,$$

for all T, so that we have

$$\lim_{T\to\infty} x'\alpha_T^{-1}\{(E\dot{G}_T^*)'(EG_T^*G_T^{*\prime})^{-1}(E\dot{G}_T^*) - (E\dot{G}_T^\triangle)'(EG_T^\triangle G_T^{\triangle\prime})^{-1}(E\dot{G}_T^\triangle)\}x = 0$$

for all $x \in R^n$, that is

$$\lim_{T\to\infty} \|D_T\| =$$

$$\lim_{T\to\infty} \|\alpha_T^{-1}\{(E\dot{G}_T^*)'(EG_T^*G_T^{*\prime})^{-1}(E\dot{G}_T^*) - (E\dot{G}_T^\triangle)'(EG_T^\triangle G_T^{\triangle\prime})^{-1}(E\dot{G}_T^\triangle)\}\| = 0.$$

Let

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

$$A_T^* = (E\dot{G}_T^*)'(EG_T^*G_T^{*\prime})^{-1}(E\dot{G}_T^*)$$

and

$$A_T^\triangle = (E\dot{G}_T^\triangle)'(EG_T^\triangle G_T^{\triangle\prime})^{-1}(E\dot{G}_T^\triangle);$$

then

$$D_T = -\alpha_T^{-1}(A_T^* - A_T^\triangle).$$

By the definition of $G_T^\triangle$,

$$\alpha_T^{-1}(A_T^* - A_T) + D_T$$

$$= (\alpha_T^{-1}A_T^* + D_T) - \alpha_T^{-1}A_T$$

is nonnegative definite. Since

$$(\alpha_T^{-1}A_T^* + D_T) = \alpha_T^{-1}A_T^\triangle > 0,$$

and

$$\alpha_T^{-1}A_T > 0,$$

we obtain

$$(\alpha_T^{-1}A_T)^{-1} - (\alpha_T^{-1}A_T^* + D_T)^{-1} \geq 0,$$

and that is

$$(\alpha_T^{-1}A_T)^{-1} - (\alpha_T^{-1}A_T^*)^{-1}(I + (\alpha_T^{-1}A_T^*)^{-1}D_T)^{-1} \geq 0.$$

The above inequality can be rewritten as

$$0 \leq (\alpha_T^{-1}A_T)^{-1} - (\alpha_T^{-1}A_T^*)^{-1}(I + (\alpha_T^{-1}A_T^*)^{-1}D_T)^{-1}$$

$$= (\alpha_T^{-1}A_T)^{-1} - (\alpha_T^{-1}A_T^*)^{-1}(I + (\alpha_T^{-1}A_T)^{-1}D_T + ((\alpha_T^{-1}A_T)^{-1}D_T)^2 +$$

$$+((\alpha_T^{-1}A_T)^{-1}D_T)^3 + \cdots)$$

$$= (\alpha_T^{-1}A_T)^{-1} - (\alpha_T^{-1}A_T^*)^{-1} - (\alpha_T^{-1}A_T^*)^{-1}((\alpha_T^{-1}A_T^*)^{-1}D_T +$$

$$+((\alpha_T^{-1}A_T^*)^{-1}D_T)^2+\cdots)$$
$$=(\alpha_T^{-1}A_T)^{-1}-(\alpha_T^{-1}A_T^*)^{-1}+\breve{D}_T.$$

Here

$$\|\breve{D}_T\| = \|(\alpha_T^{-1}A_T^*)^{-1}((\alpha_T^{-1}A_T^*)^{-1}D_T + +((\alpha_T^{-1}A_T^*)^{-1}D_T)^2+\cdots)\|$$
$$\leq \|(\alpha_T^{-1}A_T^*)^{-1}\| \sum_{k=1}^{\infty} \|(\alpha_T^{-1}A_T^*)^{-1}D_T\|^k.$$

In the following we want to prove that $\lim_{T\to\infty}\|\breve{D}_T\|=0$. Since

$$\frac{\lambda_{max}(A_T^*)}{\lambda_{min}(A_T^*)} = O(1),$$

if $\underline{\lim}_{T\to\infty}\alpha_T^{-1}\lambda_{min}(A_T^*)=0$, so is $\underline{\lim}_{T\to\infty}\alpha_T^{-1}\lambda_{max}(A_T^*)=0$, and then we have

$$\underline{\lim}_{T\to\infty}\alpha_T^{-1}\|(A_T^*)\|=0,$$

which provides a contradiction with the assumption. Therefore there exists a constant $\alpha>0$ such that, for large enough T,

$$\alpha_T^{-1}\lambda_{min}(A_T^*) \geq \alpha > 0.$$

This yields the existence of a constant c such that

$$\|\alpha_T(A_T^*)^{-1}\| \leq c,$$

for large enough T. Hence, for large enough T,

$$\|\breve{D}_T\| \leq \|\alpha_T(A_T^*)^{-1}\| \sum_{k=1}^{\infty} \|\alpha_T(A_T^*)^{-1}D_T\|^k$$
$$\leq c\frac{c\|D_T\|}{1-c\|D_T\|} \to 0, \qquad \text{as } T\to\infty.$$

We therefore obtain

$$\alpha_T(A_T)^{-1}-\alpha_T(A_T^*)^{-1}$$

is asymptotically nonnegative definite. By assumption we also have

$$\underline{\lim}_{T\to\infty}\alpha_T\|(A_T)^{-1}\|>0.$$

Thus, $\{G_T^*\}$ is a 2-asymptotic quasi-score sequence of estimating functions within $\{\mathcal{G}_T\}$ and

$$0 < \underline{\lim}_{T\to\infty} \alpha_T \|(E\dot{G}_T^*)^{-1}(EG_T^*G_T^{*\prime})(E\dot{G}_T^{*\prime})^{-1}\|$$
$$\leq \overline{\lim}_{T\to\infty} \alpha_T \|(E\dot{G}_T^*)^{-1}(EG_T^*G_T^{*\prime})(E\dot{G}_T^{*\prime})^{-1}\| < \infty.$$

Remark: If the conditions of Theorem 1 do not hold, the concept of 1-asymptotic quasi-score sequence of estimating functions is not equivalent to the concept of a 2-asymptotic quasi-score sequence of estimating functions . The following example illustrates this fact.

Example 1: Let $\{X_T\}$ be a process in R^2 and

$$X_T = \int_0^T \begin{pmatrix} \theta_1^{t^6} \\ \theta_2/T^2 \end{pmatrix} dt + \begin{pmatrix} m_1(T) \\ m_2(T) \end{pmatrix}$$
$$= \begin{pmatrix} \int_0^T \theta_1^{t^6} dt \\ \theta_2/T \end{pmatrix} + \begin{pmatrix} m_1(T) \\ m_2(T) \end{pmatrix},$$

where θ_1, $\theta_2 > 1$ are parameters; $m_1(T)$ and $m_2(T)$ are martingales with respect to some σ-fields $\{\mathcal{F}_T\}$ and $m = (m_1, m_2)'$ satisfies

$$< m(\theta) >_T = \int_0^T I dt = \begin{pmatrix} T & 0 \\ 0 & T \end{pmatrix}.$$

Let $\mathcal{G}_T$ be the set $\{\int_0^T f_t(\theta) dm_t \mid \{f_t\}$ is a predictable process with respect to $\{\mathcal{F}_t\}\}$. Following Godambe and Heyde(1987), the quasi-score estimating function in $\mathcal{G}_T$ is

$$G_T^{\triangle} = \begin{pmatrix} \int_0^T t^6\theta_1^{t^6-1} dm_1(t) \\ m_2(T)/T^2 \end{pmatrix}.$$

After some calculations we find that

$$(E\dot{G}_T^{\triangle})^{-1}(EG_T^{\triangle}G_T^{\triangle\prime})(E\dot{G}_T^{\triangle\prime})^{-1} = \begin{pmatrix} (\int_0^T t^{12}\theta_1^{2t^6-2} dt)^{-1} & 0 \\ 0 & T^3 \end{pmatrix}$$

and

$$(E\dot{G}_T^{\triangle})'(EG_T^{\triangle}G_T^{\triangle\prime})^{-1}(E\dot{G}_T^{\triangle}) = \begin{pmatrix} \int_0^T t^{12}\theta_1^{2t^6-2} dt & 0 \\ 0 & T^{-3} \end{pmatrix}.$$

Now we construct estimating functions G_T^* for every T. Let

$$G_T^* = G_T^{\triangle} + \begin{pmatrix} 0 \\ \int_0^T \theta_2 t dm_2(t) \end{pmatrix}.$$

Then we obtain

$$(E\dot{G}_T^*)^{-1}(EG_T^*G_T^{*\prime})(E\dot{G}_T^{*\prime})^{-1} = \begin{pmatrix} (\int_0^T t^{12}\theta_1^{2t^6-2}dt)^{-1} & 0 \\ 0 & \frac{T^{-3}+\theta_2+1/3\theta_2^2T^3}{(T^{-3}+1/2\theta_2)^2} \end{pmatrix}.$$

Since

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

$$= \begin{pmatrix} 0 & 0 \\ 0 & T^{-3}(\frac{(T^{-3}+1/2\theta_2)^2}{T^{-6}+\theta_2T^{-3}+1/3\theta_2^2} - 1) \end{pmatrix} \to 0, \qquad \text{as } T \to \infty$$

and

$$\lim_{T\to\infty} \|(E\dot{G}_T^*)'(EG_T^*G_T^{*\prime})^{-1}(E\dot{G}_T^*)\|$$

$$= \lim_{T\to\infty} \left\| \begin{pmatrix} \int_0^T t^{12}\theta_1^{2t^6-2}dt & 0 \\ 0 & \frac{(T^{-3}+1/2\theta_2)^2}{T^{-3}+\theta_2+1/3\theta_2^2T^3} \end{pmatrix} \right\| > 0,$$

G_T^* is a 1-asymptotic quasi-score estimating function , but it is not possible for it to be a 2-asymptotic quasi-score estimating function . If G_T^* is a 2-asymptotic quasi-score estimating function , there exists a sequence of positive functions $\alpha_T > 0$ such that

$$\alpha_T^{-1}\|(E\dot{G}_T^{\triangle})^{-1}(EG_T^{\triangle}G_T^{\triangle\prime})(E\dot{G}_T^{\triangle\prime})^{-1} - (E\dot{G}_T^*)^{-1}(EG_T^*G_T^{*\prime})(E\dot{G}_T^{*\prime})^{-1}\|$$

$$= 1/3\alpha_T^{-1}T^3O(1) \to 0, \qquad \text{as } T \to \infty.$$

This yields

$$\alpha_T^{-1}\|(E\dot{G}_T^*)^{-1}(EG_T^*G_T^{*\prime})(E\dot{G}_T^{*\prime})^{-1}\|$$

$$= \alpha_T^{-1}(o(1) + T^6O(1))^{1/2} \to 0, \qquad \text{as } T \to \infty.$$

Thus we obtain a contradiction with the definition of a 2-asymptotic quasi-score estimating function .

Similarly we can use the same technique to construct an example of a 2-asymptotic quasi-score estimating function which is not a 1-asymptotic quasi-score estimating function .

3 Asymptotic Quasi-score Estimating Functions : a New Definition

As we have discussed in Section 2, Definition 1 is not adequate to ensure that the asymptotic quasi-score estimating function will provide smaller asymptotic confidence zones than other estimating functions in a given estimating functions space. Therefore we give another definition of asymptotic quasi-score estimating function which avoids this problem.

<u>Definition</u> 3 *Suppose* $\{G_T^*\} \in \mathcal{G}_T \subseteq \Psi$. *If there is a family of nonsingular matrices* $\{B_T\}$ *such that*

$$B_T^{-1}\{(E\dot{G}_T)^{-1}(EG_TG_T{}')(E\dot{G}_T')^{-1} - (E\dot{G}_T^*)^{-1}(EG_T^*G_T^*{}')(E\dot{G}_T^*{}')^{-1}\}(B_T^{-1})'$$

is asymptotically nonnegative definite for all $\{G_T\} \in \mathcal{G}_T$, *as* $T \to \infty$ *and*

$$\varliminf_{T\to\infty} \|B_T^{-1}\{(E\dot{G}_T^*)^{-1}(EG_T^*G_T^*{}')(E\dot{G}_T^*{}')^{-1}(B_T^{-1})'\| > 0, \tag{3}$$

we shall say that $\{G_T^*\}$ *is an asymptotic quasi-score sequence of estimating functions*

.

Remarks: 1. In practice, If G_T^* is a suitable estimating function we usually have

$$\|(E\dot{G}_T^*)^{-1}(EG_T^*G_T^*{}')(E\dot{G}_T^*{}')^{-1}\| \to 0$$

as $T \to \infty$. Therefore (3) shows the convergence speed of $\|(E\dot{G}_T^*)^{-1}$ $(EG_T^*G_T^*{}')$ $(E\dot{G}_T^*{}')^{-1}\|$. Of course we may choose $\{B_T\}$ to be a sequence of functions. But if the eigenvalues of

$$(E\dot{G}_T^*)^{-1}(EG_T^*G_T^{*\prime})(E\dot{G}_T^{*\prime})^{-1}$$

converge to zero at different rates, it is better to choose $\{B_T\}$ to be a sequence of matrices from which we can take account of the different behaviour of each of the eigenvalues. This is equivalent to, after a suitable linear transformation, dealing with the different behaviour of the confidence intervals for each parameter.

2. The concept of asymptotic quasi-score estimating function is associated with some sequence of matrices $\{B_T\}$.

3. If $\{B_T\}$ is a sequence of real functions, then the new definition of asymptotic quasi-score sequence of estimating functions coincides the definition of 2-asymptotic quasi-score sequence of estimating functions .

Now we need to investigate the relations between asymptotic quasi-score estimating functions and quasi-score estimating functions and show that the definition of an asymptotic quasi-score estimating function is reasonable. The following theorems will tell us that, in a certain sense, both

$$(E\dot{G}_T^{\triangle})^{-1}(EG_T^{\triangle}G_T^{\triangle\prime})(E\dot{G}_T^{\triangle\prime})^{-1}$$

and

$$(E\dot{G}_T^*)^{-1}(EG_T^*G_T^*)(E\dot{G}_T^{*\prime})^{-1}(B_T^{-1})$$

have a same limit, where $\{G_T^{\triangle}\}$ is the sequence of the quasi-score estimating functions and $\{G_T^*\}$ is the sequence of asymptotic quasi-score estimating functions .

<u>**Theorem**</u> **2** *Suppose that $\{G_T^{\triangle}\}$ is a quasi-score sequence of estimating functions within $\{\mathcal{G}_T\}$. Then $\{G_T^*\}$ is an asymptotic quasi-score sequence of estimating functions iff there exists a sequence of matrices $\{B_T\}$ such that*

$$\|B_T^{-1}\{(E\dot{G}_T^{\triangle})^{-1}(EG_T^{\triangle}G_T^{\triangle\prime})(E\dot{G}_T^{\triangle\prime})^{-1} - (E\dot{G}_T^*)^{-1}(EG_T^*G_T^*)(E\dot{G}_T^{*\prime})^{-1}\}(B_T^{-1})\| \to 0,$$

as $T \to \infty$, and (3) holds.

Proof: Using the technique of the proof in Theorem 1, the necessity part is obvious.

The following is the proof of the sufficiency part. Since, for any $G_T \in \mathcal{G}_T$,

$$B_T^{-1}\{((E\dot{G}_T)^{-1}(EG_TG_T')(E\dot{G}_T')^{-1} - (E\dot{G}_T^*)^{-1}(EG_T^*G_T^{*\prime})(E\dot{G}_T^{*\prime})^{-1}\}(B_T^{-1})'$$

$$= B_T^{-1}\{(E\dot{G}_T)^{-1}(EG_TG_T')(E\dot{G}_T')^{-1} - (E\dot{G}_T^{\triangle})^{-1}(EG_T^{\triangle}G_T\triangle')(E\dot{G}_T^{\triangle\prime})^{-1}\}(B_T^{-1})'$$

$$+B_T^{-1}\{(E\dot{G}_T^{\triangle})^{-1}(EG_T^{\triangle}G_T\triangle')(E\dot{G}_T^{\triangle\prime})^{-1} - (E\dot{G}_T^*)^{-1}(EG_T^*G_T^{*\prime})(E\dot{G}_T^{*\prime})^{-1}\}(B_T^{-1}),$$

let

$$D_T = B_T^{-1}\{(E\dot{G}_T^{\triangle})^{-1}(EG_T^{\triangle}G_T\triangle')(E\dot{G}_T^{\triangle\prime})^{-1}$$
$$-(E\dot{G}_T^*)^{-1}(EG_T^*G_T^{*\prime})(E\dot{G}_T^{*\prime})^{-1}\}(B_T^{-1})'.$$

Then we have

$$B_T^{-1}\{(E\dot{G}_T)^{-1}(EG_TG_T')(E\dot{G}_T')^{-1} - (E\dot{G}_T^*)^{-1}(EG_T^*G_T^{*\prime})(E\dot{G}_T^{*\prime})^{-1}\}(B_T^{-1})' - D_T)$$
$$\geq 0.$$

Following the assumption

$$B_T^{-1}\{(E\dot{G}_T^{\triangle})^{-1}(EG_T^{\triangle}G_T\triangle')(E\dot{G}_T^{\triangle\prime})^{-1} - (E\dot{G}_T^*)^{-1}(EG_T^*G_T^{*\prime})(E\dot{G}_T^{*\prime})^{-1}\}B_T^{-1} \to 0,$$

as $T \to \infty$, we obtain that $\{G_T^*\}$ is an asymptotic quasi-score sequence of estimating functions .

Theorem 3 *Suppose that $G_T^* \in \mathcal{G}_T \subseteq \Psi$ for each T. If $G_T^{\triangle}$ exists for each T and there exists a sequence of nonsingular matrices $\{A_T\}$ such that*

$$A_T'\{(E\dot{G}_T^*)'(EG_T^*G_T^{*\prime})^{-1}(E\dot{G}_T^*) - (E\dot{G}_T)'(EG_TG_T')^{-1}(E\dot{G}_T)\}A_T$$

is asymptotically nonnegative definite for all $G_T \in \mathcal{G}_T$, as $T \to \infty$, and

$$0 < \varliminf_{T\to\infty} \lambda_{min}(A_T'(E\dot{G}_T^*)'(EG_T^*G_T^{*\prime})^{-1}(E\dot{G}_T^*)A_T) \tag{4}$$
$$\leq \varlimsup_{T\to\infty} \lambda_{max}(A_T'(E\dot{G}_T^*)'(EG_T^*G_T^{*\prime})^{-1}(E\dot{G}_T^*)A_T) < \infty, \tag{5}$$

then $\{G_T^*\}$ is an asymptotic quasi-score sequence of estimating functions within $\{\mathcal{G}_T\}$.

Proof: Because (7) and (8) ensure that

$$0 < \underline{\lim}_{T\to\infty} \|(A_T)^{-1}(E\dot{G}_T^*)^{-1}(EG_T^*G_T^{*\prime})(E\dot{G}_T^{*\prime})^{-1}(A_T^{-1})'\| < \infty,$$

we need only prove that

$$A_T^{-1}\{(E\dot{G}_T)^{-1}(EG_TG_T')(E\dot{G}_T')^{-1} - (E\dot{G}_T^*)^{-1}(EG_T^*G_T^{*\prime})(E\dot{G}_T^{*\prime})^{-1}\}(A_T^{-1})'$$

is asymptotically nonnegative definite.

Let

$$B_T = (E\dot{G}_T)'(EG_TG'_T)^{-1}(E\dot{G}_T),$$

$$B_T^* = (E\dot{G}_T^*)'(EG_T^*G_T^{*\prime})^{-1}(E\dot{G}_T^*),$$

$$B_T^\triangle = (E\dot{G}_T^\triangle)'(EG_T^\triangle G_T^{\triangle\prime})^{-1}(E\dot{G}_T^\triangle),$$

$$D_T = A_T'\{B_T^* - B_T^\triangle\}A_T.$$

Using the same method as in the proof of Theorem 1 , we have $\|D_T\| \to 0$ as $T \to \infty$. Since

$$0 \le A_T'(B_T^* - B_T)A_T - D_T = (A_T'B_T^*A_T - D_T) - A_T'B_TA_T$$

and

$$A_T'B_T^*A_T - D_T = A_T'B_T^\triangle A_T > 0,$$

we obtain

$$A_T^{-1}B_T^{-1}(A_T^{-1})' - (A_T'B_T^*A_T - D_T)^{-1} \ge 0,$$

i.e.

$$A_T^{-1}B_T^{-1}(A_T^{-1})' - A_T^{-1}(B_T^*)^{-1}(A_T^{-1})'(I - A_T^{-1}(B_T^*)^{-1}(A_T^{-1})'D_T$$
$$+(A_T^{-1}(B_T^*)^{-1}(A_T^{-1})'D_T)^2 - \cdots)$$
$$= A_T^{-1}(B_T)^{-1}(A_T^{-1})' - A_T^{-1}(B_T^*)^{-1}(A_T^{-1})'$$
$$+A_T^{-1}(B_T^*)^{-1}(A_T^{-1})'(A_T^{-1}(B_T^*)^{-1}(A_T^{-1})'D_T - (A_T^{-1}(B_T^*)^{-1}(A_T^{-1})'D_T)^2 + \cdots) \ge 0.$$

Following the assumptions of the theorem, there is an constant c such that

$$\|A_T^{-1}(B_T^*)^{-1}(A_T^{-1})'\| \le c < \infty,$$

for large enough T and then $\|D_T\| \to 0$, as $T \to \infty$, yields

$$\|A_T^{-1}(B_T^*)^{-1}(A_T^{-1})'(A_T^{-1}(B_T^*)^{-1}(A_T^{-1})'D_T - (A_T^{-1}(B_T^*)^{-1}(A_T^{-1})'D_T)^2 + \cdots)\|$$
$$\leq c\frac{c\|D_T\|}{1 - c\|D_T\|} \to 0, \qquad \text{as } T \to \infty.$$

Therefore

$$A_T^{-1}\{(E\dot{G}_T)^{-1}(EG_TG_T')((E\dot{G}_T)')^{-1} - (E\dot{G}_T^*)^{-1}(EG_T^*G_T^*)((E\dot{G}_T^*)')^{-1}\}(A_T^{-1})'$$
$$= A_T^{-1}(B_T^{-1} - (B_T^*)^{-1})(A_T^{-1})'$$

is asymptotically nonnegative definite for all $G_T \in \mathcal{G}_T$ and $\{G_T^*\}$ is an asymptotic quasi-score sequence of estimating functions .

Theorem 3 reveals the relation between the old and new definitions of an asymptotic quasi-score sequence of estimating functions .

4 Asymptotically E-sufficient Estimating Functions

As indicated in Chapter 2, it is desirable to choose an estimating function from the set of E-sufficient estimating functions or the set of locally E-sufficient estimating functions. However it is sometimes difficult to do this for the same reasons as we have indicated in the introduction of this chapter. Therefore we need a new set of estimation functions to substitute for the set of E-sufficient estimating functions or the set of locally E-sufficient estimating functions.

Let Ψ be an estimating function space. Every element ψ of Ψ is a sequence $\{\psi_T\}$ of estimating functions . For every T fixed, let

$$\Psi_T = \{\psi_T | \psi \in \Psi\}.$$

Without further explanation, we shall follow the definitions, notations and assumptions of Section 2 of Chapter 2 here. Following the discussion in Chapter 2, we have

$$\Psi_T = \mathcal{S}_{loc,T} \otimes \mathcal{A}_{loc,T},$$

where $\mathcal{S}_{loc,T}$ and $\mathcal{A}_{loc,T}$ are the complete locally E-sufficient space and the locally E-ancillary spaces, respectively, for each T (see McLeish and Small(1988) and Chapter 2).

Definition 4 *We say that $\psi \in \Psi$ is asymptotically E-sufficient, if, for every T, ψ_T can be decomposed as*

$$\psi_T = \psi_{T,ps} + \psi_{T,pa},$$

where $\psi_{T,ps} \in \mathcal{S}_{loc,T}$ and $\psi_{T,pa} \in \mathcal{A}_{loc,T}$, and there exits a sequence of nonsingular matrices $\{A_T\}$ such that

$$\underline{\lim}_{T\to\infty} \|A_T(E\psi_T\psi_T{}')A_T{}'\| > 0,$$

$$\lim_{T\to\infty} \|A_T(E\psi_{T,pa}\psi_{T,pa}{}')A_T{}'\| = 0.$$

In all the discussion here we shall always assume that the necessary regularity conditions hold and

$$\mathcal{A}_{loc,T} = \{\psi_T | E_\theta\dot{\psi}_T = 0, \quad \text{for all } \theta \in \Theta\}.$$

The following theorem will show that under the above conditions, the asymptotic quasi-score sequence of estimating functions is asymptotically E-sufficient . However we first need a supplementary proposition.

Proposition 1 *If A and B are $p \times p$ nonnegative definite matrices, then*

$$\|B\| \leq p^{3/2}\|A + B\|.$$

Proof: Since

$$\lambda_{max}(A+B) \geq 1/p \ \ tr(A+B) \geq 1/p \ \ tr(B) \geq 1/p\lambda_{max}(B),$$

we obtain

$$(\|B\|)^2 = tr(BB') \leq p(\lambda_{max}(B))^2 \leq p^3(\lambda_{max}(A+B))^2$$

$$= p^3\lambda_{max}((A+B)(A+B)') \leq p^3 \ \ tr((A+B)(A+B)') = p^3(\|A+B\|)^2.$$

Therefore $\|B\| \leq p^{3/2}\|A+B\|$ as required.

Borrowing the definitions of $\tilde{\mathcal{G}}_T$ and sub-quasi-score estimating functions in Chapter 2, we have the theorem below.

Theorem 4 *Suppose that $\tilde{\mathcal{G}}_T \subseteq \Psi_T$ and $G_T^{\triangle}$ is a sub-quasi-score estimating function in $\tilde{\mathcal{G}}_T$. Under regularity conditions, if $G_T^* \in \tilde{\mathcal{G}}_T$ satisfies*

$$\|B_T^{-1}\{(E\dot{G}_T^{\triangle})^{-1}(EG_T^{\triangle}(G_T^{\triangle})')(E\dot{G}_T^{\triangle\prime})^{-1} - (E\dot{G}_T^*)^{-1}(EG_T^*(G_T^*)')(E\dot{G}_T^{*\prime})^{-1}\}(B_T^{-1})'\| \tag{6}$$

$$\to 0, \qquad \text{as } T \to \infty,$$

and

$$\varliminf_{T\to\infty} \|B_T^{-1}(E\dot{G}_T^*)^{-1}(EG_T^*(G_T^*)')(E\dot{G}_T^{*\prime})^{-1}(B_T^{-1})'\| > 0,$$

for some sequence of nonsingular matrices $\{B_T\}$, then $\{G_T^*\}$ is asymptotically E-sufficient.

Proof: Assume $G_T^* = G_{T,ps}^* + G_{T,pa}^*$, where $G_{T,ps}^* \in \mathcal{S}_{loc,T}$ and $G_{T,pa}^* \in \mathcal{A}_{loc,T}$. We may assume that all elements of Ψ_T take values in R^p. The only thing we need to prove is that there is a sequence of matrices $\{A_T\}$ such that

$$\lim_{T\to\infty} \|A_T E G_{T,pa}^* G_{T,pa}^{*\prime} A_T'\| = 0.$$

Let $A_T = B_T^{-1}(E\dot{G}_T^*)^{-1}$. Since

$$B_T^{-1}(E\dot{G}_T^*)^{-1}(EG_{T,pa}^*(G_{T,pa}^*)')(E\dot{G}_T^{*\prime})^{-1}(B_T^{-1})' \geq 0$$

and following the definition of $G_T^{\triangle}$,

$$B_T^{-1}\{(E\dot{G}_T^*)^{-1}(EG_{T,ps}^*(G_{T,ps}^*)')(E\dot{G}_T^{*\prime})^{-1} - (E\dot{G}_T^{\triangle})^{-1}(EG_T^{\triangle}(G_T^{\triangle})')(E\dot{G}_T^{\triangle\prime})^{-1}\}(B_T^{-1})'$$

$$\geq 0.$$

Then, in the light of Proposition 1, we obtain

$$\begin{aligned}
&\|B_T^{-1}\{(E\dot{G}_T^*)^{-1}(EG_{T,pa}^*G_{T,pa}^{*\prime})(E\dot{G}_T^{*\prime})^{-1}\}(B_T^{-1})'\| \\
&\leq p^{3/2}\|B_T^{-1}\{(E\dot{G}_T^*)^{-1}(EG_{T,ps}^*G_{T,ps}^{*\prime})(E\dot{G}_T^{*\prime})^{-1} \\
&\quad -(E\dot{G}_T^{\triangle})^{-1}(EG_T^{\triangle}(G_T^{\triangle})')(E\dot{G}_T^{\triangle\prime})^{-1}\}(B_T^{-1})' \\
&\quad +B_T^{-1}\{(E\dot{G}_T^*)^{-1}(EG_{T,pa}^*G_{T,pa}^{*\prime})(E\dot{G}_T^{*\prime})^{-1}\}(B_T^{-1})'\|
\end{aligned}$$

$$= p^{3/2} \| B_T^{-1} \{ (E\dot{G}_T^{\triangle})^{-1} (EG_T^{\triangle}(G_T^{\triangle})')(E\dot{G}_T^{\triangle}{}')^{-1}$$

$$-(E\dot{G}_T^{*})^{-1}(EG_T^{*}(G_T^{*})')(E\dot{G}_T^{*}{}')^{-1}\}(B_T^{-1})' \| \to 0,$$

$$\text{as } T \to \infty.$$

Remark: As we have discussed in Chapter 2, in a wide range of cases, the quasi-score estimating function is a sub-quasi-score estimating function . Due to Theorem 2 and 4, under the regularity conditions, we obtain that an asymptotic quasi-score estimating function is asymptotically E-sufficient.

The following example gives an asymptotic quasi-score sequence of estimating functions which is asymptotically E-sufficient. This example is borrowed from Lloyd-Smith(1990).

Example: We suppose that $\{X_k\}$ is a process on some probability space, which satisfies

$$X_k = \theta X_{k-1} + u_k,$$

where

$u_k = (\theta_k - \theta)X_{k-1} + \varepsilon_k,$

$\{\theta_k\}_{k=0}^{\infty}$ is a stationary ergodic process,

$E\theta_k = \theta$ and $E(\theta_k - \theta)^2 = \eta,$

$E(\varepsilon_k | X_{k-1}) = 0.$ a.s.,

$E(\varepsilon_k^2 | X_{k-1}) = \sigma^2 X_{k-1},$ a.s.,

$E(\varepsilon_k(\theta_k - \theta)|X_{k-1}) = 0$

$E(u_k^2 | X_{k-1}) = \sigma^2 X_{k-1} + \eta^2 X_{k-1}^2,$ a.s.,

$E(u_k | X_{k-1}) = 0.$

Here θ is a parameter to be estimated and σ^2 and η^2 are nuisance parameters. The details of this (population) process can be found in Lloyd-Smith (1990). We can express the process in a useful form. Defining

$$y_t = \sum_{k=1}^{t} X_k,$$

$$m_t = \sum_{k=1}^{t} u_k,$$

$$f_s(\theta) = \theta X_{s-1}, \quad \text{where } X_s = X_{[s]}.$$

and $[s]$ denotes the integer part of s, then we can write

$$y_t = \theta \sum_{k=1}^{t} X_{k-1} + m_t$$

$$= \int_0^t f_s(\theta) d\lambda_s + m_t(\theta),$$

where $m_t(\theta)$ is a martingale with respect to $\mathcal{F}_t = \sigma(X_s, s \leq t)$ and

$$< m(\theta) >_t = \sum_{k=1}^{t} (\sigma^2 X_{k-1} + \eta^2 X_{k-1}^2)$$

$$= \int_0^t g_s d\lambda_s, \qquad a.s.,$$

while $g_s = \sigma^2 X_{s-1} + \eta^2 X_{s-1}^2$. Therefore

$$y_t = \int_0^t \frac{f_s(\theta)}{g_s} d < m(\theta) >_s + m_t(\theta).$$

Let $\Psi = \{\Psi \mid \Psi_T = \{\int_0^T H_s(\theta) dm_s(\theta) \mid \{H_s(\theta)\}$ is predictable with respect to $\{\mathcal{F}_s\}\}, T \geq 0\}$. In the light of Example 2 and Proposition 1 in Chapter 2, we have

$$\mathcal{A}_{loc,T} = \{\psi_T : E_\theta \dot{\psi}_T = 0, \quad \text{for all } \theta \in \Theta\}$$

and a quasi-score estimating function in $\mathcal{G}_T \subseteq \Psi_T$ is

$$G_T^{\triangle} = \sum_{k-1}^{T} X_{k-1} (\sigma^2 X_{k-1} + \eta^2 X_{k-1}^2)^{-1} (X_k - \theta X_{k-1})$$

which is a sub-quasi-score estimating function too .

Let $G_T^* = \sum_{k=1}^{T} X_{k-1}^{-1} (X_k - \theta X_{k-1})$. Lloyd-Smith(1990) has proved that $\{G_T^*\}$ is a 1-asymptotic quasi-score sequence of estimating functions . Here we shall prove that $\{G_T^*\}$ is a 2-asymptotic quasi-score sequence of estimating functions . Since G_T^* is a one dimensional function,

$$\frac{\lambda_{max}((E\dot{G}_T^*)'(EG_T^* G_T^{*\prime})^{-1}(E\dot{G}_T^*))}{\lambda_{min}((E\dot{G}_T^*)'(EG_T^* G_T^{*\prime})^{-1}(E\dot{G}_T^*))} = 1$$

and following the calculations of Lloyd-Smith(1990), we find that

$$1/T((E\dot{G}_T^*)'(EG_T^*G_T^{*\prime})^{-1}(E\dot{G}_T^*)) \to 1/\eta^2 < \infty,$$

as $T \to \infty$, Theorem 1 yields that $\{G_T^*\}$ is a 2-asymptotic quasi-score sequence of estimating functions . Therefore $\{G_T^*\}$ is a asymptotic quasi-score estimating functions . In the light of the discussions after Definition 8 in Chapter 2 and Theorem 4, $\{G_T^*\}$ is asymptotically E-sufficient.

CHAPTER 6
APPLICATIONS OF THE QUASI-LIKELIHOOD METHOD

1 Introduction

Much estimation theory is based on ad hoc application of least squares or conditional least squares methods in cases where specific distributional assumptions are avoided. However, the general quasi-likelihood theory can often suggest improved estimators. The theory deserves attention both for its concentration on efficiency issues and for the focus which it gives on the role of distributional assumptions.

In this chapter we illustrate the ideas by applying the general quasi-likelihood theory to estimation of the offspring and immigration means in a Bienaymé-Galton-Watson process with immigration and also the parameters in a heteroscedastic regression model. The former model has been used quite widely in both the physical and biological sciences as well as for traffic flows while the latter has been applied in analysis of cross-sectional data in various micro-economic contexts, such as for firms or households. We shall show how all the previous results can be improved by using quasi-likelihood methods.

In this chapter we also briefly discuss the application of the quasi-likelihood method to censoring problems. Censoring occurs when, by accident or design, the value of the process under investigation is unobserved for some of the items in the sample. How to use censored data to estimate the parameters of the process is called a censoring problem.

2 General Principles

The general framework is as follows. We are given a sample $\{Z_t, 0 \leq t \leq T\}$, say, taking values in r-dimensional Euclidean space and the set of probability measures for $\{Z_t\}$ is a union of families of models, each being indexed by a characteristic

θ belonging to an open subset of p-dimensional Euclidean space. The object is the efficient estimation of θ which is not necessarily a parameter; it could be, for example, a population mean.

We shall begin with a discussion of the case of discrete time in the interests of clarity and simplicity. Every discrete time process is a semimartingale and we may suppose that $\{Z_t\}$ has an associated filtration (past history σ-fields) $\{\mathcal{F}_t\}$ and possesses a decomposition

$$Z_t = Z_0 + A_t(\theta) + M_t(\theta) \tag{1}$$

where A_t a finite variation process, $M_t = \sum_{s=1}^{t} m_s$ is a local martingale and $A_0 = M_0 = 0$ (e.g. Rogers and Williams (1987, VI. 40)). The local martingale condition means that there exists a sequence of stopping times $\{T_n\}$ with $T_n \uparrow \infty$ such that $\{M_{T_n}\}$ is a martingale but M_t will be a martingale itself in most applications and then, if $\{Z_t\}$ is integrable, A_t can be chosen as predictable process with respect to $\{\mathcal{F}_t\}$, the decomposition (1) is unique and the differences of A_t have the simple form

$$A_t - A_{t-1} = E(Z_t \mid \mathcal{F}_{t-}) - Z_{t-1}.$$

The local martingale M_t has a natural role to play in inference as it represents a residual stochastic disturbance after fitting of the trend encapsulated in the finite variation process.

Now if we confine attention to local martingale estimating functions belonging to the class

$$\mathcal{G} = \{G_t(\theta) = \sum_{k=1}^{t} \alpha_k(\theta) m_k(\theta), \alpha_k{}'s \text{ are predictable}\}$$

where EG_tG_t' and $E\dot{G}_T$ are nonsingular, then the optimal choice is given by the quasi-score estimating function

$$Q_t(\theta) = \sum_{k=1}^{t} (E(\dot{m}_k(\theta)|\mathcal{F}_{k-1}))'(E(m_k(\theta)m_k{}'(\theta)|\mathcal{F}_{k-1}))^{+} m_k(\theta) \tag{2}$$

where, as usual, prime denotes transpose, + the Moore-Penrose generalized inverse and the dot refers to differentiation with respect to the components of θ.

In practice, various different functions of the basic data may be used to suggest different local martingales (or martingales) as a basis for different quasi-score estimating functions obtained from application of (2). Comparison and combination of estimating functions must then be considered. These matters are addressed in Heyde (1987), (1989) and are illustrated in Section 4 of this chapter, and Chapter 1 and 3.

3 Branching Processes with Immigration

As a simple illustration of the methodology take the case where $\{X_t\}$ is a subcritical Bienaymé-Galton-Watson branching process with immigration and $\theta' = (m, \lambda)$ is to be estimated on the basis of data $\{X_t, t = 0, 1, \ldots, T\}$, where m and λ are, respectively, the means of the offspring and immigration distributions. This model has been widely used in practice, for example for particle counts in colloidal solutions, and an account of various applications is given in Heyde and Seneta (1972) and Winnicki (1988).

Let $\{X_t\}$ be a sequence of random variables such that

$$X_t = \sum_{i=1}^{X_{t-1}} Y_{t,i} + I_t, \qquad n = 1, 2, \ldots, \tag{3}$$

where X_0 is a nonnegative integer valued r.v.'s and $\{Y_{,i}, i = 1, 2, \ldots, t = 1, 2, \ldots\}$ and $\{I_t, t = 1, 2, \ldots\}$ are independent families of i.i.d. nonnegative integer valued r.v.'s. $\{X_t\}$ is a branching process with immigration. Let $\mathcal{F}_{t-1} = \sigma\{X_i, i \leq t-1\}$. Then, we obtain

$$E(X_t|\mathcal{F}_{t-1}) = mX_{t-1} + \lambda$$

so that the decomposition (1) gives

$$\sum_{t=1}^{T} X_t = m\sum_{t=1}^{T} X_{t-1} + T\lambda + \sum_{t=1}^{T} m_t$$

where

$$m_t = X_t - E(X_t|\mathcal{F}_{t-1})$$

are martingale differences.

Now suppose that X_0 has a distribution for which $E{X_0}^2 < \infty$ and that the variances of the offspring and immigration distributions are $\sigma^2(< \infty)$ and $\eta^2(< \infty)$ respectively. Then, using (2), the quasi-score estimating function based on the martingale $\{\sum_{s=1}^{t} m_s\}$ is

$$Q_T = \begin{pmatrix} \sum_1^T X_{t-1}(\sigma^2 X_{t-1} + \eta^2)^{-1} \ (X_t - mX_{t-1} - \lambda) \\ \sum_1^T (\sigma^2 X_{t-1} + \eta^2)^{-1} \ (X_t - mX_{t-1} - \lambda) \end{pmatrix} \tag{4}$$

which in general involves the nuisance parameters σ^2, η^2. If no immigration is present, so that $\lambda = \eta^2 = 0$, the nuisance parameter σ^2 disappears in the estimating equation $Q_T = 0$. Furthermore, if we have a parametric model based on m, λ alone, such as in the case where the offspring and immigration distributions are both Poisson, there is no nuisance parameter problem. Indeed, in the Poisson case it is straightforward to check that (4) is a constant multiple of the score function which leads to the maximum likelihood estimators.

In general, nuisance parameters must be estimated or avoided. In this case estimation is possible using strongly consistent estimators of σ^2 and η^2 suggested by Yanev and Tchoukova-Dantcheva (1980). We can take these estimators as

$$\begin{aligned} \hat{\sigma}^2(\bar{m}_T, \bar{\lambda}_T) &= \frac{T\sum_{t=1}^T X_{t-1}\hat{U}_t^2 - \sum_{t=1}^T X_{t-1}\sum_{t=1}^T \hat{U}_t^2}{T\sum_{t=1}^T X_{t-1}^2 - (\sum_{t=1}^T X_{t-1})^2}, \\ \hat{\eta}_T^2(\bar{m}_T, \bar{\lambda}_T) &= \frac{\sum_{t=1}^T \hat{U}_t^2 \sum_{t=1}^T X_{t-1}^2 - \sum_{t=1}^T X_{t-1}\sum_{t=1}^T X_{t-1}\hat{U}_t^2}{T\sum_{t=1}^T X_{t-1}^2 - (\sum_{t=1}^T X_{t-1})^2} \end{aligned}$$

where $\hat{U}_t = X_t - \bar{m}_T X_{t-1} - \bar{\lambda}_T$, $\bar{m}_T, \bar{\lambda}_T$ being strongly consistent estimators of m, λ respectively.

Wei and Winnicki (1989) studied an estimating function closely related to (4) in which the term $\sigma^2 X_{t-1} + \eta^2$ is replaced by $X_{t-1} + 1$. Details of the corresponding limit theory are given in Theorem 3.C of Winnicki (1988). Furthermore, for $\hat{m}_T, \hat{\lambda}_T$ based on (4) with estimated σ^2 and η^2 as suggested above (using $\hat{m}_T$, $\hat{\lambda}_T$), a similar analysis to that undertaken by Wei and Winnicki shows that

$$T^{1/2}\begin{pmatrix} \hat{m}_T - m \\ \hat{\lambda}_T - \lambda \end{pmatrix} \xrightarrow{d} N(0, W^{-1})$$

as $T \to \infty$ where, if X has the stationary limit distribution of the $\{X_t\}$ process,

$$W = \begin{pmatrix} EX^2(\sigma^2 X + \eta^2)^{-1} & EX(\sigma^2 X + \eta^2)^{-1} \\ EX(\sigma^2 X + \eta^2)^{-1} & E(\sigma^2 X + \eta^2)^{-1} \end{pmatrix}$$

the same result as one obtains if σ^2 and η^2 are known. The detailed proof can be found in the Appendix to this Chapter(Section 6, Part 1, p.118). The quasi-likelihood framework ensures that this estimation procedure is optimal from the point of view of asymptotic efficiency.

The substitution of estimated values for the nuisance parameters amounts to replacing the quasi- likelihood estimator by an asymptotic quasi-likelihood estimator which has the same asymptotic confidence zones for the unknown parameter. For an account of asymptotic quasi-likelihood see Heyde and Gay (1989) and Chapter 5.

Now it should be noted that earlier approaches to this estimation problem have used conditional least squares or ad hoc methods producing essentially similar results. This amounts to using the estimating function (4) with the terms $\sigma^2 X_{t-1} + \eta^2$ removed. For references and details of the approach see Hall and Heyde (1980, Chapter 6.3). The essential point is that quasi- likelihood or asymptotic quasi-likelihood will offer advantages in asymptotic efficiency and these may be substantial.

The character of the limiting covariance matrices associated with different estimators precludes direct general comparison other than via inequalities. We therefore give a numerical example as a concrete illustration. This concerns the Smoluchowski model for particles in a fluid in which the offspring distribution is assumed to be Poisson and the immigration distribution to be Bernoulli; for a discussion see Heyde and Seneta (1972). The data we use is a set of 505 observations from Fürth (1918, Tabelle 1). We shall compare the asymptotic covariance matrices which come from the methods (i) conditional least squares (e.g. Hall and Heyde (1980, Chapter 6.3), Winnicki (1988, Theorem 3.B)), (ii) Wei and Winnicki's weighted conditional least squares (e.g. Winnicki (1988, Theorem 3.C)), (iii) asymptotic quasi- likelihood. All quantities are estimated from the data and since the covariance matrices are all functions of m, λ common point estimates of m, λ (those which come from (i)) are used as a basis for the comparison. Methods (ii), (iii) produce slightly different point estimates.

We find that $\hat{m} = 0.68, \hat{\lambda} = 0.51$ and that the estimated asymptotic covariance matrices for $(505)^{1/2}\ (\hat{m} - m,\ \hat{\lambda} - \lambda)'$ in the cases (i), (ii), (iii) are, respectively

$$\begin{pmatrix} 0.81 & -1.04 \\ -1.04 & 2.05 \end{pmatrix}, \begin{pmatrix} 0.76 & -0.74 \\ -0.74 & 1.57 \end{pmatrix}, \begin{pmatrix} 0.75 & -0.73 \\ -0.73 & 1.56 \end{pmatrix}.$$

The reduction in variance estimates obtained by using (iii) rather than (i) or (ii) in this particular case is not of practical significance. Approximate 95% confidence intervals for m are 0.68 ± 0.08 in each case while for λ, (i) gives 0.51 ± 0.12 while (ii) and (iii) improve this to 0.51 ± 0.11. Although quasi-likelihood does not perform better than the simpler methods here, it nevertheless remains as the benchmark for minimum width confidence intervals and it does provide a definite improvement in some examples such as the one given below.

In model (3), assume that $Y_{n,i}$, $i = 1, 2, \ldots$, have a Poisson distribution with parameter $\lambda = 0.9$, I_n has Poisson distribution with parameter $\lambda = 3$, $n = 1, 2, \ldots$ and $X_0 = 1$. After simulating, we obtained four groups of 500 data $\{m_{1,i}\}$, $\{\lambda_{1,i}\}$, $\{m_{2,i}\}$ and $\{\lambda_{2,i}\}$, $i = 1, \cdots, 500$, where $\{m_{1,i}\}$ and $\{\lambda_{1,i}\}$ are the estimators coming from the quasi-likelihood method and $\{m_{2,i}\}$ and $\{\lambda_{2,i}\}$ are estimators coming from the formulas in Winnicki(1988). Then we obtained the standard deviation of $\{m_{1,i} - 0.9\}$ as 0.02072 which is less than 0.02123, the standard deviation of $\{m_{2,i} - 0.9\}$, while the standard deviation of $\{\lambda_{1,i} - 3\}$ is 0.5566 which is less than 0.5754 the standard deviation of $\{\lambda_{2,i} - 3\}$. Thus we have an example where the quasi-likelihood method does provide usefully smaller confidence intervals than Wei and Winnicki's method.

Theoretical calculations are given in Part I Section 6 of the Appendix to this chapter.

4 Heteroscedastic Regressions

For the general linear regression model

$$y = X\beta + u \tag{5}$$

where y is an $n \times 1$ vector of observations, X is an $n \times k$ matrix of known constants of rank $k < n$, and β is a $k \times 1$ vector of unknown parameters, heteroscedasticity exists if the diagonal elements of the covariance matrix of u are not all identical. The assumption of heteroscedasticity is necessary in analyses of cross-sectional data on a number of micro economic units, such as firms or households. For example, Prais and Houthakker(1955) found in their analysis of family budgets that the variation of expenditures tends to increase as household income increases.

Theil(1951), Prais(1953), Prais and Houthakker(1955), in their studies of household expenditure functions, assumed that the variances of expenditure observations are proportional to the square of their expectations, that is,

$$E(u_t^2) = \sigma^2(X_t\beta)^2.$$

However, a more general assumption is

$$E(u_t^2) = \sigma^2(X_t\beta)^\gamma,$$

which had been previously suggested by Prais and Aitchison(1954) and Kakwani and Gupta(1967).

Here we consider the model (5) in which u is an $n \times 1$ vector of independent residuals with mean zero and covariance matrix

$$\Omega = \mathrm{diag}(g_1(\theta), \ldots, g_n(\theta))$$

where

$$g_t(\theta) = \sigma^2(X_t\beta)^{2(1-\alpha)},$$

$$X_t = (X_{t,1}, \ldots, X_{t,k}), \quad t = 1, 2, \ldots, n, \quad X = (X_1' \vdots \ \ldots \ \vdots X_n')'.$$

The objective is the efficient estimation of the $(k+2) \times 1$ vector $\theta = (\beta', \sigma^2, \alpha)'$. Using a nonlinear least squares method, Anh(1988) has studied this model and provided estimators of θ which are strongly consistent and asymptotically normally distributed. Here we improve these results by applying quasi-likelihood methods to provide smaller confidence zones than the nonlinear least squares method.

The obvious martingales to use for the model (5) are $\{U_t\}$ and $\{V_t\}$ given by

$$U_t = \sum_{s=1}^{t} u_s, \qquad V_t = \sum_{s=1}^{t}(u_s{}^2 - g_s(\theta)).$$

Using (2), the quasi-score estimating function based on $\{U_t\}$ is

$$\sum_{t=1}^{n} \begin{pmatrix} X_t & 0 & 0 \end{pmatrix}' g_t{}^{-1}(\theta)u_t = \begin{pmatrix} u'\Omega^{-1}X & \vdots & 0 & 0 \end{pmatrix}' \tag{6}$$

which focuses on β, provides no information on σ^2 and involves α as a nuisance parameter. On the other hand, the quasi-score estimating function based on $\{V_t\}$ is

$$\sum_{t=1}^{n} (\dot{g}_t(\theta))'(var\, u_t^2)^{-1}(u_t^2 - g_t(\theta)) = (\frac{\partial\, vec\,\Omega}{\partial\,\theta})' vec\, F^{-1}(\Omega - uu') \tag{7}$$

where

$$F = \mathrm{diag}(var\ u_1^2, \ldots, var\ u_n^2),$$

and this allows estimation of all components of θ provided F is a known function of θ. However, efficiency in the estimation is enhanced if the martingales $\{U_t\}$ and $\{V_t\}$ are used in combination and, using results of Heyde (1987), the combined quasi-score estimating function is

$$\sum_{t=1}^{n} \begin{pmatrix} X_t & 0 & 0 \end{pmatrix}' (g_t(\theta))^{-1}(1 - R_tS_t)^{-1}[u_t + R_t(u_t^2 - g_t(\theta))]$$

$$+ \sum_{t=1}^{n}(\dot{g}_t(\theta))'(var\, u_t^2)^{-1}(1 - R_tS_t)^{-1}[S_tu_t + (u_t^2 - g_t(\theta))] \tag{8}$$

where $R_t = Eu_t^3/varu_t{}^2$, $S_t = Eu_t^3/g_t(\theta)$. This simplifies considerably when $Eu_t^3 = 0$ for each t for then the martingales $\{U_t\}$ and $\{V_t\}$ are orthogonal and (8) is a sum of the quasi-score estimating functions (6) and (7). In this case we may separate the corresponding estimating equation into the two components

$$\begin{aligned} X'\Omega' u \ + \ (\tfrac{\partial\, vec\,\Omega}{\partial\,\beta})' vecF^{-1}(\Omega - uu') &= 0 \\ (\tfrac{\partial\, vec\,\Omega}{\partial\,\gamma})' vecF^{-1}(\Omega - uu') &= 0 \end{aligned} \tag{9}$$

where $\gamma = (\alpha, \sigma^2)'$.

If the u_t are normally distributed, $F = 2\Omega$ and the quasi-likelihood estimator for θ coincides with the maximum likelihood estimator. The equation (9) then corresponds to the true score equations (4.2), (4.3) of Anh (1988). The martingale information (see Heyde (1987)) contained in (5) is

$$I_{QL(U)} = \begin{pmatrix} X'\Omega^{-1}X & O_{k\times 2} \\ O_{2\times k} & O_{2\times 2} \end{pmatrix}$$

which clearly contributes only to the estimation of β while that in (7) is

$$I_{QL(V)} = \sum_{t=1}^{n}(var\, u_t{}^2)^{-1}(\dot{g}_t(\theta))'(\dot{g}_t(\theta)) = A'F^{-1}A,$$

where $A = (\frac{g_i(\theta)}{\partial\theta_j})$, and the combined information in (8) is

$$\begin{aligned} I_{QL(U,V)} = & \begin{pmatrix} X'\Omega^{-1}(I-RS)^{-1}X & \vdots & O_{k\times 2} \\ O_{2\times k} & \vdots & O_{2\times 2} \end{pmatrix} \\ & +A'F^{-1}(I-RS)^{-1}A \\ & + \begin{pmatrix} X'\Omega^{-1}(I-RS)^{-1}RA \\ O_{2\times(k+2)} \end{pmatrix} \\ & + \begin{pmatrix} (A'R(I-RS)^{-1}\Omega^{-1}X & \vdots & O_{(k+2)\times 2} \end{pmatrix} \end{aligned} \tag{10}$$

where $R = \text{diag}(R_1, \ldots, R_n)$, $S = \text{diag}(S_1, \ldots, S_n)$, which of course is $I_{QL(U)} + I_{QL(V)}$ when $Eu_t{}^3 = 0$ for each t.

By comparison with the above estimating functions, Anh (1988) has used non-linear least squares based on minimization of

$$\sum_{t=1}^{n}(e_t^2 - Ee_t{}^2)^2 \tag{11}$$

where the e_t are the (observable) elements of the vector

$$e = (I - X(X'X)^{-1}X')u = Pu = Py,$$

and

$$P = I - X(X'X)^{-1}X.$$

Note that (11) amounts to using the estimating function

$$\sum_{t=1}^{n} \dot{f}_t(\theta)\ (e_t^2 - f_t(\theta))$$

where $f_t(\theta) = E{e_t}^2$. Anh has shown that this leads to a strongly consistent and asymptotically normal estimator $\hat{\theta}_A$, say, satisfying

$$((A'A)(AFA)^{-1}(A'A))^{1/2}(\hat{\theta}_A - \theta) \overset{d}{\approx} N(0, I)$$

for large n under his Conditions 1-3. However, under a similar conditions, we can also obtain

$$(A'F^{-1}A)^{1/2}(\hat{\theta}_{QL(V)} - \theta) \overset{d}{\approx} N(0, I) \tag{12}$$

and

$$I_{QL(U,V)}^{1/2}(\hat{\theta}_{QL(U,V)} - \theta) \overset{d}{\approx} N(0, I) \tag{13}$$

where $\hat{\theta}_{QL(V)}$ and $\hat{\theta}_{QL(U,V)}$ are, respectively, the quasi-likelihood estimators based on (7) and (8). The proof of (12) can be found in Part II Section 6 of this chapter and the proof of (13), which is omitted, is similar to that of (12). By construction we have that

$$I_{QL(V)}^{-1} - I_{QL(U,V)}^{-1}$$

is non-negative definite, while

$$(A'A)^{-1}A'FA(A'A)^{-1} - (A'F^{-1}A)^{-1}$$

is also non-negative definite. This last result holds since, for a covariance matrix F and $n \times p$ matrix A, the Gauss-Markov Theorem gives non-negativity of

$$BFB' - (A'F^{-1}A)^{-1}$$

for every $p \times n$ matrix B satisfying $BA = I_p$ (e.g. Heyde (1989)) and the result holds in particular for

$$B = (A'A)^{-1}A.$$

Thus, from the point of view of asymptotic variance, the order of preference is $\hat{\theta}_{QL(U,V)}$, $\hat{\theta}_{QL(V)}$, $\hat{\theta}_A$.

5 Censoring Problems

In general, censored data depends on censoring r.v.'s $\{U_i\}_{i=1,\ldots,n}$, or $\{V_i\}_{i=1,\ldots,n}$ or both, where n is the size of sample, U_i is the right stopping time and V_i is the left stopping time for the i-th sampled value. Sometimes we also say that the censored data depends on a censoring process $\{C_i(s) = I_{(s\leq U_i)}\}_{i=1,\ldots,n}$ or $\{C_i(s) = I_{(s\geq V_i)}\}_{i=1,\ldots,n}$ or$\{C_i(s) = I_{(V_i\leq s\leq U_i)}\}_{i=1,\ldots,n}$, and the censored data are called right censoring data or left censoring data or symmetrically censored data(even though it may not be exactly symmetrical) respectively.

Let $\{X_i(t)\}_{i=1,\ldots,n}$ be processes in a space $(\Omega, \mathcal{F})$ and $\{\tilde{X}_i(t)\}_{i=1,\ldots,n}$ be censored processes in another space $(\Omega', \mathcal{F}')$. The literature in the censoring field is extensive, such as Arjas and Haara (1984), Jacobsen (1989), Keiding and Gill (1987), Chang and Hsiung (1990) and Andersen, Borgan, Gill, Keiding(1988), and many papers have given methods for estimation of the parameters in $\{X_i(t)\}_{i=1,\ldots,n}$ using censored data and assuming $\{X_i(t)\}_{i=1,\ldots,n}$ are counting processes or the likelihood functions of $\{X_i(t)\}_{i=1,\ldots,n}$ are known. However, it is not only counting processes which can have censoring problems and also it sometimes happens that not all the likelihood functions of $\{X_i(t)\}_{i=1,\ldots,n}$ can be written down . Therefore we face the general question of how to solve the censoring problem when the process is not a counting process or the likelihood function of the process is unknown.

We shall introduce a method based on quasi-likelihood to solve this question under certain circumstances.

Here we shall only consider right censoring problems; similar ideas can be employed for other kinds of censoring problems.

Consider a multivariate process composed of n individual processes, $X(t) = (X_1(t), \ldots, X_n(t))$. Right censoring will often introduce extra variation in which case we first have to enlarge the filtrations compared with the $\{\mathcal{F}_t\}$ considered for the uncensored sample.

Suppose that $X(t) = (X_1(t), \ldots, X_n(t))$ is defined on some space $(\Omega, \mathcal{F})$ with respect to some filtration $\{\mathcal{F}_t\}$ and a family of probability measures

$$\mathcal{P} = \{P_\theta, \theta \in \Theta\}.$$

Right censoring of $X(t)$ is the situation in which observation of $X_i(t)$ is ceased after some (possibly random) time U_i, $i = 1, 2, \ldots, n$, i.e. $X_i(t)$ is only observed on the random set $E_i = \{t \leq U_i\} \subseteq \mathcal{T}$, where $\mathcal{T}$ is some σ-field, or equivalently where the process

$$C_i(t) = I_{(t \in E_i)} = I_{(t \leq U_i)}$$

is unity. Thus, right censoring is imposed on X by individual right censoring processes $C_1(\cdot), \ldots, C_n(\cdot)$. We may assume that the censoring process $C = (C_i(\cdot), i = 1, 2, \ldots, n)$ is predictable with respect to $\{\mathcal{F}_t\}$(see, Andersen, Borgan, Gill, Keiding (1988)).

Suppose that, for each $i = 1, 2, \ldots, n$,

$$X_i(t) = \int_0^t f_i(\theta, s) d\lambda_s + M_{i,t}, \tag{14}$$

where, for each i fixed, $\{M_{i,t}\}$ is a martingale with respect to $\{\mathcal{F}_t\}$, $\{f_i(t)\}$ is a predictable process with respect to $\{\mathcal{F}_t\}$ and λ_t is an increasing function. This model is more general than those discussed in Andersen, Borgan, Gill, Keiding (1988), because it is free from the distribution. Before applying the quasi-likelihood method to this model, we need the following additional assumptions:

(i) $\{\mathcal{F}_t\}$ can be enlarged to $\{\mathcal{F}_{i,t}\}_{i=1,2,\ldots,n}$ such that, for i fixed, $M_{i,t}$ is a martingale with respect to $\mathcal{F}_{\cdot,t} \hat{=} \bigvee_i \mathcal{F}_{i,t}$ and, for each t fixed, $M_{i,t}$ is a martingale difference with respect to $\mathcal{F}_{i,\cdot} \hat{=} \bigvee_t \mathcal{F}_{i,t}$.

(ii) $f_i(\theta, t)$ is predictable with respect to $\mathcal{F}_{i,t}$.

These assumptions are obviously true when $\{X_i\}_{i=1,2,\ldots,n}$ are mutually independent.

Let $\tilde{X}_i(t)$ be the censored process, $i = 1, 2, \ldots, n$. Then

$$\tilde{X}_i(t) = \int_0^t C_i(s) dX_i(s)$$

$$= \int_0^t C_i(s) f_i(\theta, s) d\lambda_s + \int_0^t C_i(s) dM_{i,s}.$$

Since $C_i(s)$ is $\mathcal{F}_{.,s-1}$ measurable, $\int_0^t C_i(s)dM_{i,s}$ is a martingale with respect to $\mathcal{F}_{.,t}$, $i = 1, 2, \ldots, n$.

Proposition 1 *If U_i is $\mathcal{F}_{i-1,.}$-measurable, and $E(|M_i(U_i)||\mathcal{F}_{i-1,.}) < \infty$, then*

$$E(M_i(U_i)|\mathcal{F}_{i-1,.}) = 0.$$

Proof: Since U_i is $\mathcal{F}_{i-1,.}$-measurable, there are simple functions $\{\sum_{j=1}^n \alpha_{i,j}^{(n)} I_{E_{i,j}^{(n)}}\}$ such that

$$U_i = \lim_{n\to\infty} \sum_{j=1}^n \alpha_{i,j}^{(n)} I_{E_{i,j}^{(n)}},$$

where the $E_{i,j}^{(n)}$ are $\mathcal{F}_{i-1,.}$-measurable and the $\alpha_{i,j}^{(n)}$ are constants. Following the assumption $E(|M_i(U_i)||\mathcal{F}_{i-1,.}) < \infty$ and $E(M_i(t)|\mathcal{F}_{i-1,.}) = 0$ for all t fixed, we obtain $E(M_i(U_i)|\mathcal{F}_{i-1,.}) = 0$ by Fubini's Theorem.

Using Proposition 1, if U_i is $\mathcal{F}_{i-1,.}$-measurable and $E(|M_i(U_i)||\mathcal{F}_{i-1,.}) < \infty$, then, for all t fixed, $\{\int_0^t C_i(s)dM_i(s)\}$ is a martingale difference sequence with respect to $\{\mathcal{F}_{i,.}\}$. Therefore, when $\{U_i\}_{i=1,2,\ldots,n}$ are mutually independent and $E(M_i(t)) < \infty$ for all t, the assumptions:

(iii) U_i is $\mathcal{F}_{i-1,.}$-measurable, $i = 1, 2, \ldots, n$,

(iv) $\{\int_0^t C_i(s)dM_i(s)\}$ is a martingale difference sequence with respect to $\{\mathcal{F}_{i,.}\}$, will always hold.

Now consider

$$\sum_{i=1}^n \tilde{X}_i(t) = \sum_{i=1}^n \int_0^t C_i(s)f_i(\theta, s)d\lambda_s + \sum_{i=1}^n \int_0^t C_i(s)dM_{i,s}.$$

If Assumptions (i)-(iv) hold, then $\int_0^t C_i(s)f_i(\theta, s)d\lambda_s$ is $\mathcal{F}_{i-1,.}$-measurable.

Let $\mathcal{G}_{n,t}$ be an estimating function space in which every element $G_n(t)$ of $\mathcal{G}_{n,t}$ has the form:

$$\sum_{i=1}^n \beta_i(t) \int_0^t C_i(s)dM_{i,s},$$

where $\beta_i(t)$ is $\mathcal{F}_{i-1,.}$-measurable, and $EG_n(t)G_n'(t)$, $E(\dot{G}_n(t))$ are nonsingular.

Under suitable regularity conditions, a quasi-score estimating function

$$G_n^*(t) = \sum_{i=1}^n \left(\frac{\partial}{\partial\theta} \int_0^t C_i(s)f_i(\theta, s)d\lambda_s\right) \frac{\tilde{X}_i(t) - \int_0^t C_i(s)f_i(\theta, s)d\lambda_s}{E((\int_0^t C_i(s)dM_{i,s})^2|\mathcal{F}_{i-1,.})}$$

can be obtained within $\mathcal{G}_{n,t}$. Then, letting $G_n^*(t) = 0$, we can get the quasi-likelihood estimator $\theta^*_{n,t}$ for the parameter θ in equation (14).

Remarks: 1. If $\{\int_0^t C_i(s)dM_{i,s}\}_{i=1,2,\ldots,n}$ are mutually independent, then

$$E((\int_0^t C_i(s)dM_{i,s})^2|\mathcal{F}_{i-1,.}) = E(\int_0^t C_i(s)dM_i(s))^2$$
$$= E\int_0^t C_i^2(s)d < M_i >_s .$$

2. For $G_n^*(t)$, we can provide the conditions under which $\{\theta_n^*(t)\}$ is strongly consistent and also to obtain confidence intervals for θ using the results of Chapters 3 and 8.

Next we give an example. Suppose that $X(t)$ satisfies the model

$$X(t) = \int_0^t s^\theta ds + B_t, \qquad 0 \le t \le 3,$$

where B_t is Brownian motion. Assume that the true parameter $\theta = 2$. Let $X_1(t), \ldots, X_{100}(t)$ be samples from $X(t)$. Suppose that the censoring variables U_i, $i = 1, 2, \ldots, 100$ are given by:

$$\begin{aligned} U_i &= 2, & i &= 1, \ldots, 30, \\ U_i &= 2.5, & i &= 31, \ldots, 60, \\ U_i &= 3, & i &= 61, \ldots, 100. \end{aligned}$$

After simulating, we obtained the simulated censored samples

$$\begin{aligned} \tilde{X}_i(3) &= X_i(2), & i &= 1, \ldots, 30, \\ \tilde{X}_i(3) &= X_i(2.5), & i &= 31, \ldots, 60, \\ \tilde{X}_i(3) &= X_i(3), & i &= 61, \ldots, 100. \end{aligned}$$

Then the quasi-score estimating function is

$$G_n^*(t) = \sum_{i=1}^n (\frac{\partial}{\partial\theta}\int_0^3 C_i(s)s^\theta ds)\frac{\tilde{X}_i(3) - \int_0^3 C_i(s)s^\theta ds}{\int_0^3 C_i^2 ds}$$
$$= \sum_{i=1}^n (-\frac{U_i^{\theta+1}}{(\theta+1)^2} + \frac{U_i^{\theta+1} log U_i}{\theta+1})\frac{\tilde{X}_i(3) - U_i^{\theta+1}/(\theta+1)}{U_i}.$$

Let $G_n^*(t) = 0$. From the graph below we get $\theta_n^*(t) = 1.9$ which is very close to the true value 2.

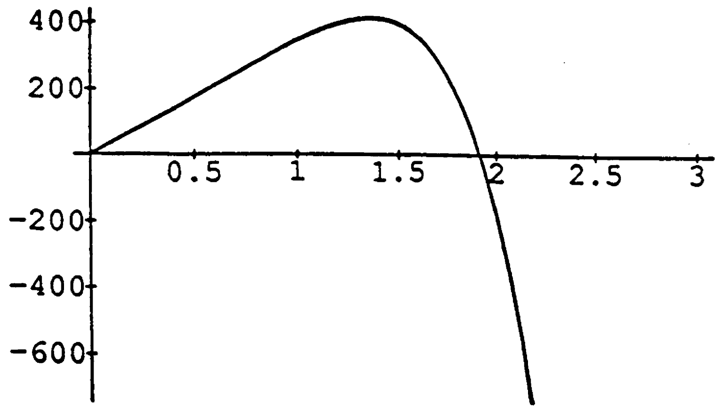

FIGURE 1

6 Appendix

Part I

Here we provide the detailed theoretical calculations of the last example in Section 3 of this chapter.

We discuss the model (3) under the following assumptions: $m_0 = EY_{n,i} < 1$, $\lambda_0 = EI_n < \infty$, $\sigma^2 = var(Y_{n,i}) < \infty$ and $\eta^2 = var(I_n) < \infty$. We also suppose that $E|Y_{n,i}|^k < \infty$ and $E|I_n|^k < \infty$ for $k \leq 4$, and X_0 has the stationary distribution.

The terminology " the assumptions of the model" means the above assumptions.

Following the result of Heathcote (1965, 1966), augmented by Quine (1970), we have

Theorem 1 *If $m_0 < 1$, then $\{X_n\}$ is an irreducible aperiodic, positive recurrent Markov chain. In this case $X_n \xrightarrow{p} X$ where X is a random variable whose support is $\{k|P(X_1 = k) > 0\}$.*

Under the assumptions of the model, the process $\{X_n\}$ is stationary and ergodic. Ergodicity follows from the uniqueness of the stationary distribution(see, Billingsley(1961), p52, or Hall and Heyde(1980), p179). From the ergodic theorem we obtain that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \frac{X_i^2}{\sigma^2 X_i + \eta^2} = E\frac{X^2}{\sigma^2 X + \eta^2}, \qquad a.s., \tag{15}$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \frac{X_i}{\sigma^2 X_i + \eta^2} = E\frac{X}{\sigma^2 X + \eta^2}, \qquad a.s., \tag{16}$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sigma^2 X_i + \eta^2} = E\frac{1}{\sigma^2 X + \eta^2}, \qquad a.s.. \tag{17}$$

We also have the following proposition.

Proposition 2 *Under the assumptions of the model,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \frac{X_i^2}{(\sigma^2 X_i + \eta^2)^2}(X_{i+1} - m_0 X_i - \lambda_0)^2 = E\frac{X^2}{\sigma^2 X + \eta^2}, \qquad (P), \tag{18}$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \frac{X_i}{(\sigma^2 X_i + \eta^2)^2}(X_{i+1} - m_0 X_i - \lambda_0)^2 = E\frac{X}{\sigma^2 X + \eta^2}, \qquad (P), \tag{19}$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(\sigma^2 X_i + \eta^2)^2}(X_{i+1} - m_0 X_i - \lambda_0)^2 = E\frac{1}{\sigma^2 X + \eta^2}, \qquad (P), \tag{20}$$

Proof: Because of the similarity of the method of proof only that of (18) is given here.

Due to (15), we only need to prove

$$\frac{1}{n}\sum_{i=1}^{n} \frac{X_i^2}{(\sigma^2 X_i + \eta^2)^2}(X_{i+1} - m_0 - \lambda_0)^2 - \sum_{i=1}^{n} \frac{X_i^2}{(\sigma^2 X_i + \eta^2)}$$

$$\xrightarrow{P} 0, \qquad as \quad n \to \infty.$$

Following Theorem 2.23 in Hall and Heyde (1980), the only things we need to check are

$$\sup_n P\left(\frac{1}{n}\sum_{i=1}^{n} E\left(\frac{X_i^2}{(\sigma^2 X_i + \eta^2)^2}(X_{i+1} - m_0 - \lambda_0)^2 | \mathcal{F}_i\right) > \lambda\right) \to 0, \tag{21}$$

$$as \quad \lambda \to \infty,$$

and, for all $\varepsilon > 0$,

$$\sum_{i=1}^{n} E[\frac{1}{n}\frac{X_i^2}{(\sigma^2 X_i + \eta^2)^2}(X_{i+1} - m_0 - \lambda_0)^2 \tag{22}$$

$$\times I_{(|\sum_{i=1}^{n} \frac{X_i^2}{(\sigma^2 X_i + \eta^2)^2}(X_{i+1} - m_0 - \lambda_0)^2| > n\varepsilon)} |\mathcal{F}_i] \xrightarrow{p} 0, \qquad as \quad n \to \infty.$$

Since

$$\sup_n P(\frac{1}{n}\sum_{i=1}^{n} E(\frac{X_i^2}{(\sigma^2 X_i + \eta^2)^2}(X_{i+1} - m_0 - \lambda_0)^2 |\mathcal{F}_i) > \lambda)$$

$$\leq \sup_n (\sum_{i=1}^{n} E\frac{X_i^2}{(\sigma^2 X_i + \eta^2)})/(n\lambda),$$

and

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} E\frac{X_i^2}{(\sigma^2 X_i + \eta^2)} = E\frac{X^2}{(\sigma^2 X + \eta^2)},$$

(21) is obviously true.

Now we consider (22). For any $\varepsilon > 0$, and $\delta > 0$,

$$P(|\sum_{i=1}^{n} E[\frac{1}{n}\frac{X_i^2}{(\sigma^2 X_i + \eta^2)^2}(X_{i+1} - m_0 - \lambda_0)^2$$

$$\times I_{(|\sum_{i=1}^{n} \frac{X_i^2}{(\sigma^2 X_i + \eta^2)^2}(X_{i+1} - m_0 - \lambda_0)^2| > n\varepsilon)} |\mathcal{F}_i]| > \delta)$$

$$\leq 1/\delta \sum_{i=1}^{n} E[\frac{1}{n}\frac{X_i^2}{(\sigma^2 X_i + \eta^2)^2}(X_{i+1} - m_0 - \lambda_0)^2$$

$$\times I_{(|\sum_{i=1}^{n} \frac{X_i^2}{(\sigma^2 X_i + \eta^2)^2}(X_{i+1} - m_0 - \lambda_0)^2| > n\varepsilon)}]$$

$$\leq (n\delta)^{-1} \sum_{i=1}^{n} [\sqrt{E\frac{X_i^4}{(\sigma^2 X_i + \eta^2)^4}(X_{i+1} - m_0 - \lambda_0)^4}$$

$$\times \sqrt{P(|\sum_{i=1}^{n} \frac{X_i^2}{(\sigma^2 X_i + \eta^2)^2}(X_{i+1} - m_0 - \lambda_0)^2| > n\varepsilon)}\,]$$

$$\leq (n\delta)^{-1} \sum_{i=1}^{n} \sqrt{(n\varepsilon)^{-1} E\frac{X_i^4}{(\sigma^2 X_i + \eta^2)^4}(X_{i+1} - m_0 - \lambda_0)^4 E\frac{X_i^2}{(\sigma^2 X_i + \eta^2)}}$$

$$\leq (n\delta)^{-1} \sqrt{(\varepsilon)^{-1} \sum_{i=1}^{n} (E\frac{X_i^4}{(\sigma^2 X_i + \eta^2)^4}(X_{i+1} - m_0 X_i - \lambda_0)^4 E\frac{X_i^2}{(\sigma^2 X_i + \eta^2)})}$$

$$\leq (\sqrt{n}\sqrt{\varepsilon}\delta)^{-1}\{n^{-1} \sum_{i=1}^{n} [E\frac{X_i^2}{(\sigma^2 X_i + \eta^2)}(\frac{E(Y - m_0)^4}{\sigma^8} EX_i + \frac{E(X_i(X_i + 1))}{2!\sigma^4}$$

$$+\frac{6\eta^2 EX_i}{\sigma^4}+\frac{1}{\sigma^8}E(I-\lambda_0)^4)]]\}^{1/2}$$

$$\to 0, \qquad as \quad n\to\infty.$$

Therefore we obtain (22) and Proposition 2 is proved.

From the model (3), we can construct a martingale $\{M_n\}$ given by

$$M_n=\sum_{i=1}^{n}(X_{i+1}-mX_i-\lambda),$$

with respect to the natural σ-fields $\mathcal{F}_n$ generated by $\{M_n\}$. Let $\mathcal{G}_n$ be an estimating function space in which every element has form $G_n=\sum_{i=1}^{n}\alpha_i(M_i-M_{i-1})$, where $\{\alpha_i\}$ is predictable with respect to $\{\mathcal{F}_n\}$ and G_n satisfies the conditions that $EG_nG_n{}'$ and $E\dot{G}_n$ are nonsingular. Following the theory of the quasi-likelihood method, if there is a quasi-score estimating function G_n^* in $\mathcal{G}_n$, then the G_n^* is

$$G_n^*=-\sum_{i=1}^{n}\begin{pmatrix}X_i\\1\end{pmatrix}\frac{X_{i+1}-mX_i-\lambda}{\sigma X_i+\eta^2}.$$

Let $\hat{m}_n$ and $\hat{\lambda}_n$ be the quasi-likelihood estimators of m_0 and λ_0 respectively. Then

$$0=G_n^*(\hat{m}_n,\hat{\lambda}_n)=G_n^*(m_0,\lambda_0)+\dot{G}_n^*(m_0,\lambda_0)\begin{pmatrix}\hat{m}_n-m_0\\ \hat{\lambda}_n-\lambda_0\end{pmatrix},$$

where $\dot{G}_n^*=(\frac{\partial}{\partial m}G_n(m,\lambda),\frac{\partial}{\partial\lambda}G_n(m,\lambda))'$.

In the following, we shall prove

$$(EG_n^*G_n^{*\prime})^{-1/2}G_n^*(m_0,\lambda_0)\xrightarrow{d}N(0,I), \qquad as \quad n\to\infty.$$

Denoting

$$(EG_n^*G_n^{*\prime})^{-1/2}G_n^*(m,\lambda)=\sum_{i=1}^{n}\begin{pmatrix}a_{i,1}\\a_{i,2}\end{pmatrix},$$

and following Theorem 1 in Chapter 8, we need to prove

$$\max_{i\le n}|a_{i,j}|\xrightarrow{p}0, \qquad as \quad n\to\infty, \quad j=1,2, \tag{23}$$

$$\sum_{i=1}^{n}\begin{pmatrix}a_{i,1}\\a_{i,2}\end{pmatrix}\begin{pmatrix}a_{i,1} & a_{i,2}\end{pmatrix}\xrightarrow{p}I, \qquad as \quad n\to\infty, \tag{24}$$

and

$$E\max_{i\le n}(|a_{i,j}a_{i,k}|) \tag{25}$$

is bounded in n, where $j,k=1,2$.

Utilizing (15)-(20), it is easy to show that (24) is true. Now we consider (23). For any $\varepsilon>0$, due to (24), there is a $N_0(\varepsilon)>0$ such that, if $n\ge N_0$,

$$P(|tr\sum_{i=N_0}^{n}\begin{pmatrix} a_{i,1}\\ a_{i,2}\end{pmatrix}\begin{pmatrix} a_{i,1} & a_{i,2}\end{pmatrix}|>\varepsilon)<\varepsilon/2.$$

Fixing this $N_0(\varepsilon)$, since

$$EG_n^*G_n^{*\prime}=\begin{pmatrix} E\sum_{i=1}^n\frac{X_i^2}{\sigma^2X_i+\eta^2} & E\sum_{i=1}^n\frac{X_i}{\sigma^2X_i+\eta^2}\\ E\sum_{i=1}^n\frac{X_i}{\sigma^2X_i+\eta^2} & E\sum_{i=1}^n\frac{1}{\sigma^2X_i+\eta^2}\end{pmatrix}$$

$$\sim n\begin{pmatrix} E\frac{X^2}{\sigma^2X+\eta^2} & E\frac{X}{\sigma^2X+\eta^2}\\ E\frac{X}{\sigma^2X+\eta^2} & E\frac{1}{\sigma^2X+\eta^2}\end{pmatrix},$$

we obtain, after some calculations,

$$P(tr\begin{pmatrix} a_{i,1}\\ a_{i,2}\end{pmatrix}\begin{pmatrix} a_{i,1} & a_{i,2}\end{pmatrix}>\varepsilon)$$

$$\le(\varepsilon)^{-1}tr[(EG_n^*G_n^{*\prime})^{-1/2}\begin{pmatrix} E\frac{X_i^2}{\sigma^2X_i+\eta^2} & E\frac{X_i}{\sigma^2X_i+\eta^2}\\ E\frac{X_i}{\sigma^2X_i+\eta^2} & E\frac{1}{\sigma^2X_i+\eta^2}\end{pmatrix}(EG_n^*G_n^{*\prime})^{-1/2}]$$

$$\sim(n\varepsilon)^{-1}tr[(\begin{pmatrix} E\frac{X_i^2}{\sigma^2X_i+\eta^2} & E\frac{X_i}{\sigma^2X_i+\eta^2}\\ E\frac{X_i}{\sigma^2X_i+\eta^2} & E\frac{1}{\sigma^2X_i+\eta^2}\end{pmatrix})^{-1/2}\begin{pmatrix} E\frac{X_i^2}{\sigma^2X_i+\eta^2} & E\frac{X_i}{\sigma^2X_i+\eta^2}\\ E\frac{X_i}{\sigma^2X_i+\eta^2} & E\frac{1}{\sigma^2X_i+\eta^2}\end{pmatrix}$$

$$\times(\begin{pmatrix} E\frac{X_i^2}{\sigma^2X_i+\eta^2} & E\frac{X_i}{\sigma^2X_i+\eta^2}\\ E\frac{X_i}{\sigma^2X_i+\eta^2} & E\frac{1}{\sigma^2X_i+\eta^2}\end{pmatrix})^{-1/2}]\to 0,\qquad as\quad n\to\infty,$$

for all $i\le N_0$. Therefore there is a $N_1(\varepsilon)>N_0(\varepsilon)>0$ such that, if $n\ge N_1(\varepsilon)$,

$$P(\max_i tr(\begin{pmatrix} a_{i,1}\\ a_{i,2}\end{pmatrix}\begin{pmatrix} a_{i,1} & a_{i,2}\end{pmatrix})>\varepsilon)$$

$$\le P(|tr\sum_{i=N_0}^{n}\begin{pmatrix} a_{i,1}\\ a_{i,2}\end{pmatrix}\begin{pmatrix} a_{i,1} & a_{i,2}\end{pmatrix}|>\varepsilon)+\sum_{i=1}^{N_0}P(|tr\begin{pmatrix} a_{i,1}\\ a_{i,2}\end{pmatrix}\begin{pmatrix} a_{i,1} & a_{i,2}\end{pmatrix}|>\varepsilon)$$

$$\leq \varepsilon,$$

i.e. (23) is true.

Finally, we prove (25). Since for all $j, k = 1, 2$,

$$E \max_{i \leq n}(|a_{i,j} a_{i,k}|)$$

$$\leq 1/2 E(\max_{i \leq n}(a_{i,j}^2 + a_{i,k}^2))$$

$$\leq \sum_{i=1}^{n} trE[\begin{pmatrix} a_{i,1} \\ a_{i,2} \end{pmatrix} \begin{pmatrix} a_{i,1} & a_{i,2} \end{pmatrix}] \to 2, \qquad as \quad n \to \infty.$$

which forces (25) to hold. Therefore

$$(EG_n^* G_n^{*\prime})^{-1/2} G_n^* \xrightarrow{d} N(0, I), \qquad as \quad n \to \infty,$$

and

$$(EG_n^* G_n^{*\prime})^{-1/2} \dot{G}_n^* \begin{pmatrix} \hat{m}_n - m_0 \\ \hat{\lambda}_n - \lambda_0 \end{pmatrix} \xrightarrow{d} N(0, I), \qquad as \quad n \to \infty.$$

Because of

$$\dot{G}_n^* = \begin{pmatrix} \sum_{i=1}^n \frac{X_i^2}{\sigma^2 X_i + \eta^2} & \sum_{i=1}^n \frac{X_i}{\sigma^2 X_i + \eta^2} \\ \sum_{i=1}^n \frac{X_i}{\sigma^2 X_i + \eta^2} & \sum_{i=1}^n \frac{1}{\sigma^2 X_i + \eta^2} \end{pmatrix}$$

and

$$(EG_n^* G_n^{*\prime})^{-1/2} \dot{G}_n^* (EG_n^* G_n^{*\prime})^{-1/2} \xrightarrow{p} I,$$

by (15)-(17), we also can obtain

$$(EG_n^* G_n^{*\prime})^{1/2} \begin{pmatrix} \hat{m}_n - m_0 \\ \hat{\lambda}_n - \lambda_0 \end{pmatrix} \xrightarrow{d} N(0, I), \qquad as \quad n \to \infty,$$

i.e.

$$n^{1/2} \begin{pmatrix} \hat{m}_n - m \\ \hat{\lambda}_n - \lambda \end{pmatrix} \xrightarrow{d} N(0, W^{-1})$$

as $n \to \infty$, where

$$W = \begin{pmatrix} EX^2(\sigma^2 X + \eta^2)^{-1} & EX(\sigma^2 X + \eta^2)^{-1} \\ EX(\sigma^2 X + \eta^2)^{-1} & E(\sigma^2 X + \eta^2)^{-1} \end{pmatrix}.$$

We now leave the above results and consider a result of Winnicki(1988) in which he used the weighted conditional least squares method to deal with this same model. Let $\tilde{m}_n$ and $\tilde{\lambda}_n$ be the weighted conditional least squares estimators of m_0 and λ_0 respectively. Winnicki showed that

$$\sqrt{n}(\tilde{m}_n - m_0, \tilde{\lambda}_n - \lambda_0) \xrightarrow{d} N(0, \Phi), \qquad as \quad n \to \infty,$$

where $\Phi = V^{-1}WV^{-1}$,

$$V = \begin{pmatrix} EX & 1 \\ E\frac{X}{1+X} & E\frac{1}{1+X} \end{pmatrix},$$

and

$$W = \begin{pmatrix} E(\sigma^2 X + \eta^2) & E\frac{\sigma^2 X+\eta^2}{1+X} \\ E\frac{\sigma^2 X+\eta^2}{1+X} & E\frac{\sigma^2 X+\eta^2}{(1+X)^2} \end{pmatrix}.$$

For the convenience of calculating, we may assume $\sigma^2 = \eta^2 = 1$. Therefore

$$V'W^{-1}V = \lim_{n\to\infty} 1/nB_n,$$

where

$$B_n = \begin{pmatrix} \sum_{i=1}^n EX_i & \sum_{i=1}^n E\frac{X_i}{1+X_i} \\ n & \sum_{i=1}^n E\frac{1}{1+X_i} \end{pmatrix} \begin{pmatrix} \sum_{i=1}^n E(1+X_i) & n \\ n & \sum_{i=1}^n E\frac{1}{1+X_i} \end{pmatrix}^{-1}$$

$$\times \begin{pmatrix} \sum_{i=1}^n EX_i & n \\ \sum_{i=1}^n E\frac{X_i}{1+X_i} & \sum_{i=1}^n E\frac{1}{1+X_i} \end{pmatrix}.$$

In this case, we obtain

$$B_n^{-1/2} \begin{pmatrix} \tilde{m}_n - m_0 \\ \tilde{\lambda}_n - \lambda_0 \end{pmatrix} \xrightarrow{d} N(0, I), \qquad as \quad n \to \infty.$$

Therefore the asymptotic confidence zone under the Winnicki's approach is determined by B_n.

Choosing an estimating function

$$G_n = \sum_{i=1}^n \begin{pmatrix} b_{i,1} \\ b_{i,2} \end{pmatrix} (X_{i+1} - mX_i - \lambda).$$

from $\mathcal{G}_n$. Then after some algebra we have

$$E\dot{G}_n = -\sum_{i=1}^{n} E \begin{pmatrix} b_{i,1}X_i & b_{i,1} \\ b_{i,2}X_i & b_{i,2} \end{pmatrix},$$

and

$$EG_nG_n' = \begin{pmatrix} \sum_{i=1}^{n} Eb_{i,1}^2(X_i+1) & \sum_{i=1}^{n} Eb_{i,1}b_{i,2}(X_i+1) \\ \sum_{i=1}^{n} Eb_{i,1}b_{i,2}(X_i+1) & \sum_{i=1}^{n} Eb_{i,2}^2(X_i+1) \end{pmatrix}.$$

Let $b_{i,1} = 1$ and $b_{i,2} = (1+X_i)^{-1}$; it is interesting to get

$$(E\dot{G}_n)'(EG_nG_n')^{-1}(E\dot{G}_n) = B_n.$$

By the definition of the quasi-score estimating function $\dot{G}_n^*$, we have

$$(E\dot{G}_n^*)'(EG_n^*G_n^{*\prime})^{-1}(E\dot{G}_n^*) - (E\dot{G}_n)'(EG_nG_n')^{-1}(E\dot{G}_n)$$

$$= (EG_n^*G_n^{*\prime}) - B_n \geq 0.$$

Therefore, following Theorem 2 in Chapter 5, the asymptotic confidence zones determined by the quasi-likelihood method are smaller than those determined by the weighted least squares method.

Part II

Here we establish (12) under certain conditions.

Consider a linear model

$$y = X'\beta + u, \tag{26}$$

where y is an $n \times 1$ random vector, $X = (X_1, \ldots, X_n)'$ is an $n \times k$ nonrandom matrix($k < n$), where $X_t = (X_{t,1}, \ldots, X_{t,k})'$, β is a $k \times 1$ vector of unknown parameter and u is an $n \times 1$ random vector with diagonal covariance matrix

$$\Omega = \sigma^2 diag((X_1\beta)^{2(1-\alpha)}, \ldots, (X_n\beta)^{2(1-\alpha)}).$$

The vector y is observable while the vector u is unobservable. Denoted by $\theta = (\beta', \sigma^2, \alpha)'$ a $(k+2) \times 1$ vector.

Assumption 1. The elements of x are uniformly bound. The parameter space Θ is a compact subset of R_{k+2}. The true parameter θ_0 is an interior point of Θ.

Assumption 2. For each i, Eu_i^4 is uniformly bounded.

Let $\Psi_n = \{\sum_{i=1}^n \alpha_i(u_i^2 - E(u_i^2|\mathcal{F}_{i-1})\}$ be an estimating function space, where $\{\alpha_i\}$ is a predictable process with respect to $\{\mathcal{F}_i\}$ and $\mathcal{F}_i = \sigma\{u_1, u_2, \ldots, X_0, \ldots, X_{i+1}\}$. Let $\mathcal{G}_n$ be a subset of Ψ such that $G_n \in \mathcal{G}_n$ iff EG_nG_n' and $E\dot{G}_n$ are nonsingular. Denote by G_n^* the quasi-likelihood estimating function in $\mathcal{G}_n$, then following the theory of the quasi-likelihood method we have

$$G_n^* = \sum_{i=1}^n \begin{pmatrix} (\alpha-1)\sigma^2(X_i\beta)^{1-2\alpha}X_{i,1} \\ \vdots \\ (\alpha-1)\sigma^2(X_i\beta)^{1-2\alpha}X_{i,k} \\ 1/2(X_i\beta)^{2(1-\alpha)} \\ \sigma^2(X_i\beta)^{2(1-\alpha)}log(X_i\beta) \end{pmatrix} \frac{u_i^2 - \sigma^2(X_i\beta)^{2(1-\alpha)}}{\sigma^4(X_i\beta)^{4(1-\alpha)}}.$$

Let

$$\begin{aligned} C_n(\theta) &= \sum_{i=1}^n \frac{2}{\sigma^4(X_i\beta)^{4(1-\alpha)}} \begin{pmatrix} (\alpha-1)\sigma^2(X_i\beta)^{1-2\alpha}X_{i,1} \\ \vdots \\ (\alpha-1)\sigma^2(X_i\beta)^{1-2\alpha}X_{i,k} \\ 1/2(X_i\beta)^{2(1-\alpha)} \\ \sigma^2(X_i\beta)^{2(1-\alpha)}log(X_i\beta) \end{pmatrix} \begin{pmatrix} (\alpha-1)\sigma^2(X_i\beta)^{1-2\alpha}X_{i,1} \\ \vdots \\ (\alpha-1)\sigma^2(X_i\beta)^{1-2\alpha}X_{i,k} \\ 1/2(X_i\beta)^{2(1-\alpha)} \\ \sigma^2(X_i\beta)^{2(1-\alpha)}log(X_i\beta) \end{pmatrix}' \\ &= A_n' F_n^{-1} A_n. \end{aligned}$$

Assumption 3. As $n \to \infty$, $n^{-1}A_n'F_n^{-1}A_n = n^{-1}C_n$ exists and is positive definite for all $\theta \in \Theta$.

After some algebra we can obtain $EG_n^* =< G^* >_n= C_n$ and, following Assumptions 1-3 and applying Theorem 2.7.5.a in Révész(1968), we can also get the following facts:

(i)

$$C_n^{-1}(\theta)\sum_{i=1}^{n}\frac{\partial}{\partial\theta}[\begin{pmatrix}(\alpha-1)\sigma^2(X_i\beta)^{1-2\alpha}X_{i,1}\\ \vdots\\ (\alpha-1)\sigma^2(X_i\beta)^{1-2\alpha}X_{i,k}\\ 1/2(X_i\beta)^{2(1-\alpha)}\\ \sigma^2(X_i\beta)^{2(1-\alpha)}log(X_i\beta)\end{pmatrix}\frac{1}{\sigma^4(X_i\beta)^{4(1-\alpha)}}](u_i-\sigma^2(X_i\beta)^{2(1-\alpha)})$$

$$=(1/nC_n)^{-1}1/n\sum_{i=1}^{n}\frac{\partial}{\partial\theta}[\begin{pmatrix}(\alpha-1)\sigma^2(X_i\beta)^{1-2\alpha}X_{i,1}\\ \vdots\\ (\alpha-1)\sigma^2(X_i\beta)^{1-2\alpha}X_{i,k}\\ 1/2(X_i\beta)^{2(1-\alpha)}\\ \sigma^2(X_i\beta)^{2(1-\alpha)}log(X_i\beta)\end{pmatrix}\frac{1}{\sigma^4(X_i\beta)^{4(1-\alpha)}}](u_i-\sigma^2(X_i\beta)^{2(1-\alpha)})$$

$$\xrightarrow{a.s.} 0, \qquad \text{as } n\to\infty,$$

i.e.

$$(EG_n^*)^{-1}\dot{G}_n^* =< G^* >_n^{-1} \dot{G}_n^* = C_n^{-1}\dot{G}_n^* \xrightarrow{a.s.} I, \text{ as } n\to\infty \tag{27}$$

where I is a unit matrix;

(ii) $\sum_{i=1}^{n} < G^* > n^{-1/2}(G_i^* - G_{i-1}^*)(G_i^* - G_{i-1}^*)' < G^* > n^{-1/2} \xrightarrow{p} I$, as $n\to\infty$;

(iii) $\max_{0<i\le n}|(< G^* > n^{-1/2}(G_i^* - G_{i-1}^*))_j| \xrightarrow{p} 0$, as $n\to\infty, j=1,2,\ldots,k+2$;

(iv) $E\max_{0<i\le n}(|(< G^* > n^{-1/2}(G_i^* - G_{i-1}^*))_l(< G^* > n^{-1/2}(G_i^* - G_{i-1}^*))_j|)$
is bounded in n, for $\theta_0=\theta$ and $1\le l,j\le k+2$, where $(\cdot)_j$ denotes the j−th element of the vector $(\cdot)$.

Therefore, applying Theorem 1 in Chapter 4, we obtain

$$< G^*(\theta_0) >_n^{-1/2} G_n^*(\theta_0) \xrightarrow{p} N(0,I), \qquad as \quad n\to\infty.$$

Since

$$0 = G_n^*(\hat{\theta}_n) = G_n^*(\theta_0) + \dot{G}_n^*(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0),$$

where $\tilde{\theta}_n$ is a point on the line connecting θ_0 and $\hat{\theta}_n$, following Assumptions 1-4 and Theorem 2.7.5a in Révész(1968), we obtain

$$(\hat{\theta}_n - \theta_0) = -(1/n\dot{G}_n^*(\tilde{\theta}))^{-1}(1/n\dot{G}_n^*(\tilde{\theta}))(\hat{\theta}_n - \theta_0)$$

$$= -(1/n\dot{G}_n^*(\tilde{\theta}))^{-1}(1/nG_n^*(\theta_0))$$

$$= -(1/nC_nC_n^{-1}\dot{G}_n^*(\tilde{\theta}_n))^{-1}(1/nG_n^*(\theta_0)) \xrightarrow{a.s.} 0, \qquad as \quad n \to \infty.$$

This yields

$$< G^*(\theta_0) >_n^{-1/2} \dot{G}_n^*(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)$$

$$\sim < G^*(\theta_0) >_n^{-1/2} \dot{G}_n^*(\theta_0)(\hat{\theta}_n - \theta_0)$$

$$= < G^*(\theta_0) >_n^{-1/2} C_n(\theta_0)C_n(\theta_0)^{-1}\dot{G}_n^*(\theta_0)(\hat{\theta}_n - \theta_0)$$

$$\sim < G^*(\theta_0) >_n^{-1/2} C_n(\theta_0)(\hat{\theta}_n - \theta_0).$$

Furthermore, since

$$< G^*(\theta_0) >_n^{-1/2} \dot{G}_n^*(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I), \qquad \text{as } n \to \infty,$$

we obtain

$$< G^*(\theta_0) >_n^{-1/2} C_n(\theta_0)(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I), \qquad \text{as } n \to \infty,$$

i.e.

$$(A'F^{-1}A)^{1/2}(\hat{\theta}_n - \theta_\circ) \xrightarrow{d} N(0, I), \qquad \text{as } n \to \infty$$

CHAPTER 7
EXTENSIONS OF QUASI-LIKELIHOOD

1 Introduction

In this chapter, we deal with three different subjects, generalized information, estimating functions without the martingale property and estimating functions without zero mean.

As we have mentioned before, the quasi-likelihood method developed from the ML and LS methods. The quasi-score estimating function can provide smaller asymptotic confidence zones for an unknown parameter than other estimating functions in the same estimating function space. This is an important property and we might say that the optimality of the quasi-likelihood method just comes from this property. Following this view-point, a more general definition of the quasi-score estimating function will be given in this chapter.

Another topic which will be discussed here is extension of quasi-likelihood from the case where the estimating functions are based on families of martingales to the case where no particular structure is assumed. Godambe and Heyde(1987) and Heyde(1988) provided a formula to determine the quasi-score estimating function from an estimating function space $\mathcal{G}$ in which each element is a martingale with respect to some σ-fields. Godambe and Thompson(1989) gave another formula for a more general: mutually orthogonal case (for which the martingale case is a particular example). However, an estimating function space can be of diverse character. We need more formulas to determine the quasi-score estimating function for different kinds of estimating function spaces. Following Godambe and Heyde's idea, another formula will be presented here to determine the quasi-score estimating function from an estimating function space $\mathcal{G}$ in which the elements may not even be generated by a sequence of mutually orthogonal estimating functions .

The final subject is the development of the quasi-likelihood method in a more general estimating function space. Since the estimating function G is initially used to obtain an estimator by solving the equation $G = 0$, the unbiasedness condition

$$E_\theta(G) = 0, \qquad \forall\, \theta \in \Theta,$$

becomes a natural one(see Godambe and Kale(1991)). However, it does not mean that the estimating function G should necessarily satisfy the unbiasedness condition. In the quasi-likelihood framwork, the quasi-score estimating function G is usually a martingale(see Godambe and Heyde(1987)). However, in practice, G may not be a martingale, for example G may be a semimartingale. An example will be provided to explain this fact and a framework will be given for this situation.

2 Optimal Estimating Functions

Modeling by a stochastic process and estimating the parameters of the process are common problems. If the likelihood function of the process is known, the ML method is generally used. The LS method(including the weighted least-squares method, the conditional least-squares method and so on) is another method which is widely used especially for prediction problems. In a sense, these two methods come from some basic idea such as the likelihood principle or the best prediction idea of minimizing the residual sum of squares. But, from the view-point of estimating function theory, these two methods are just methods for choosing an estimating function from some estimating function space under some specified rules. For the ML method, the rule is the maximum likelihood principle; for the LS method, there are two rules: one is the minimum residual sum of squares and the other is the form of the estimating function. We also can say that the ML method is a method for choosing an estimating function from the whole estimating function space under the likelihood principle and the LS method is a method for choosing an estimating function from a subset of estimating function space which has minimum residual sum of squares.

Why can the ML method choose an estimating function from the whole estimating function space but the LS method only choose an estimating function from a subset of an estimating function space? Because, when using the ML method , we should know the likelihood function of the process, which means that we obtain all the information about the process. If we do not know the likelihood function of the process, we can only choose an estimating function from some subset of an estimating function space. The subset of the estimating function space depends on the information which we obtain from the process and experiments. For example, if we consider the LS method for fitting a model $y_t = f(x_s, s < t) + \varepsilon_t$ and we know nothing about the moments and mixed moments of $\{x_s\}$ for order $k > 2$, then only the set of linear estimating functions can be considered.

Briefly, using an estimation method is equivalent to finding an estimating function from some subset of an estimating function space under some rules. So, the estimation method is determined by the subset of the estimating function space and the rules which we consider.

Also, we can say that the likelihood score function and the least squares estimating function are, respectively, " optimal" estimating functions with reference to some appropriate subset of estimating functions. Following these ideas, we here give a general definition for the quasi-score estimating function.

Suppose that Ψ_T is a subset of estimating function space in which T is the sample size, Θ is a parameter space which is a subset of R^p and A is a " matrix functional" on $\Psi_T \otimes \Theta$ such that $A : \Psi_T \otimes \Theta \ni G \otimes \theta \longmapsto R^{p\times p}$ and $(A(G_T, \theta)'A(G_T, \theta))^{-1}$ are nonsingular for $G_T \in \Psi_T$, $\theta \in \Theta$. We call (Ψ_T, Θ, A) an estimating function space.

Definition: We say that G_T^* is a quasi-score estimating function in (Ψ, Θ, A), if, for all $G_T \in \Psi_T$ and $\theta \in \Theta$

$$A'(G_T^*, \theta)A(G_T^*, \theta) \geq A'(G_T, \theta)A(G_T, \theta).$$

We denote θ_T^* as a quasi-likelihood estimator which satisfies

$$G_T^*(\theta_T^*) = 0.$$

Example 1: Suppose that $X = \{X_t\}$ is a process and, for every T fixed, $\{X_t\}_{t\leq T}$ belongs to an exponential family. Let $\Psi_T = \{G_T = G_T(X_t, 0 \leq t \leq T, \theta)\}$, in which $EG_T(\theta) = 0$ for each $P_\theta \in \mathcal{P}$ with index θ , $\{G_T, \mathcal{B}_T\}$ are martingales for each

$P_\theta \in \mathcal{P}$, G_T are almost surely differentiable with respect to the components of θ and $E\dot{G}_T(\theta) = (E(\partial G_{T,i}(\theta)/\partial\theta_i))$ and $E(G_T(\theta)G_T'(\theta))$ are nonsingular. Let A be a matrix functional defined by

$$A : \Psi_T \otimes \Theta \ni G_T \otimes \theta \longmapsto (E_\theta G_T G_T')^{-1/2} E_\theta \dot{G}_T.$$

Then there is a quasi-score estimating function G_T^* in (Ψ, Θ, A). In fact, G_T^* is the likelihood score function.

Example 2: Let us consider an AR model. Assume that $\sum_1^T \varepsilon_T = \sum_1^T (y_t - (\beta_1 x_{t-1} + \ldots + \beta_k x_{t-k}))$ is a martingale in which $E(\varepsilon_t|\mathcal{F}_{t-1}) = 0$ and $E(\varepsilon_t^2|\mathcal{F}_{t-1}) = c$, c being a constant and $\mathcal{F}_t = \sigma(x_1, \ldots, x_{t-1})$. Let $\Psi_T = \{G_T = \sum_1^T f_t(\theta)\varepsilon_t| \ \{f_t(\theta)\}$ is a predictable process with respect to $\{\mathcal{F}_t\}\}$ and A be a matrix functional defined by

$$A : \Psi_T \otimes \Theta \ni G_T \otimes \theta \longmapsto (E_\theta G_T G_T')^{-1/2} E_\theta \dot{G}_T.$$

Then there is a quasi-score estimating function G_T^* in (Ψ_T, Θ, A) and

$$G_T^* = \frac{1}{c} \sum_1^T \begin{pmatrix} -x_{t-1} \\ \vdots \\ -x_{t-k} \end{pmatrix} (y_t - (\beta_1 x_{t-1} + \ldots + \beta_k x_{t-k})).$$

The only difference between G_T^* and the least-squares estimating function is the constant coefficient.

Example 3: (Godambe and Heyde(1987)) Let $\mathcal{G}_T$ be a subspace of Ψ in which every element G_T is almost surely differentiable with respect to the components of θ and $E\dot{G}_T(\theta)$ and $E(G_T(\theta)G_T'(\theta))$ are nonsingular. Let μ_1 denote the subset of $\mathcal{G}_T$ consisting of square integrable martingales with respect to $\{\mathcal{F}_t\}$. Let A be a matrix defined by

$$A : \mu_1 \times \Theta \ni G_T \otimes \theta \mapsto < G(\theta) >_T^{-1/2} \bar{G}_T(\theta),$$

where $\bar{G}_T(\theta) = \int_0^T E(d\dot{G}_s(\theta)|\mathcal{F}_{s-})$. Then the quasi-score estimating function in (μ_1, Θ, A) is same as the quasi-score estimating function under Criterion 4 in Godambe and Heyde(1987).

Therefore, following the extending definition of the quasi-score estimating function the quasi-likelihood method is a method to choose an estimating function from Ψ_T under the rules:

(1) A is a given matrix functional,

(2) $A'(G_T^*, \theta)A(G_T^*, \theta) \geq A'(G_T, \theta)A(G_T, \theta)$ for all $G_T \in \Psi$, $\theta \in \Theta$.

We can also say that G_T^* is an optimal estimating function in Ψ. But "optimal" has its full meaning only when the following facts are true:

(1) $\theta_T^* \xrightarrow{a.s.} \theta_\circ$, where $\theta_\circ$ is the true parameter,

(2) $\|A(G_T^*, \theta_\circ)(\theta_T^* - \theta_\circ)\|^2 \xrightarrow{d} \mathcal{L}(X)$, for some proper r.v. X.

Following the discussion in Chapter 5, an $(1-\alpha) \times 100\%$ percent asymptotic confidence zone for $\theta_\circ$ will be determined by

$$P((\theta_T^* - \theta_\circ)((\theta_T^* - \theta_\circ)' \leq \epsilon_\alpha(A'(G_T^*, \theta_\circ)A(G_T^*, \theta_\circ))^{-1}),$$

where $P(X \leq \epsilon_\alpha) = 1 - \alpha$. If there is another estimating function $G_T \in \Psi$ such that,

$$\theta_T \xrightarrow{a.s.} \theta,$$

where θ_T satisfies $G_T(\theta_T) = 0$ and

$$\|A(G_T, \theta_\circ)(\theta_T - \theta_\circ)\|^2 \xrightarrow{d} \mathcal{L}(X),$$

then $A(G_T^*, \theta_\circ)$ will provide smaller asymptotic confidence zones for $\theta_\circ$ than those provided by $A(G_T, \theta_\circ)$. In this sense the $A'(G_T, \theta)A(G_T, \theta)$ is playing the role of Fisher Information .

Here we have discussed three kinds of optimal estimating functions. There are also others. However it is difficult to find an universal optimal estimating function except the likelihood score function(in practice the likelihood score function may work as a universal optimal estimating function if it can be written down). In general, the optimal estimating function property is only a local property or one relative to $(\Psi, \Theta, (rules))$. The Ψ is determined by the kind of model we want to fit to the data and the kind of information we can obtain from the history. The rule is determined by the kinds of criteria we choose for optimality and the kinds of tools we use.

3 The Form of a Quasi-score Estimating Function and an Equivalence Criterion

How is a quasi-likelihood method applied in practice? The key task is to find the quasi-score estimating function from a given estimating function space. However the definition of the quasi-score estimating function given by Godambe and Heyde(1987) is too abstract to suggest a general form of the quasi-score estimating function .

This question was considered by Godambe and Heyde(1987) who presented a formula for a quasi-score estimating function from an estimating function space of the form $\mathcal{G}_T = \{\int_0^T \alpha_s dM_s | \{\alpha_s\}$ is a predictable process with respect to $\{\mathcal{F}_t\}\}$, where $\{M_t\}$ being a fixed martingale with respect to some σ-fields $\{\mathcal{F}_t\}$. They proved that in the above case , if

$$G_T^* = \int_0^T E(dM_s|\mathcal{F}_{s-})(d < M >_s)^+ dM_s \tag{1}$$

belongs to $\mathcal{G}_T$, then G_T^* is a quasi-score estimating function in $\mathcal{G}_T$. This formula is very convenient to use in practice.

Godambe and Thompson(1989) extended (1) and gave another formula for " the mutually orthogonal" case. Let $\Omega = \{x\}$ be an abstract sample space and $\mathcal{P} = \{p\}$ be a class of distribution functions on Ω. Let $\theta = (\theta_1, \ldots, \theta_m)$ be a vector parameter with real components defined on $\mathcal{P}$ such that $\{\theta\} \equiv \Theta$. Let further f_j, $j = 1, \ldots, n$, with arbitrary n, be real functions on $\Omega \times \Theta$ such that

$$E_p\{f_j(x, \theta(p))|\mathcal{F}_j\} = 0, \qquad p \in \mathcal{P}, \tag{2}$$

where $\mathcal{F}_j$ is a specified partition (or technically a σ-field generated by a partition) of $\mathcal{X}$, $j = 1, 2, \ldots, k$. For simplicity we use the notation

$$E_p\{\cdot|\mathcal{F}_j\} = E_j\{\cdot\}.$$

Consider an estimating function space $\mathcal{G}_n = \{\sum_{j=1}^n \alpha_j f_j | \ \alpha_j \in \mathcal{F}_j\}$ and let $\tilde{G}_n^* = \sum_{j=1}^n \tilde{\alpha}_j^* f_j \in \mathcal{G}_n$, where $\tilde{\alpha}_j^* = E_j(\dot{f}_j)/E_j(f_j^2)$.

Definition : The estimating functions f_j, $j = 1, \ldots, k$, satisfying (2) are said to be mutually orthogonal, if

$$E_j(\tilde{\alpha}_j^{*(r)} f_j f_i \tilde{\alpha}_i^{*(r')}) = 0,$$

$i \neq j$, $i, j = 1, \ldots, n$ and $r, r' = 1, \ldots, m$, where $\tilde{\alpha}_j^{*(r)}$ is the r-th component of $\tilde{\alpha}_j^*$.

Godambe and Thompson(1989) gave the following result.

Theorem 1 *The estimating function $\tilde{G}_n^* = \sum_{j=1}^n \tilde{\alpha}_j^* f_j$ is a quasi-score estimating function in $\mathcal{G}_n$ if $\{f_i\}$ are mutually orthogonal.*

We can see that the martingale case belongs to "the mutually orthogonal" case.

Now our question is that if $\mathcal{G}_T$ does not consist of estimating functions $\{\sum \alpha_i f_i\}$, where α_i are $\mathcal{F}_i$ measurable and $\{f_i\}$ are orthogonal estimating functions with respect to $\{\mathcal{F}_i\}$, how we can determine the quasi-score estimating function G_T^* from the $\mathcal{G}_T$. In this section, we shall discuss this question when the restrictions on $\mathcal{G}_T$ are somewhat reduced. As usual the regularity conditions are always assumed true.

Before discussing this question, we need a new definition. Assume that $\mathcal{G}_n$ is an estimating function space.

Definition 1 *G_n^* is said to be a pseudo-quasi-score estimating function in $\mathcal{H}_n$ if*

(a) $E\dot{G}_n^*$ is nonsingular;

(b) $E(G_n G_n') \geq (E\dot{G}_n)(E\dot{G}_n^*)^{-1}(EG_n^* G_n^{*\prime})(E\dot{G}_n^{*\prime})^{-1}(E\dot{G}_n)'$

for any $G_n \in \mathcal{H}_n$.

Remarks: 1. The choice of the particular form (b) is motivated by the fact that the element G_n of $\mathcal{H}_n$ may not satisfy the conditions that

$$EG_n G_n' \qquad \text{and} \qquad E\dot{G}_n \tag{3}$$

are nonsingular.

2. If $\mathcal{G}_n \subseteq \mathcal{H}_n$ and every element G_n of $\mathcal{G}_n$ satisfies the conditions (3), the pseudo-quasi-score estimating function G_n^* in $\mathcal{H}$ is the quasi-score estimating function in $\mathcal{G}_n$ if $G_n^* \in \mathcal{G}_n$.

Let $\{f_n\} \in R^k$ be a sequence of random vectors having the property that there exists a sequence of σ-fields $\mathcal{F}_n$ which is generated from some subsets of the sample space and is such that

$$E_\theta(f_n(\theta)|\mathcal{F}_n) = 0, \tag{4}$$

where $\theta \in \Theta \subseteq R^d$, Θ being a parameter space. Let $\mathcal{G}_n$ be a set of $\mathcal{H}_n = \{G_n = \sum_{i=1}^n \alpha_i f_i\}$, where $\{\alpha_i\}$ are $\{\mathcal{F}_i\}$ measurable matrices.

Remarks: 3. The σ-fields $\{\mathcal{F}_n\}$ may have no order, i.e., $n \leq m$ does not mean that $\mathcal{F}_n \subseteq \mathcal{F}_m$.

4. If $\{f_j\}$ is a martingale difference sequence and $\{\mathcal{F}_j\}$ are natural σ -fields, then the $\{f_j\}$ satisfy (4). But the converse is not always true. Because a martingale should be an adapted process but the process $\{\sum_{i=1}^n f_i\}$ is not necessarily adapted.

Following the proof of Theorem 1 in Heyde(1988), we have the result below.

Theorem 2 *Suppose that $G_n^* = \sum_{j=1}^n \alpha_j^* f_j \in \mathcal{H}_n$ and that both $E(G_n^* G_n^{*\prime})$ and $E\dot{G}_n^*$ are nonsingular. Then, the following statements are equivalent:*

(1) G_n^* is a pseudo-quasi-likelihood estimating function in $\mathcal{H}_n$;

(2) $EG_nG_n^{*\prime} = E\dot{G}_nA$ for all $G_n \in \mathcal{H}_n$, for some nonsingular nonrandom matrix A;

(3) $E\sum_{i=1}^n \alpha_i \sum_{j\neq i}^n f_i f_j' \alpha_j^{*\prime} = 0$, if $\alpha_j^* = AE_j\dot{f}_j(E_jf_jf_j')^{-1}$, for some nonsingular nonrandom matrix A.

The above theorem shows us how to construct a quasi-score estimating function from $\mathcal{G}_n \in \mathcal{H}$. As an application we give the following theorem.

Theorem 3 *If $\{\alpha_j^*\}$ is such that $G_n^* = \sum_{j=1}^n a_{j-1}\alpha_i^* f_j \in \mathcal{G}_n$, where $\{a_{j-1}$ are constants, and*

$$E_i \sum_{j=1}^n a_{j-1} f_i f_j' \alpha_j^{*\prime} = E_i \dot{f}_i A$$

for some nonrandom matrix A, then G_n^* is a quasi-score estimating function in $\mathcal{G}_n$.

If we denote

$$b_{i,j}(\theta) = E_\theta(f_i(\theta)f_j'(\theta)|\mathcal{F}_i \vee \mathcal{F}_j)$$
$$i, j = 1, 2, \ldots,$$

where $\mathcal{F}_i \vee \mathcal{F}_j = \sigma\{\mathcal{F}_i, \mathcal{F}_j\}$ and

$$c_i = c_i(\theta) = E_i(\dot{f}_i)$$
$$i = 1, 2, \ldots,$$

then applying the above theorem, we obtain a formula for quasi-score estimating function in the nonorthogonal case.

Corollary 1 *Suppose that $\{f_i\}$ satisfy (4). Let $\mathcal{G}_n = \{\sum_{i=1}^n \alpha_i f_i | \alpha_i \in \mathcal{F}_i\}$ be an estimating function space. If $\alpha^* = (\alpha_1^*, \ldots, \alpha_n^*)'$ satisfies*

$$\alpha^* = \begin{pmatrix} b_{1,1} & b_{12} & \ldots & b_{1,n} \\ \vdots & \vdots & \vdots & \vdots \\ b_{n,1} & b_{n,2} & \ldots & b_{n,n} \end{pmatrix}^+ \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix},$$

where for any matrix B, B^+ denotes the Moore-Penrose pseudo inverse of B, then $G_n^* = \sum_{i=1}^n \alpha_i^* f_i$ is a quasi-score estimating function in $\mathcal{G}_n$ if $G_n^* \in \mathcal{G}_n$.

Proof: From the condition of theorem, we have

$$\begin{pmatrix} b_{1,1} & b_{12} & \ldots & b_{1,n} \\ \vdots & \vdots & \vdots & \vdots \\ b_{n,1} & b_{n,2} & \ldots & b_{n,n} \end{pmatrix} \begin{pmatrix} \alpha_1^{*\prime} \\ \vdots \\ \alpha_n^{*\prime} \end{pmatrix} = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix}.$$

Therefore,

$$\sum_{j=1}^n b_{i,j}\alpha_j^{*\prime} = c_i, \qquad i = 1, \ldots, n,$$

i.e.

$$E_i(\sum_{j=1}^n f_i f_j' \alpha_j^{*\prime}) = \sum_{j=1}^n E_i(b_{i,j}\alpha_j^{*\prime}) = E_i(c_i) = c_i,$$
$$i = 1, \ldots, n.$$

Applying Theorem 3, $G_n^* = \sum_{i=1}^n \alpha_i^* f_i$ is a quasi-score estimating function in $\mathcal{G}_n$.

Corollary 2 *If $b_{i,j} = 0$ for all $i \neq j$, and*

$$G_n^* = \sum_{i=1}^{n} a_{i-1}(E_\theta(\dot{f}_i|\mathcal{F}_{r(i)}))'(a_{i-1}E_\theta(f_i f_i'|\mathcal{F}_{r(i)}))^+ f_i$$

belongs to $\mathcal{G}_n$, then G_n^* is a quasi-score estimating function in $\mathcal{G}_n$.

Remarks: 1. If $\{f_j\}$ is a martingale difference sequence and $\{\mathcal{F}_j\}$ are natural σ-fields generated from $\{f_j\}$, the $b_{i,j}$ are always 0, $i \neq j$.

2. If $b_{i,j}$, $i \neq j$, are not always equal to 0, it is usually very difficult to check whether $\alpha_j^* \in \mathcal{F}_j$, $j = 1, \ldots, n$.

Therefore we are still interested the conditions which can force

$$\alpha_j^* = AE_j(\dot{f}_j)(E_j(f_j f_j'))^{-1}.$$

From Theorem 2, it is easy to get the following relations.

Corollary 3 *Suppose that $\alpha_j^* = AE_j\dot{f}_j(E_j f_j f_j')^{-1}$ for some nonsingular nonrandom matrix A. Then*

(1) $E_i(\sum_{j\neq i}^n f_i f_j' \alpha_j^{*\prime}) = 0$, $i = 1, \ldots, n$, implies that G_n^* is a pseudo-quasi-score estimating function in $\mathcal{G}_n$;

(2) G_n^* is a pseudo-quasi-score estimating function in $\mathcal{G}_n$ implies that

$$E(\sum_{j\neq i}^{n} f_i f_j' \alpha_j^{*\prime}) = 0;$$

(3) $E_i(f_i f_j' \alpha_j^{*\prime}) = 0$, $i \neq j$, $i, j = 1, \ldots, n$, implies that G_k^* is a pseudo-quasi-score estimating function in $\mathcal{G}_k$, $k = 1, \ldots, n$;

(4) G_k^* is a pseudo-quasi-score estimating function in $\mathcal{G}_k$, $k = 1, \ldots, n$, implies that $Ef_i f_j' \alpha_j^{*\prime} = 0$, $i \neq j$, $i, j = 1, \ldots, n$.

Remarks: 1. Under the assumption that $E\dot{G}_n^*$ is nonsingular, the conditions $E_i(f_i f_j' \alpha_j^{*\prime}) = 0$, $i \neq j$, $i, j = 1, \ldots, n$ are equivalent to the " mutually orthogonal" condition introduced by Godambe and Thompson(1989).

2. Following the discussion in Corollary 3, we may say that in a certain sense the " mutually orthogonal" condition is a rather weak condition to force $\alpha_j^* = AE_j(\dot{f}_j)(E_j(f_j f_j'))^{-1}$.

Let $\mathcal{G}_n$ be a set of $\{G_n\}$, where G_n is a Borel mapping from $(R^m)^n \times \Theta \to R^k$ such that, for any $\theta \in \Theta$,

$$E_\theta G_n(y_1, \ldots, y_n, \theta) = \int_{(R^m)^n} G_n(y_1, \ldots, y_n, \theta) dp_{n,\theta} = 0.$$

The following theorem gives us a criterion for the quasi-score estimating function which is equivalent to the criteria in Godambe and Heyde(1987).

<u>**Theorem**</u> **4** *If there exists a quasi-score estimating function within $\mathcal{G}_n$, G_n^* is a quasi-score estimating function iff, for all $\theta \in \Theta$ and $G_n \in \mathcal{G}_n$,*

$$tr((E_\theta \dot{G}_n^*)^{-1}(E_\theta(G_n^*)(G_n^*)')((E_\theta \dot{G}_n^*)')^{-1}) \leq tr((E_\theta \dot{G}_n)^{-1}(E_\theta(G_n)(G_n)')((E_\theta \dot{G}_n)')^{-1}).$$

Proof: " $\Rightarrow$ " is trivial.

Now we prove the " $\Leftarrow$ " part. Assume that there is a $\tilde{G}_n$ in $\mathcal{G}_n$ such that

$$tr((E_\theta \dot{\tilde{G}}_n)^{-1}(E_\theta(\tilde{G}_n)(\tilde{G}_n)')((E_\theta \dot{\tilde{G}}_n)')^{-1}) \leq tr((E_\theta \dot{G}_n)^{-1}(E_\theta(G_n)(G_n)')((E_\theta \dot{G}_n)')^{-1})$$

for all $\theta \in \Theta$ and $G_n \in \mathcal{G}_n$. Following the definition of the quasi-score estimating function we should have

$$tr((E_\theta \dot{\tilde{G}}_n)^{-1}(E_\theta(\tilde{G}_n)(\tilde{G}_n)')((E_\theta \dot{\tilde{G}}_n)')^{-1}) = tr((E_\theta \dot{G}_n^*)^{-1}(E_\theta(G_n^*)(G_n^*)')((E_\theta \dot{G}_n^*)')^{-1}).$$

By the definition of the quasi-score estimating function we also have

$$(E_\theta \dot{\tilde{G}}_n)^{-1}(E_\theta(\tilde{G}_n)(\tilde{G}_n)')((E_\theta \dot{\tilde{G}}_n)')^{-1} - (E_\theta \dot{G}_n^*)^{-1}(E_\theta(G_n^*)(G_n^*)')((E_\theta \dot{G}_n^*)')^{-1} \geq 0.$$

Combining these two statements, we obtain that the maximum eigenvalue of

$$(E_\theta \dot{\tilde{G}}_n)^{-1}(E_\theta(\tilde{G}_n)(\tilde{G}_n)')((E_\theta \dot{\tilde{G}}_n)')^{-1} - (E_\theta \dot{G}_n^*)^{-1}(E_\theta(G_n^*)(G_n^*)')((E_\theta \dot{G}_n^*)')^{-1}$$

is equal to 0; that is

$$(E_\theta \dot{\tilde{G}}_n)^{-1}(E_\theta(\tilde{G}_n)(\tilde{G}_n)')((E_\theta \dot{\tilde{G}}_n)')^{-1} = (E_\theta \dot{G}_n^*)^{-1}(E_\theta(G_n^*)(G_n^*)')((E_\theta \dot{G}_n^*)')^{-1}.$$

Therefore $\tilde{G}_n$ is also a quasi-score estimating function in $\mathcal{G}_n$.

4 Extending the Quasi-likelihood Method to a Model with a Semimartingale Error Process

Firstly, let us consider an example. Suppose that there are n identical pieces of equipment in a factory. The average of the measurement errors of the this kind of equipment is zero. However each piece of equipment may have a measurement bias. Now we do a series of experiments to fit a model

$$X = f(\theta) + M$$

where M is a r.v. called the error part. The form of the function f is known and θ is a parameter which is going to be estimated. When we do the experiment, the pieces of equipment need to be used. Assume that each piece of equipment is randomly chosen for use and is only used once. From each piece experiment we can get an observation on X. If we take n samples from X, then we obtain

$$X_i = f_i(\theta) + M_i, \qquad i = 1, 2, \ldots, n.$$

For the ideal model, we would assume that $E(M_i|\mathcal{F}_{i-1}) = 0$, where $\mathcal{F}_i = \sigma\{X_j, j \leq i\}$. However, when we use the above system with measurement bias to observe X_i, $i = 1, 2, \ldots, n$, we can only obtain

$$\begin{aligned} \tilde{X}_i &= f_i(\theta) + M_i + N_i \\ &= f_(\theta) + A_i, \end{aligned}$$

where N_i is a r.v. which comes from the bias. Because of the finite amount of equipment, the finite usability of the equipment and bias of each piece of equipment, it is obvious that $E(N_i|\mathcal{F}_{i-1}) \neq 0$ and $E(N_i|\mathcal{F}_{i-1})$ is possibly a r.v. for each fixed i. Thus

$$\sum_{i=1}^{n} \tilde{X}_i = \sum_{i=1}^{n} f_i(\theta) + \sum_{i=1}^{n} A_i$$

gives us a model having a semimartingale error process $\sum_{i=1}^{n} A_i$. This example tells us, in practice, that there exists a model:

$$X_t = \int_0^t f_s(\theta) d\lambda_s + M(t), \tag{5}$$

where $\{M_t\}$ is a semimartingale with respect to some σ-fields $\{\mathcal{F}_t\}$ and $\{f_t(\theta)\}$ is a predictable process with respect to $\{\mathcal{F}_t\}$. For each t fixed, $f_t(\theta)$ is a function of θ, where θ is a parameter to be estimated.

In order to estimate θ, we focus our attention on two questions: (i) what kind of estimating function space do we choose? (ii) how do we choose an optimal estimating function from that estimating function space?

It is well known that, for a distribution free inference problem, to choose a suitable estimating function space $\mathcal{G}$ is quite important. If $\mathcal{G}$ is too large, it seems difficult to find an optimal estimating function from $\mathcal{G}$; if $\mathcal{G}$ is too small, the optimal estimating function chosen from $\mathcal{G}$ could be seriously lacking in efficiency.

Due to the $\{M_t\}$ in (5) being a semimartingale, we prefer to focus our attention on an estimating function space $\mathcal{G}_T$:

$$\mathcal{G}_T = \{G_T | \dot{G}_T(\theta) = \bar{G}_T(\theta) + o_p(\bar{G}_T(\theta)),$$

$$\bar{G}_T(\theta) = \int_0^T E(d\bar{G}_s(\theta)|\mathcal{F}_{s-})\} \tag{6}$$

where $\dot{G}_T(\theta) = \partial(G_T(\theta))/\partial\theta$. Godambe and Heyde(1987) have discussed this kind of estimating function space. They called this space a class μ_4. If $\{M_t\}$ is a martingale, a criterion for choosing optimal estimating function was given in that paper. However there is a difference between (6) and μ_4. Unlike what is supposed for μ_4, we cannot assume that $E_\theta G_T = 0$, because, in general case, G_T is a semimartingale — a stochastic integral with respect to $\{M_t\}$.

Avoiding a number of complex questions involved in the model (5), we here only consider a simple case

$$X_T = \int_0^T f_s(\theta)d\lambda_s + M_T \tag{7}$$

where $\{M_t\}$ is a semimartingale with $\lambda_s =< M >_s$ and $f_t(\theta)$ is a predictable process which is a linear function of θ, for each fixed s.

Before discussing (7), we list some basic definitions and properties of semimartingales.

We denote by $\mathcal{V}$ the space of bounded variation processes and denote by $\mathcal{M}_{loc}$ the space of local martingales. let M be a semimartingale admitting the representation

$$M_t = M_0 + A_t + \tilde{M}_t \tag{8}$$

where $A = (A_t)_{t \geq 0} \in \mathcal{V}$ with $A_0 = 0$ and $\tilde{M} = (\tilde{M}_t)_{t \geq 0} \in \mathcal{M}_{loc}$ with $\tilde{M}_0 = 0$. Suppose that a predictable process H possesses the following properties:

$$|H| \circ Var(A) \in \mathcal{V}^+, \tag{9}$$

$$(H^2 \circ [\tilde{M}, \tilde{M}])^{1/2} \in \mathcal{A}^+_{loc} \tag{10}$$

where $\mathcal{A}^+_{loc}$ is the set of local integrable variation and $\mathcal{V}^+$ is the set of positive bounded variation processes. We shall mention the definition of the above stochastic integral because it will be used here. Details can be found in any stochastic analysis textbook like Liptser and Shiryaev(1986).

<u>**Definition**</u> **2** *The process*

$$H.M = H \circ A + H.\tilde{M}$$

is called the stochastic integral of H with respect to a semimartingale M.

Sometimes $H.M_t$ will also be denoted as

$$\int_0^t H_s dM_s \qquad or \qquad \int_{(0,t]} H_s dM_s.$$

<u>**Theorem**</u> **5** *(Liptser and Shiryaev(1986)) Let M be a semimartingale with decomposition (8) and let H be a predictable process with the properties (9) and (10). Then $H.M$ is a semimartingale.*

An important characteristic of a semimartingale M is represented by its quadratic variation $[M, M]$ with $(\triangle M_0 = 0)$

$$[M, M]_t =< M^c >_t + \sum_{0 \leq s \leq t} (\triangle M_s)^2 \tag{11}$$

where M^c is the continuous martingale component of M. Following the definition, $[M, M]$ is an adapted increasing process. If $[M, M]$ is a local integrable variation process, following the theory of dual process, there is a dual predictable projection process of $[M,M]$ denoted by $< M,M >$ such that $[M,M]- < M,M >$ is a uniformly local integral martingale(see, Dellacherie and Meyer(1975), Meyer(1976, 1977) and Yan (1987)).

Let $\mathcal{G}_T$ be an estimating function space and

$$\mathcal{G}_T = \{G_T | G_T = \int_0^T \alpha_s dM_s, \qquad \alpha_s \text{ satisfies (9) and (10)},$$
$$\dot{G}_T(\theta) = \bar{G}_T(\theta) = o_p(\bar{G}_T(\theta)),$$
$$\bar{G}_T(\theta) = \int_0^T E(d\dot{G}_s(\theta)|\mathcal{F}_{s-})\}.$$

Let μ' be a subspace of $\mathcal{G}_T$. Each element G_T of μ' satisfies the condition

$$\bar{G}_T(\theta) = \int_0^T \alpha_s E(d\dot{M}_s|\mathcal{F}_{s-}) + o_p(\int_0^T \alpha_s E(d\dot{M}_s|\mathcal{F}_{s-})).$$

For a special case, if α_s is not a function of θ, then

$$\bar{G}_T = \int_0^T \alpha_s E(d\dot{M}_s|\mathcal{F}_{s-}).$$

Let μ'' be a subset of μ' such that each element G_T in μ'' satisfies the condition that $[G.,G.]$ is a local integral variation process.

Theorem 6 *Suppose that μ is a measure generated by an increasing process A on $\mathcal{B}(R^+) \times \mathcal{F}$ and μ^p is the predictable projection of μ. Then μ^p is generated by a predictable increasing process iff A is local integral, i.e. there is sequence of stopping time $\{T\}$ such that A_T is integral.*

Therefore, following the definition of dual predictable projection process, $< G,G >$ is a predictable increasing process.

Theorem 7 *Suppose that X and Y are two semimartingales and H and K are two measurable processes, while p and q satisfy $1/p + 1/q = 1$. Then*

$$|\int_{(0,\infty[} |H_s K_s||d[X,Y]_s|$$

$$\leq \left(\int_{[0,\infty[} H_s^2 d[X,X]_s\right)^{1/2} \left(\int_{[0,\infty[} K_s^2 d[Y,Y]_s\right)^{1/2}, \qquad a.s., \tag{12}$$

$$E\left[\int_{(0,\infty[} |H_s K_s| |d[X,Y]_s|\right]$$

$$\leq \left\|\sqrt{\int_{[0,\infty[} H_s^2 d[X,X]_s}\right\|_p \left\|\sqrt{\int_{[0,\infty[} K_s^2 d[Y,Y]_s}\right\|_q, \qquad a.s., \tag{13}$$

The above theorems can be obtained from Dellacherie and Meyer(1975), Meyer(1976, 1977) and Yan (1987).

Sometimes we call the above inequalities the K-W (Kunita-Watanabe)inequalities.

Remark : If $< X, X >$, $< X, Y >$ and $< Y, Y >$ exist, there are corresponding K-W inequalities for $< X, X >$, $< X, Y >$ and $< Y, Y >$.

Due to the above properties of the quadratic characteristic, a criterion for choosing an optimal estimating function $\dot{G}_T$ from μ'' can be given.

<u>**Criterion :**</u> We call G_T^* the quasi-score estimating function in μ', if

$$\tilde{G}_T^*(\theta) < G^* >_T^{-1} \tilde{G}_T^*(\theta)' - \tilde{G}_T(\theta) < G >_T^{-1} \tilde{G}_T(\theta)'$$

is nonnegative definite for all $G_T \in \mu'$, $\theta \in \Theta$ and $p_\theta \in \mathcal{P}$, where $\tilde{G}_T$ has the form

$$\tilde{G}_T = \int_0^T \alpha_s E(d\dot{M}_s|\mathcal{F}_{s-}).$$

Consider the model (4). If $[G_\cdot^*, G_\cdot^*]$ is a local integral variation process, then under the above Criterion, the optimal estimating function in μ'' is:

$$G_T^* = \int_0^T E(d\dot{M}_s|\mathcal{F}_{s-})(d < M >_s)^+ dM_s, \tag{14}$$

where $(d < M >_s)^+$ denotes the Moore-Penrose pseudo inverse.

Because, for any $G \in \mu''$, we have $< G, G^* >_T^2 \; \leq \; < G, G >_T < G^*, G^* >_T$ and

$$\begin{pmatrix} < G, G >_T & < G, G^* >_T \\ < G, G^* >_T & < G^*, G^* >_T \end{pmatrix}$$

is nonnegative-definite, then the method of Rao(1965, p266) gives the nonnegative-definiteness of

$$< G >_T - < G, G^* >_T < G^* >_T^{-1} < G, G^* >_T$$

i.e.

$$< G, G^* >_T^{-1} < G >_T < G, G^* >_T^{-1} - < G^* >_T^{-1} \geq 0. \tag{15}$$

Since (14) leads to

$$\tilde{G}^*_T = \int_0^T E(d\dot{M}_s|\mathcal{F}_{s-})(d < M >_s)^+ E(d\dot{M}_s|\mathcal{F}_{s-})$$

$$= < G^*, G^* >_T$$

and

$$< G, G^* >_T = < \alpha.M, E(d\dot{M}|\mathcal{F}_{.-})(d < M >_.)^+ M >$$

$$= \int_0^T \alpha_s E(d\dot{M}_s|\mathcal{F}_{s-})(d < M >_s)^+ d < M >_s$$

$$= \tilde{G}_T,$$

we obtain

$$(\tilde{G}_T)^{-1} < G >_T (\tilde{G}_T)^{-1} - (\tilde{G}^*_T)^{-1} < G^* >_T (\tilde{G}^*_T)^{-1} \geq 0,$$

i.e. G^*_T is the quasi-score estimating function in μ'' under the criterion.

Remark : If M is a martingale, then under the model (6),

$$\bar{G}_T(\theta) = \int_0^T E(d\dot{G}_s(\theta)|\mathcal{F}_{s-})$$

$$= \int_0^T E(\frac{\partial}{\partial\theta}(\alpha_s dM_s(\theta))|\mathcal{F}_{s-})$$

$$= \int_0^T E(\dot{\alpha}_s dM_s(\theta)|\mathcal{F}_{s-}) + \int_0^T E(\alpha_s d\dot{M}_s(\theta)|\mathcal{F}_{s-})$$

$$= \int_0^T E(\alpha_s d\dot{M}_s(\theta)|\mathcal{F}_{s-})$$

$$= \tilde{G}_T(\theta).$$

So, in this situation, the criterion coincides with Criterion 4 in Godambe and Heyde(1987).

Following the above discussion, if in model (7), $f_s(\theta)$ is a linear function of θ, then

$$G_T^* = \int_0^T \frac{\partial}{\partial\theta} f_s(\theta) d\lambda_s (d < M >_s)^+ dM_s$$
$$= \int_0^t \frac{\partial}{\partial\theta} f_s(\theta) dM_s \in \mu''$$

belongs to μ''. If $[G_T^*]$ is a local integrable variation process, then G_T^* is a quasi-score estimating function in μ''.

CHAPTER 8

BASIC LIMIT RESULTS

1 The strong law of large numbers for multivariate martingales

1. **Introduction**

As we have mentioned in Chapter 4, the strong (or weak) law of large numbers for martingales is an important tool for studying the consistency of quasi-likelihood estimators. In this section we shall discuss and present many sufficient conditions for the strong (or weak) law of large numbers for martingales.

Let $(\Omega, \{\mathcal{F}_t\}, \{P_t\})$, $t \in D$, be a stochastic probability space. Sometimes the index set D is a discrete set but sometimes it is not. We always assume that $\{S_t\}$ is a martingale in R^p, $(1 \le p < \infty)$, with respect to $\{\mathcal{F}_t\}$. If the index set is a discrete set, sometimes we write $S_t = \sum_{i=1}^{t} X_i$, where $\{X_i\}$ is a martingale difference with respect to $\{\mathcal{F}_t\}$. Additionally, we assume $S_0 = 0$. Let B_t be a $\mathcal{F}_{t-}$-measurable $p \times p$ matrix. In the following, we shall ask for conditions under which

$$\lim_{t \to \infty} B_t S_t = 0, \qquad a.s.. \tag{1}$$

This question has been investigated by many people for a long time and various sufficient conditions for (1) have been obtained. But some of the commonly used conditions are unnecessarily restrictive. For example, Anderson and Taylor(1974) or Anderson and Moore(1976) introduced a condition in which $\lambda_{max}(B_t) = O(\lambda_{min}(B_t))$ or B_t is diagonal for all t. Also, some of the conditions appear nice but are only suitable for nonrandom matrices B_t (see Kaufmann(1987)). We know that when the quasi-likelihood method is applied to estimate parameters, we usually need a random normalizing matrix in (1)(see Chapter 5). That is why, in the following sections, we shall focus our attentions on the random norm case. In ~~the meantime~~ due course we also discuss the nonrandom norm cases in which the sufficient conditions do not require $\lambda_{max}(B_t) = O(\lambda_{min}(B_t))$.

For any square matrix A, the determinant of A is denoted by $|A|$; the minimum eigenvalue of A and the maximum eigenvalue of A are denoted by $\lambda_{min}(A)$ and $\lambda_{max}(A)$ respectively. The symbol $\|x\|$ and $\|A\|$ will be reserved for the Euclidean vector norm $(x'x)^{1/2}$ and the matrix norm $(\lambda_{max}(A'A))^{1/2}$ respectively. If A is positive semi-definite, then this norm reduces to $\|A\| = \lambda_{max}(A)$. We also define another norm for A by $\|A\|_1 = (tr(AA'))^{1/2}$. It is easy to prove that $\|\cdot\|$ and $\|\cdot\|_1$ are equivalent. For a sequence of matrices $\{B_s\}$, we say that $\{B_s\}$ is monotonic decreasing iff $\|B_s x\| \geq \|B_t x\|$, for all $x \in R^p$, when $s \leq t$.

2. Random Norm Case

Kaufmann(1987) has discussed the the strong law of large numbers for martingales when the norm matrix is a non-random matrix. Many results from his paper can not be transformed to the random norm case directly. However the basic idea he used is quite important and we shall employ it to discuss the same question in the random norm case.

Theorem 1 *Suppose that $\{S_n = \sum_{i=1}^n X_i\}$ is a martingale and $\{B_n\}$ are predictable matrices. If*

(1) $\{B_n\}$ are a.s. monotonic,

(2) $\sup_n E\sum_{i=1}^n \|B_i x_i\|^\alpha < \infty$, for some $1 \leq \alpha \leq 2$,

(3) $E(\|B_n S_n\|^2 | \mathcal{F}_{n-1}) \xrightarrow{p} 0$, as $n \to \infty$,

then

$$B_n S_n \xrightarrow{a.s.} 0, \qquad \text{as } n \to \infty.$$

Proof: Due to Corollary 1 in Kaufmann(1987), Conditions 1 and 2 imply that $\{B_n S_n\}$ converges almost surely.

In the following, we only need to prove $B_n S_n \xrightarrow{p} 0$, as $n \to \infty$. Since

$$E(\|B_n S_n\|^2 | \mathcal{F}_{n-1}) \xrightarrow{p} 0, \qquad \text{as } n \to \infty,$$

we get $E((B_n S_n)(S_n B_n)' | \mathcal{F}_{n-1}) \xrightarrow{p} 0$, as $n \to \infty$. For any $x \in R^p$ and $\varepsilon > 0$ fixed, let $y_n(\varepsilon) = P(|x' B_n S_n| > \varepsilon | \mathcal{F}_{n-1})$. Then $y_n(\varepsilon) \leq 1$ and

$$0 \leq y_n(\varepsilon) \leq \frac{x' E(B_n S_n S_n' B_n' | \mathcal{F}_{n-1}) x}{\varepsilon^2} \xrightarrow{p} 0, \qquad \text{as } n \to \infty.$$

Therefore, $\lim_{n\to\infty} P(|x'B_nS_n| > \varepsilon) = \lim_{n\to\infty} E(y_n(\varepsilon)) = 0$, for all $x \in R^p$, and we obtain $B_nS_n \xrightarrow{p} 0$, as $n \to \infty$. This completes the proof.

Remark: Theorem 2 in Kaufmann(1987) can be deduced from this theorem. Because, if B_n is nonrandom and $B_n \to 0$, we have

$$E\|B_nS_n\|^2 \le \sum_{i=1}^{n} E\|B_nX_i\|^2 \to 0, \qquad \text{as } n \to \infty,$$

and

$$P(E(\|B_nS_n\|^2|\mathcal{F}_{n-1}) > \varepsilon) \le \frac{E\|B_nS_n\|^2}{\varepsilon^2} \to 0, \qquad \text{as } n \to \infty,$$

i.e. $E(\|B_nS_n\|^2|\mathcal{F}_{n-1}) \xrightarrow{p} 0$, as $n \to \infty$.

Using the same technique as in Lemma 4 of Kaufmann(1987), we have the following lemma.

Lemma 1 *Suppose that $\{S_n = \sum_{i=1}^n X_i\}$ is a martingale in R^p. Then*

$$E\| < S >_n^{-1/2} X_n\|^2 \le pE(1 - \frac{|S_{n-1}S_{n-1}'|}{|E(S_nS_n'|\mathcal{F}_{n-1})|}).$$

In the proof of this lemma, we only need notice that

$$E\| < S >_n^{-1/2} X_n\|^2 = trE(< S >_n^{-1/2} (X_nX_n') < S >_n^{-1/2})$$

$$= trE(< S >_n^{-1/2} E(X_nX_n'|\mathcal{F}_{n-1}) < S >_n^{-1/2})$$

$$= trE(E(S_nS_n'|\mathcal{F}_{n-1})^{-1/2}(E(S_nS_n'|\mathcal{F}_{n-1}) - E(S_{n-1}S_{n-1}'|\mathcal{F}_{n-1}))E(S_nS_n'|\mathcal{F}_{n-1})^{-1/2}).$$

This is a key lemma for transforming the non-random norm results to random norm results and reveals the main difference between the non-random norm case and random norm case. In Lemma 1, if $< S >_n$ is non-random then we obtain Lemma 4 of Kaufmann.

Remark : In Lemma 1, the expectation " $E(\cdot)$" can be replaced by " $E(\cdot|\mathcal{F}_{n-1})$".

Assumption 1 There is sequence $\{c_n\}$ such that

$$(| < S >_n | - |S_{n-1}S_{n-1}'|)/| < S >_n | \le c_n, \qquad a.s.,$$

$n \ge N$ for some $N > 0$, where $E\sum_{n=1}^{\infty} c_n < \infty$.

Following Lemma 1, we can obtain the following theorems.

Theorem 2 *Suppose that* $\{S_n = \sum_{i=1}^n X_i\}$ *is a martingale. Under Assumption 1, if*

(1) $\lambda_{min} < S >_n \to \infty$, *as* $n \to \infty$, *uniformly for* $\omega \in \Omega$,

(2) $(ES_nS_n')^{-1} < S >_n \xrightarrow{p} A$, *where* A *is a nonsingular matrix,*

(3) $h(x)$ *is a positive increasing function for* $x \geq x_0$ *and*

$$\int_0^\infty \frac{dx}{xh(x)} < \infty,$$

then

$$\lim_{n\to\infty} [h(| < S >_n |)]^{-1/2} < S >_n^{-1/2} S_n = 0, \qquad a.s..$$

Proof: In the light of $\lim_{n\to\infty} \lambda_{min} < S >_n = \infty$, uniformly for $\omega \in \Omega$, for any large $M > 0$, there is a $N(M) > 0$ such that, if $n > N(M)$, then

$$\lambda_{min} < S >_n \geq M.$$

Therefore

$$\begin{aligned} 0 &\leq E(h^{-1}(| < S >_n |)E^{-1}(< S >_n)(S_nS_n')) \\ &\leq \frac{E^{-1}(< S >_n)E(S_nS_n')}{h(M^p)} = (h(M^p))^{-1}, \end{aligned}$$

for large enough n. Let $M \to \infty$; we obtain

$$\lim_{n\to\infty} E(h^{-1}(| < S >_n |)E^{-1}(< S >_n)(S_nS_n')) = 0.$$

Since $(h(| < S >_n |))^{-1/2} < S >_n^{-1/2}$ is monotonic and, by assumption,

$$\sum_{n=1}^\infty E\|(h(| < S >_n |))^{-1/2} < S >_n^{-1/2} X_n\|^2$$

$$\begin{aligned} &= \sum_{n=1}^\infty trE[(h(| < S >_n |))^{-1}(< S >_n^{-1/2} (X_nX_n') < S >_n^{-1/2})] \\ &\leq p\sum_{n=1}^\infty E((h(| < S >_n |))^{-1}(\frac{| < S >_n | - | < S >_{n-1} |}{| < S >_n |} \\ &\quad + \frac{| < S >_{n-1} | - |S_{n-1}S_{n-1}'|}{| < S >_n |})) < \infty \end{aligned}$$

(see Kaufmann(1987)), we have $(h(| < S >_n |))^{-1/2} < S >_n^{-1/2} S_n$ converges a.s..

Moreover, following the assumptions, we still have

$$\|(h(| < S >_n |))^{-1/2} < S >_n^{-1/2} S_n\|^2$$

$$\leq p\, tr(E^{1/2}(< S >_n) < S >_n^{-1} E^{1/2}(< S >_n))$$

$$\times tr(E^{-1/2}(< S >_n)(S_n S_n')E^{-1/2}(< S >_n)h^{-1}(| < S >_n |))$$

$$= p\, tr(< S >_n^{-1} E(< S >_n))\, tr(E^{-1}(< S >_n)(S_n S_n')h^{-1}(| < S >_n |))$$

$$\xrightarrow{p} 0, \qquad \text{as } n \to \infty,$$

and therefore $\lim_{n\to\infty}(h(| < S >_n |))^{-1/2} < S >_n^{-1/2} S_n = 0$, a.s..

Remark: If condition (2) is replaced by $(ES_nS_n')^{-1} < S >_n \xrightarrow{a.s.} A$, as $n \to \infty$, where A is a nonsingular matrix, then Assumption 1 can be removed.

Theorem 3 *Suppose that* $\{S_n = \sum_{i=1}^n X_i\}$ *is a martingale . Under Assumption 1, if*

(1) $\lambda_{min} E(< S >_n) \to \infty$, *as* $n \to \infty$,

(2) $(E^{-1} < S >_n) < S >_n \xrightarrow{p} A$, *where* A *is a nonsingular matrix,*

(3) $\lambda_{min}(< S >_n) = O(| < S >_n |^\delta)$, $\delta > 0$,

then $$< S >_n^{-1} S_n \xrightarrow{a.s.} 0, \text{ as } n \to \infty.$$

Proof: Since

$$\sum_{n=1}^{\infty} E\| < S >_n^{-1} X_n\|^2 < \infty,$$

we have that $< S >_n^{-1} S_n$ converges almost surely. Furthermore, we still have

$$\| < S >_n^{-1} S_n\|^2 = tr < S >_n^{-1} S_n S_n' < S >_n^{-1}$$

$$= tr(< S >_n^{-1} E < S >_n E^{-1} < S >_n S_n S_n' E^{-1} < S >_n E < S >_n < S >_n^{-1})$$

$$\leq p\, tr(E < S >_n < S >_n^{-2} E < S >_n)\, tr(E^{-1} < S >_n S_n S_n' E^{-1} < S >_n)$$

$$= p\, tr((< S >_n^{-1} E < S >_n)(< S >_n^{-1} E < S >_n)')$$

$$\times tr(E^{-1} < S >_n S_n S_n' E^{-1}) < S >_n \quad \xrightarrow{p} 0, \qquad \text{as } n \to \infty.$$

Therefore we obtain $\lim_{n\to\infty} < S >_n^{-1} S_n = 0$, a.s..

Remark: If $(E^{-1} < S >_n) < S >_n \xrightarrow{a.s.} A$, where A is a nonsingular matrix, and $\lambda_{min}(E < S >_n) = O(|E < S >_n |^\delta)$, $\delta > 0$, then Assumption 1 can be removed.

Theorem 4 *Suppose that* $\{S_n = \sum_{i=1}^n X_i\}$ *is a martingale . Under Assumption 1, if*

(1) $\lambda_{min} < S >_n \to \infty$, *a. s. , as* $n \to \infty$,

(2) *for any stopping time* T_n,

$$(E^{-1} < S >_{T_n}) < S >_{T_n} \xrightarrow{p} A,$$

where A *is a nonsingular matrix,*

(3) $h(x)$ *is a positive increasing function, such that*

$$\int_0^\infty \frac{dx}{xh(x)} < \infty,$$

then

$$h^{-1/2}(| < S >_n |) < S >_n^{-1/2} S_n \xrightarrow{a.s.} 0, \qquad as \quad n \to \infty.$$

Proof: It is obvious that Condition 3 yields

$$h^{-1/2}(| < S >_n |) < S >_n^{-1/2} S_n$$

converges a.s..

In the following we only need to prove that $h^{-1/2}(| < S >_n |) < S >_n^{-1/2} S_n$ is convergent in probability. Let $T_n = \min\{k | \lambda_{min}(< S >_k) > c_n\}$, where the nonrandom sequence $\{c_n\}$ tends to ∞. Then $\{T_n\}$ is a sequence of stopping times and

$$\|h^{-1/2}(| < S >_{T_n} |) < S >_{T_n}^{-1/2} S_{T_n}\|^2$$
$$= tr(h^{-1}(| < S >_{T_n} |) < S >_{T_n}^{-1/2} S_{T_n} S_{T_n}' < S >_{T_n}^{-1/2})$$
$$\leq p\, tr(< S >_{T_n}^{-1} E < S >_{T_n})\, tr(h^{-1}(| < S >_{T_n} |) E^{-1} < S >_{T_n} (S_{T_n} S_{T_n}')).$$

Since $tr(< S >_{T_n}^{-1} E < S >_{T_n}) \xrightarrow{p} tr(A^{-1})$, as $n \to \infty$, and

$$P(tr(h^{-1}(| < S >_{T_n} |) E^{-1} < S >_{T_n} (S_{T_n} S_{T_n}')) > \varepsilon)$$

$$\leq tr(E(h^{-1}(| < S >_{T_n} |) E^{-1} < S >_{T_n} (S_{T_n} S_{T_n}')/\varepsilon)$$

$$\leq p(\varepsilon h(c_n^p))^{-1} \to 0, \qquad as \quad n \to \infty,$$

we obtain $h^{-1/2}(|<S>_n|) <S>_n^{-1/2} S_n \xrightarrow{p} 0$, as $n \to \infty$ and the theorem follows.

Remark : The difference between Theorem 2 and Theorem 4 is the condition on $\lambda_{min} <S>_n$. In Theorem 2 the condition involves uniform convergence but in Theorem 4 almost sure convergence is required.

Theorem 5 *Suppose that $\{S_n = \sum_{i=1}^n X_i\}$ is a martingale. Under Assumption 1, if*

(1) $\lambda_{min} <S>_n \xrightarrow{a.s.} \infty$,

(2) there is a positive increasing function $h(x)$ such that

$$\int_0^\infty \frac{dx}{xh(x)} < \infty$$

and

$$h(|<S>_n|) = o(\lambda_{min} <S>_n),$$

then $<S>_n^{-1} S_n \xrightarrow{a.s.} 0$, as $n \to \infty$.

Proof: Following the assumptions, we have $h^{-1/2}(|<S>_n|) <S>_n^{-1/2} S_n$ converges a.s. and

$$\| <S>_n^{-1/2} h^{1/2}(|<S>_n|)\|^2$$

$$= tr(h(|<S>_n|) \begin{pmatrix} \lambda_{n,1}^{-1} & & 0 \\ & \ddots & \\ 0 & & \lambda_{n,p}^{-1} \end{pmatrix}) \xrightarrow{a.s.} 0, \qquad as \quad n \to \infty,$$

where $\{\lambda_{n,i}\}$ are the eigenvalues of $<S>_n$. Therefore

$$<S>_n^{-1} S_n = <S>_n^{-1/2} h^{1/2}(|<S>_n|) h^{-1/2}(|<S>_n|) <S>_n^{-1/2} S_n$$

$$\xrightarrow{a.s.} 0, \qquad as \quad n \to \infty.$$

Remark: 1. If $h(|<S>_n|) = o((\lambda_{min}(<S>_n))^{\alpha-1/2})$, where $\alpha > 2$, then $\lim_{n\to\infty} <S>_n^{-\alpha} S_n = 0$, a.s. .

2. Sometimes $h(x)$ can be chosen as $(log x)^{\delta}$, $\delta > 1$.

For the more general case, without Assumption 1, we have the following theorem.

Theorem 6 *Suppose that* $\{M_t = (M_{t-1}, \cdots, M_{t_p})'\}$ *is a martingale and* $\{B_t\}$ *are monotonic predictable matrices such that* $\|B_t\| \xrightarrow{a.s.} 0$, *as* $t \to \infty$. *If*

$$trE \int_0^t B_s d < M >_s B_s' < \infty, \qquad as \quad t \to \infty,$$

then, for any $\delta > 0$,

$$B_t^{1+\delta} M_t \xrightarrow{a.s.} 0, \qquad as \quad t \to \infty;$$

or

$$A_t B_t M_t \xrightarrow{a.s.} 0, \qquad as \quad t \to \infty,$$

where $\{A_t\}$ *are predictable matrices and* $\|A_t\| \to 0$.

Remark: For the second statement, $\|B_t\| \xrightarrow{a.s.} 0$ is not necessary.

The proof of this theorem is similar to those of the above theorems and will be omitted.

Finally, let us consider a special case which is often met when we use the quasi-likelihood method or the least squares method to estimate parameters.

Proposition 1 *Let* $\{u_t\}$, $\{X_t\}$ *be r.v.'s in* R , $\mathcal{F}_0 = \sigma\{X_1\}$ *and* $\mathcal{F}_t = \sigma \{u_1, \ldots, u_t, X_1, \ldots, X_{t-1}\}$, $t = 1, 2, \ldots$. *If* $E(u_t|\mathcal{F}_{t-1}) = 0$, $E(u_t^2|\mathcal{F}_{t-1}) = 1$ *for all* t *and* $\lim_{T\to\infty} \sum_{t=1}^T X_t^2 = \infty, a.s.$, *then*

$$E(X_1^2)^{1-2\alpha} < \infty, \qquad \alpha > 1/2$$

implies

$$\lim_{T\to\infty} \frac{\sum_{t=1}^T X_t u_t}{(\sum_{t=1}^T X_t^2)^{\alpha}} = 0, \qquad a.s..$$

Proof: Let

$$X^{(t)} = \frac{X_t u_t}{(\sum_{i=1}^t X_i^2)^{\alpha}}.$$

Since $EX^{(t)} = 0$ and

$$\sum_{t=1}^{\infty} E[X^{(t)}]^2 = E \sum_{t=1}^{\infty} \frac{X_t^2}{(\sum_{i=1}^t X_i^2)^{2\alpha}}$$

$$\leq E[1+\int_{X_1^2}^{\infty}\frac{dx}{x^{2\alpha}}]$$
$$=1+(2\alpha-1)^{-1}E(X_1^2)^{1-2\alpha}<\infty,$$

using the Kolmogrov Convergence Theorem and the Kronecker Lemma, we have

$$\sum_{t=1}^{\infty}\frac{X_t u_t}{(\sum_{i=1}^{t} X_i^2)^{\alpha}}<\infty, \qquad a.s.$$

and

$$\frac{1}{(\sum_{i=1}^{T} X_t^2)^{\alpha}}\sum_{t=1}^{T} X_t u_t \xrightarrow{a.s.} 0, \qquad as \quad T\to\infty.$$

Corollary 1 *Assume that* $\{u_t\}$ *and* $\{X_t\}$ *are random vectors in* R^p. *let* $\mathcal{F}_0 = \sigma\{X_1\}$, $\mathcal{F}_t=\sigma\{u_1,\ldots,u_t,X_1,\ldots,X_{t-1}\}$, $t=1,2,\ldots$ *and* $A_t=\sum_{s=1}^{t} X_s X_s'$. *If*

(1) $\sum_{s=1}^{t} X_{s,i}^2 \to\infty$, *as* $t\to\infty$, $i=1,2,\ldots,p$,

(2) $E(X_{1.i}^2)^{1-2\alpha}<\infty$, $\alpha>1/2$, $1\leq i\leq p$,

(3) $E(u_t|\mathcal{F}_{t-1})=0$,

(4) $E(u_t u_t'|\mathcal{F}_{t-1})=I$, $t=1,2,\ldots,$

then

$$\lim_{T\to\infty} D_T^{-\alpha}\sum_{t=1}^{T} X_t u_t'=0, \qquad a.s.,$$

where D_T *is the diagonal matrix of* A_T.

Here we have used a diagonal matrix as a norm, but using the technique discussed in the following section, we can change the result to deal with the non-diagonal matrix norm case under certain conditions.

3. Nonrandom Norm Case

In the following we shall discuss (1) where B_n is nonrandom.

Lemma 2 *If* $\{X_n\}$ *is a martingale difference sequence in* R^p, *then, for any* $\varepsilon>0$,

$$P(\max_{0\leq n\leq k}\|\sum_{l=1}^{n} X_l\|^2>\varepsilon)\leq p/\varepsilon E(\|\sum_{l=1}^{k} X_l\|^2).$$

Proof: Let $X_l=(X_{l,1},\cdots,X_{l,p})'$; then

$$P(\max_{0\leq n\leq k}\|\sum_{l=1}^{n} X_l\|^2>\varepsilon)$$

$$= P(\max_{0\le n\le k}\sum_{j=1}^{p}(\sum_{l=1}^{n}X_{l,j})^2 > \varepsilon)$$

$$\le P(\sum_{j=1}^{p}\max_{0\le n\le k}(\sum_{l=1}^{n}X_{l,j})^2 > \varepsilon)$$

$$\le \sum_{j=1}^{p}P(\max_{0\le n\le k}(\sum_{l=1}^{n}X_{l,j})^2 > \varepsilon/p)$$

$$\le p/\varepsilon\sum_{j=1}^{p}\sum_{l=1}^{k}EX_{l,j}^2$$

$$= p/\varepsilon E(\|\sum_{l=1}^{k}X_l\|^2).$$

By virtue of Lemma 2, we have the following theorems.

Theorem 7 *Assume that $\{X_n\}$ is a martingale difference sequence in R^p. If there is a subsequence $\{n_k\}$ such that*

(1) $\{B_{n,k}\}$ are symmetric and monotonic,

(2) $\|B_{n,k}\| \to 0$, as $n_k \to \infty$,

(3) for $n_{k-1} < n < n_k$, $B_{n_{k-1}}{}'B_{n_{k-1}} - B_n{}'B_n$ is semi-positive definite,

(4) if $n_k - n_{k-1} > 1$, $\lambda_{max}(B_{n_k}^{-1}B_{n_{k-1}}^2B_{n_k}^{-1}) \le c$ for some constant c,

then

$$\sum_{k=1}^{\infty}\lambda_{max}(B_{n_k}E[(\sum_{l=n_{k-1}+1}^{n_k}X_l)(\sum_{l=n_{k-1}+1}^{n_k}X_l)']B_{n_k}) < \infty \tag{2}$$

implies

$$B_nS_n \xrightarrow{a.s.} 0, \qquad as \quad n\to\infty.$$

Proof: Because of

$$\sum_{k=1}^{\infty}\lambda_{max}(B_{n_k}E[(\sum_{l=n_{k-1}+1}^{n_k}X_l)(\sum_{l=n_{k-1}+1}^{n_k}X_l)']B_{n_k}) < \infty$$

and Corollary 2 in Kaufmann(1987), we obtain $\lim_{k\to\infty}B_{n_k}S_{n_k} = 0$ a.s. . Let $T_n = B_nS_n$; then

$$\max_{n_{k-1}<n\le n_k}\|T_n - T_{n_{k-1}}\| \le \max_{n_{k-1}<n\le n_k}\|B_n\sum_{l=n_{k-1}+1}^{n}X_l\|$$

$$+\max_{n_{k-1}<n\le n_k} \|(B_n - B_{n_{k-1}})\sum_{l=1}^{n_{k-1}} X_l\|.$$

Therefore the sufficiency conditions for the theorem are

$$\lim_{k\to\infty}\max_{n_{k-1}<n\le n_k} \|B_n \sum_{l=n_{k-1}+1}^{n} X_l\| = 0, \qquad a.s. \tag{3}$$

and

$$\lim_{k\to\infty}\max_{n_{k-1}<n\le n_k} \|(B_n - B_{n_{k-1}})\sum_{l=1}^{n_{k-1}} X_l\| = 0, \qquad a.s.. \tag{4}$$

Firstly we prove (4). Because $B_{n_{k-1}}{}'B_{n_{k-1}} - B_n{}'B_n$ is positive semi-definite, so is $I - (B_{n_{k-1}}^{-1})'B_n{}'B_n(B_{n_{k-1}}^{-1})$. Then we get $tr(B_{n_{k-1}}^{-1})B_n^2(B_{n_{k-1}}^{-1}) \le p$. Furthermore, we have

$$\begin{aligned} tr(B_{n_{k-1}}^{-1}B_n) &= tr(B_{n_{k-1}}^{-1/2}B_nB_{n_{k-1}}^{-1/2}) \\ &= tr[(B_n^{1/2}B_{n_{k-1}}^{-1/2})'(B_n^{1/2}B_{n_{k-1}}^{-1/2})] \\ &\ge \lambda_{min}[(B_n^{1/2}B_{n_{k-1}}^{-1/2})'(B_n^{1/2}B_{n_{k-1}}^{-1/2})] \\ &\ge \lambda_{min}(B_n)\lambda_{min}(B_{n_{k-1}}^{-1}) \ge 0 \end{aligned}$$

(see Bodewig(1956), p66). Therefore we obtain

$$\begin{aligned} &\|B_nB_{n_{k-1}}^{-1} - I\|_1^2 \\ &= tr[B_{n_{k-1}}^{-1}B_n^2B_{n_{k-1}}^{-1} + I - 2B_{n_{k-1}}^{-1}B_n] \\ &\le tr[B_{n_{k-1}}^{-1}B_n^2B_{n_{k-1}}^{-1} + I] \le 2p, \end{aligned}$$

and then

$$\begin{aligned} &\max_{n_{k-1}<n\le n_k} \|(B_n - B_{n_{k-1}})\sum_{l=1}^{n_{k-1}} X_l\| \\ &\le \max_{n_{k-1}\le n\le n_k} \|(B_nB_{n_{k-1}}^{-1} - I)\|\|B_{n_{k-1}}\sum_{l=1}^{n_{k-1}} X_l\| \\ &\le \max_{n_{k-1}\le n\le n_k} \|(B_nB_{n_{k-1}}^{-1} - I)\|_1\|B_{n_{k-1}}\sum_{l=1}^{n_{k-1}} X_l\| \\ &\qquad \xrightarrow{a.s.} 0, \qquad \text{as } n_k \to \infty. \end{aligned}$$

Now we consider (3). In the light of Lemma 2,

$$P(\max_{n_{k-1}<n\le n_k} \|B_n \sum_{l=n_{k-1}+1}^{n} X_l\|^2 > \varepsilon)$$

$$\leq P(\max_{n_{k-1}<n\leq n_k} \|B_{n_{k-1}} \sum_{l=n_{k-1}+1}^{n} X_l\|^2 > \varepsilon)$$

$$\leq p/\varepsilon E((B_{n_{k-1}} \sum_{l=n_{k-1}+1}^{n} X_l)'(B_{n_{k-1}} \sum_{l=n_{k-1}+1}^{n} X_l))$$

$$= p/\varepsilon \; tr[(B_{n_k}^{-1} B_{n_{k-1}}^2 B_{n_k}^{-1})(B_{n_k} E((\sum_{l=n_{k-1}+1}^{n} X_l)(\sum_{l=n_{k-1}+1}^{n} X_l)')B_{n_k})(B_{n_k}^{-1} B_{N_{k-1}})]$$

$$\leq p^2 c/\varepsilon \lambda_{max}(B_{n_k} E((\sum_{l=n_{k-1}+1}^{n} X_l)(\sum_{l=n_{k-1}+1}^{n} X_l)')B_{n_k}).$$

This means that there is a constant $c_0 > 0$ such that

$$\sum_{k=1}^{\infty} P(\max_{n_{k-1}<n\leq n_k} \|B_n \sum_{l=n_{k-1}+1}^{n} X_l\|^2 > \varepsilon)$$

$$\leq c_0 \sum_{k=1}^{\infty} \lambda_{max}(B_{n_k} E((\sum_{l=n_{k-1}+1}^{n} X_l)(\sum_{l=n_{k-1}+1}^{n} X_l)')B_{n_k}) < \infty,$$

i.e. $\max_{n_{k-1}<n\leq n_k} \|B_n \sum_{l=n_{k-1}+1}^{n} X_l\| \overset{a.s.}{\to} 0$, as $n \to \infty$. Then we get the theorem.

Remark: For a sequence of matrices $\{B_n\}$, usually

$$\frac{\lambda_{max}(B_n)}{\lambda_{min}(B_n)} = O(1)$$

is not equivalent to

$$\lambda_{max}(B_n^{-1} B_{n-1}^2 B_n^{-1}) \leq c,$$

for some c, $n = 1, 2, \cdots$. Here is an example. Assume

$$B_n = \begin{pmatrix} 1/n & 0 \\ 0 & 1/n^2 \end{pmatrix}.$$

It is obvious that

$$\frac{\lambda_{max}(B_n)}{\lambda_{min}(B_n)} = n \to \infty, \qquad as \quad n \to \infty.$$

But

$$\lambda_{max}(B_n^{-1} B_{n-1}^2 B_n^{-1})$$

$$= \lambda(\begin{pmatrix} n^2/(n-1)^2 & 0 \\ 0 & n^4/(n-1)^4 \end{pmatrix}) \to 1, \qquad as \quad n \to \infty,$$

i.e. $\lambda_{max}(B_n^{-1}B_{n-1}^2B_n^{-1}) \le c$, for some c, $n = 2, 3, \cdots$. Therefore the condition (4) is different from the condition put forward by Anderson and Taylor(1974) or Anderson and Moore(1976).

Using the same idea, we can also obtain the following theorem.

Theorem 8 *Assume that $\{X_n\}$ is a martingale difference sequence in R^p. If there is a subsequence $\{n_k\}$ such that*

(1) $\{B_{n_k}\}$ is monotonic,

(2) $\|B_{n_k}\| \to 0$, as $k \to \infty$,

(3) for $n_{k-1} < n < n_k$, $B_{n_{k-1}}{}'B_{n_{k-1}} - B_n{}'B_n$ is semi-positive definite,

then

$$\sum_{k=1}^{\infty} \lambda_{max}(B_{n_{k-1}}E[(\sum_{l=n_{k-1}+1}^{n_k} X_l)(\sum_{l=n_{k-1}+1}^{n_k} X_l)']B_{n_{k-1}}) < \infty \tag{5}$$

implies

$$B_nS_n \xrightarrow{a.s.} 0, \qquad as \quad n \to \infty.$$

The difference between Theorem 7 and Theorem 8 is that in (2) the condition is on B_{n_k} but in (5) the condition is on $B_{n_{k-1}}$.

Theorem 9 *Assume that $\{X_n\}$ is a martingale difference sequence in R^p and $\{B_n\}$ is a sequence of $p \times p$ matrices. If*

(1) $\|B_n^{-1}\|^2 = tr(B_n^{-1\prime}B_n^{-1}) \to 0$, as $n \to \infty$,

(2) there is a constant $c > 1$ such that

$$\sum_{n=1}^{\infty} c^{-2n}E\|S_{i_n} - S_{i_{n-1}}\|^2 < \infty,$$

where $i_n = \max\{k \| \|B_k^{-1}\| \ge c^{-n}\}$,

then $\lim_{n\to\infty} B_n^{-1}S_n = 0$, a.s. .

Proof: From the assumptions and by virtue of Lemma 2,

$$\sum_{n=1}^{\infty} P(\max_{i_{n-1}<k\le i_n} c^{-2n}\|S_k - S_{i_{n-1}}\|^2 > \varepsilon)$$

$$\leq \sum_{n=1}^{\infty} \frac{p}{c^{2n}\varepsilon} E\|S_{i_n} - S_{i_{n-1}}\|^2 < \infty,$$

i.e.

$$\lim_{n\to\infty} \max_{i_{n-1}<k\leq i_n} c^{-2n}\|S_k - S_{i_{n-1}}\|^2 = 0, a.s.. \tag{6}$$

For any n, we may assume that $i_{n-1} < n < i_n$. Then

$$B_n^{-1}S_n = B_n^{-1}(S_n - S_{i_{n-1}} + S_{i_{n-1}} + \cdots + S_0),$$

where $S_0 = 0$. Therefore (6) yields that

$$\begin{aligned} &\|B_n^{-1}S_n\| \\ &\leq p[\|B_n^{-1}\|\|S_n - S_{i_{n-1}}\| + \cdots + \|B_n^{-1}\|\|S_{i_1} - S_0\|] \\ &\leq pc^{-(n-1)}\sum_{k=1}^{n} c^{k-1}\frac{1}{c^{k-1}} \max_{i_{k-1}<l\leq i_k} \|S_l - S_{i_{k-1}}\| \\ &\stackrel{a.s.}{\to} 0, \qquad as \quad n\to\infty, \end{aligned}$$

i.e. $\lim_{n\to\infty} B_n^{-1}S_n = 0$, a.s. .

Remark: The matrices $\{B_n\}$ in Theorem 9 are not necessarily symmetric matrices.

Theorem 10 *Assume that* $\{S_n = (S_{n_1}, \cdots, S_{n_p})'\}$ *is a martingale and* $\{A_n\} = (a_{i,j}^{(n)})$ *is a sequence of positive matrices, which satisfies* $\lambda_{max}(A_n) \to 0$*, as* $n \to \infty$*. If*

$$\sum_{n=1}^{\infty}(a_{i,i}^{(n)})ES_{n_i}^2 < \infty,$$

then

$$A_nS_n \stackrel{a.s.}{\to} 0, \qquad as \quad n\to\infty.$$

In fact Theorem 10, which is a corollary of Theorem 2.1.3. of Stout(1974), tells us that sometimes the diagonal elements of A_n can play an important role. This motivates us to seek some sufficient conditions which do not focus on eigenvalues but on the diagonal elements.

Before discussing this, we need some new notation. Let A be a matrix. If A is square, the matrix of diagonal elements of A is denoted by $diag(A)$. The set of elements of $diag(A)$ denoted by $\{diag(A)\}$. If $\{S_n\} = \{(S_{n,1}, \cdots, S_{n,p})\}$ is a martingale in R^p, the covariance matrix of S_n is denoted by F_n.

Lemma 3 *Assume that A_n is a $p \times p$ positive definite matrix. Then*

$$\frac{\lambda_{max}(A_n)}{\lambda_{min}(A_n)} = O(1)$$

induces

$$\frac{\max\{diag(A_n)\}}{\min\{diag(A_n)\}} = O(1),$$

but the converse is not true.

Proof: Let $A_n = (a_{i,j}^{(n)})$. Suppose that $a_{1,1}^{(n)} = \min\{diag(A_n)\}$. Since A_n is positive definite, then

$$\begin{aligned}
0 < |A_n| &= a_{1,1}^{(n)}|A_{1,1}^*| - a_{1,2}^{(n)}|A_{1,2}^*| + \cdots + (-1)^{p-1}a_{1,p}^{(n)}|A_{1,p}^*| \\
&\leq |a_{1,1}^{(n)}|A_{1,1}^*|| + |a_{1,2}^{(n)}|A_{1,2}^*|| + \cdots + |a_{1,p}^{(n)}|A_{1,p}^*|| \\
&\leq p|a_{1,1}^{(n)}|A_{1,1}^*|| \leq pa_{1,1}^{(n)} \max_i |A_{i,i}^*|.
\end{aligned}$$

This forces

$$\begin{aligned}
\max_n\{diag(A_n^{-1})\} &= \frac{\max_i |A_{i,i}^*|}{|A_n|} \\
&\geq (pa_{1,1}^{(n)})^{-1} = (p\min\{diag(A_n)\})^{-1}.
\end{aligned}$$

Therefore

$$\begin{aligned}
&\frac{\max\{diag(A_n)\}}{\min\{diag(A_n)\}} \leq p\max\{diag(A_n^{-1})\}\max\{diag(A_n)\} \\
&\leq p^2\lambda_{max}(A_n^{-1})\max\{diag(A_n)\} \leq p^3\frac{\lambda_{max}(A_n)}{\lambda_{min}(A_n)} = O(1).
\end{aligned}$$

But the converse is not true. For example, let

$$A_n = \begin{pmatrix} n & n-\delta \\ n-\delta & n \end{pmatrix},$$

where $0 < \delta < 2n$. Following the definition of A_n, it is positive definite, $|A_n| = \delta(2n-\delta)$ and $\max\{diag(A_n)\} = \min\{diag(A_n)\}$. It is easy to write down the eigenvalues of A_n which are

$$\lambda_{max}(A_n) = n + n\sqrt{1 - o(1)}$$

and

$$\lambda_{min}(A_n) = n - n\sqrt{1 - o(1)}.$$

Therefore we obtain

$$\lim_{n\to\infty} \frac{\lambda_{max}(A_n)}{\lambda_{min}(A_n)} = \infty.$$

<u>Lemma</u> 4 *If*

$$\frac{\max\{diag(F_n)\}}{\min\{diag(F_n)\}} = O(1)$$

and

$$\left|\frac{ES_{n,i}S_{n,j}}{\sqrt{ES_{n,i}^2 ES_{n,j}^2}}\right| \le K_0 < \frac{1}{p-1} < 1,$$

then $\|F_n^{-1}D_n\|$ *is bounded, where* $D_n = diag(F_n)$.

Proof: Since

$$\begin{aligned}
\|F_n^{-1}D_n\|^2 &= \lambda_{max}(D_nF_n^{-2}D_n) \\
&\le tr(D_nF_n^{-2}D_n) \\
&= tr(D_n^{1/2}F_n^{-1}D_n^{1/2}D_n^{-1}D_n^{1/2}F_n^{-1}D_n^{1/2}D_n) \\
&\le ptr(D_n^{1/2}F_n^{-1}D_n^{1/2}D_n^{-1}D_n^{1/2}F_n^{-1}D_n^{1/2})\lambda_{max}(D_n) \\
&= ptr(D_n^{1/2}F_n^{-1}D_n^{1/2}D_n^{1/2}F_n^{-1}D_n^{1/2}D_n^{-1})\lambda_{max}(D_n) \\
&\le p^2\lambda_{max}(D_n^{1/2}F_n^{-1}D_n^{1/2}D_n^{1/2}F_n^{-1}D_n^{1/2})\frac{\lambda_{max}(D_n)}{\lambda_{min}(D_n)} \\
&\le cp^2(\lambda_{max}(D_n^{1/2}F_n^{-1}D_n^{1/2}))^2,
\end{aligned}$$

for some constant c, the sufficient condition for the boundness of $\|F_n^{-1}D_n\|$ is the boundness of

$$\lambda_{max}(D_n^{1/2}F_n^{-1}D_n^{1/2}).$$

However the boundness of $\lambda_{max}(D_n^{1/2}F_n^{-1}D_n^{1/2})$ is equivalent to

$$\lambda_{min}(D_n^{-1/2}F_nD_n^{-1/2}) > 0.$$

Hence it is sufficient to prove $\lambda_{min}(D_n^{-1/2}F_nD_n^{-1/2}) > 0$. Since

$$D_n^{-1/2}F_nD_n^{-1/2} =$$

$$= \begin{pmatrix} 1 & \frac{ES_{n,1}S_{n,2}}{\sqrt{ES_{n,1}^2 ES_{n,2}^2}} & \cdots & \frac{ES_{n,1}S_{n,p}}{\sqrt{ES_{n,1}^2 ES_{n,p}^2}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{ES_{n,p}S_{n,1}}{\sqrt{ES_{n,p}^2 ES_{n,1}^2}} & \frac{ES_{n,p}S_{n,2}}{\sqrt{ES_{n,p}^2 ES_{n,2}^2}} & \cdots & 1 \end{pmatrix}$$

and

$$\left|\frac{ES_{n,i}S_{n,j}}{\sqrt{ES_{n,i}^2 ES_{n,j}^2}}\right| \le K_0 < \frac{1}{p-1} < 1,$$

we get

$$\lambda_{min}(D_n^{-1/2} F_n D_n^{-1/2}) \ge \min_i \left(1 - \sum_{k \ne i} \left|\frac{ES_{n,i}S_{n,k}}{\sqrt{ES_{n,i}^2 ES_{n,k}^2}}\right|\right)$$
$$\ge 1 - (p-1)K_0 > 0.$$

Therefore $\|F_n^{-1}D_n\|^2$ is bounded, and so is $\|F_n^{-1}D_n\|$.

Using Lemma 4, we have the following theorem.

Theorem 11 *Assume that $\{S_n\}$ is a martingale and $D_n = diag(F_n)$. If $ES_{n,i}^2 \to \infty$, as $n \to \infty$, $1 \le i \le p$, then*

(1) $D_n^{-\alpha} S_n \xrightarrow{a.s.} 0$, as $n \to \infty$, $\alpha > 1/2$.

(2) If F_n satisfies the conditions in Lemma 4,

$$F_n^{-1} S_n \xrightarrow{a.s.} 0, \qquad as \quad n \to \infty.$$

Proof: The result $\lim_{n\to\infty} D_n^{-\alpha} S_n = 0$, a.s., follows immediately from the one-dimensional results $\lim_{n\to\infty}(ES_{n,i}^2)^{-\alpha} S_{n,i} = 0$, a.s., $i = 1, 2, \cdots, p$.

Since F_n satisfies the conditions of Lemma 4, there is a constant C_0 such that $\|F_n^{-1}D_n\| < C_0$, and using the result of first part of the theorem,

$$\|F_n^{-1}S_n\| \le p\|F_n^{-1}D_n\|\|D_n^{-1}S_n\|$$
$$\le C_0 p\|D_n^{-1}S_n\| \qquad \xrightarrow{a.s.} 0, \qquad \text{as } n \to \infty.$$

2 The central limit theorem for semimartingales

1. Introduction

Given a process $\{X_t\}$, central limit theory considers the problem of finding conditions under which one can obtain

$$A_T X_T \xrightarrow{d} N(0,1), \qquad \text{as } T \to \infty. \tag{7}$$

where A_T is a normalizing matrix, deterministic or random. This is a classical question. There are comprehensive results for the case where $\{X_t\}$ are sums of independent random variables or martingale differences or for which the moments satisfy certain conditions. Here, we want to discuss other kinds of conditions for semimartingales such that (7) holds. The emphasis is on conditions on the characteristics of the semimartingale.

It is well known that a semimartingale can be determined by its characteristics. As discussed in the functional central limit theorem for semimartingales (see Lipster and Shiryaev (1980, 1981)), we focus our conditions on the characteristics of semimartingale.

We denote by $\mathcal{L}$ the set of all local martingales M such that $M_0 = 0$ and by $\mathcal{V}$ the set of all real-valued processes A that are cádlag, adapted, with $A_0 = 0$ and for which each path $t \to A_t(\omega)$ has a finite variation over each finite interval [0,t]. A process X is called a semimartingale if X has the form $X = X_0 + M + A$ where X_0 is finite-valued and $\mathcal{F}_0$-measurable, $M \in \mathcal{L}$ and $A \in \mathcal{V}$. We denote by $\mathcal{S}$ the space of all semimartingales. A special semimartingale is a semimartingale X which admits a decomposition $X = X_0 + M + A$ as above, with a process A that is predictable. $\mathcal{S}_p$ denotes the set of all special semimartingales. If X is a special semimartingale, the unique decomposition $X = X_\circ + M + A$ such that $M \in \mathcal{L}$ and A is a predictable element of $\mathcal{V}$ is called the canonical decomposition of X.

Write b_t^d for the class of all functions $h : R^d \to R^d$ which are bounded, with compact support, and satisfy $h(x) = x$ in a neighbourhood of 0. We call h a truncation function if $h \in b_t^d$.

Suppose that X is a semimartingale in R^d . Let $h \in b_t^d$. Then $\triangle X_s - h(\triangle X_s) \neq 0$ only if $\| \triangle X_s\| \geq b$ for some $b \geq 0$ and the formulae

$$\check{X}(h)_t = \sum_{s \leq t} [\triangle X_s - h(\triangle X_s)],$$

$$X(h) = X - \check{X}(h),$$

define a d-dimensional process $\check{X}(h)$ in $\mathcal{V}^d$ (i.e. its components are in $\mathcal{V}$) and a d-dimensional semimartingale $X(h)$. From the construction of $X(h)$, $X(h)$ is a special semimartingale (i.e. its components are in $\mathcal{S}_p$). Then there is a canonical decomposition of $X(h)$:

$$X(h) = X_\circ + M(h) + B(h). \tag{8}$$

Definition: $(B(h), C, \nu)$ are the characteristics (associated with h) of X, where $C = (C^{i,j})_{i,j \leq d}$ is a continuous process in $\mathcal{V}^{d \times d}$, namely

$$C^{i,j} = \langle X^{i,c}, X^{j,c} \rangle$$

X^c being the continuous martingale part of X and ν being a predictable random measure on $R^d \times R^d$, namely the compensator of the random measure μ^X associated with the jumps of X.

It is known that C and ν do not depend on the choice of the function h, while $B = B(h)$ does (Jacod and Shiryaev(1987), p76).

Definition: Let $h \in b_t^d$. The predictable process $\tilde{C}$ of $\mathcal{V}^{d \times d}$ defined by

$$\tilde{C}^{i,j} = \langle M(h)^i, M(h)^j \rangle$$

where $M(h)$ is defined in (7) called the modified second characteristic of X (associated with h).

We denote by $C_2(R^d)$ the set of all continuous bounded functions:$R^d \to R$ which are 0 around 0 and have a limit at infinity. Also $C_1(R^d)$ is a subclass of $C_2(R^d)$ having only nonnegative functions,which contains all functions $g_a(x) = (a|x|-1)^+ \bigwedge 1$ for all positive rationals a, and with the property that if η_n and η are positive measures on

R^d which do not charge $\{0\}$ and are finite on the complement of any neighbourhood of 0; then $\eta_n(f) \to \eta(f)$ for all $f \in C_1(R^l)$ implies $\eta_n(f) \to \eta(f)$ for all $f \in C_2(R^d)$ (so $C_2(R^d)$ is a convergence-determining class for the weak convergence induced by $C_2(R^d)$)(Jacod and Shiryaev(1987), p354).

2. Results

In the following we just consider the one-dimensional case. All the results presented here will hold in the multi-dimensional case by changing "$|\cdot|$" to "$\|.\|$ as appropriate. Suppose $X = (X_t)_{t\geq 0}$ is a semimartingale in R with the characteristics $(B(h), C, \nu)$ where $h \in b_t^1$.

Let $Y = AX$,where A is a positive r.v. and $h'(x) = A^{-1}h(Ax)$. Then

$$\check{Y}_t(h) = \sum_{s\leq t}[\triangle Y_s - h(\triangle Y_s)]$$

$$= \sum_{s\leq t}[A \triangle X_s - h(A \triangle X_s)]$$

$$= A\sum_{s\leq t}[\triangle X_s - h'(\triangle X_s)]$$

$$= A\check{X}_t(h'),$$

and therefore

$$Y(h) = Y - \check{Y}(h) = A(X - \check{X}(h'))$$

$$= A(X_\circ + M(h') + B(h')). \tag{9}$$

Suppose the characteristics of Y associated with h are $(B'(h)$,C', $\nu')$. Comparing with (9), we find

$$B'(h) = AB(h') = A(B(h) + (h' - h) * \nu').$$

Also we obtain

$$C' = A^2C,$$

because of

$$A^2\tilde{C}(h') = \tilde{C}'(h) = C' + (h^2) * \nu' - \sum_{s\leq\cdot}(\triangle B'_s)^2$$

$$= C' + h^2(AX) * \nu - \sum_{s\leq\cdot}(A \triangle B_s(h'))^2$$

$$= C' + A^2 h'^2 * \nu - \sum_{s\leq \cdot} A^2(\triangle B_s(h'))^2,$$

and

$$A^2\tilde{C}(h') = A^2(C + h'^2 * \nu - \sum_{s\leq \cdot}(\triangle B_s(h'))^2).$$

Therefore we can write the characteristics of Y associated with h as

$$(B'(h), C', \nu') = (AB(h'), A^2C, \nu')$$

$$= (A(B(h) + (h' - h) * \nu), A^2C, \nu').$$

After the above discussion which makes the relations between (B, C, ν) and (B', C', ν') clear, we now establish the following theorem.

Theorem 12 *Suppose that X is a semimartingale with characteristics $(B(h), C, \nu)$ associated with $h \in b_t^d$. Let $A = (A_t)_{t\geq 0}$ be a positive process. If*

(1) $\quad A_T \xrightarrow{a.s.} 0, \qquad$ *as* $T \to \infty$,

(2) $\quad A_T(B_T(h) - (X - h) * \nu_t) \xrightarrow{P} 0, \qquad$ *as* $T \to \infty$,

(3) $\quad A_T^2 C_T + (A_T X)^2 * \nu_T \xrightarrow{P} 1, \qquad$ *as* $T \to \infty$,

(4) $\quad A_T^2 \int_0^T \int_{|x|\geq\lambda} x^2 \nu(ds, dx) \xrightarrow{P} 0, \qquad$ *as* $T \to \infty$, $\lambda \to \infty$,

then

$$A_T X_T \xrightarrow{d} N(0,1), \qquad \textit{as } T \to \infty.$$

Proof: Let $h'_T(x) = A_T^{-1} h(A_T x)$ and $Y_s^T = A_T X_{s\cdot T}$, $0 \leq s \leq 1$. Then $Y^T = (Y_s^T)_{0\leq s\leq 1}$ is a semimartingale on some stochastic base (Ω, $\mathcal{F}$, F, P) and $h'_T \in b_t$. Without loss of generality, we can suppose that (Ω, $\mathcal{F}$, F, P) is the same for all Y^T, $T = 1, 2, \cdots$. From the discussion above, the characteristic of Y^T associated with h are

$$(B^T(h), C^T, \nu^T) = (A_T B_{T\cdot}(h'_T), A'_T C_{T\cdot}, \nu^T).$$

Let Y be a Wiener process in $[0, 1]$. The only thing that needs to be proved is

$$Y_1^T = A_T X_T \xrightarrow{d} Y_1 \qquad \text{as } T \to \infty.$$

Following Theorem VIII 2.4 in Jacod and Shiryaev(1987), it is sufficient to prove the following facts:

(i) $B_1^T(h) \xrightarrow{p} 0,$ as $T \to \infty$,

(ii) $\tilde{C}_1^T \xrightarrow{p} 1,$ as $T \to \infty$,

(iii) $\sup_{s\le 1} \nu^T(\{s\} \times \{|x| > \varepsilon\}) \xrightarrow{p} 0,$ for $\varepsilon \ge 0$, as $T \to \infty$,

(iv) $g * \nu_1^T \xrightarrow{p} 0,$ as $T \to \infty$, for all $g \in C_1(R^1)$.

Firstly, we prove (iii). Due to the definition of ν, we have

$$\sup_{s\le 1} \nu^T(\{s\} \times \{|x| > \varepsilon\}) = \sup_{s\le 1} \nu(\{sT\} \times \{|x| > \varepsilon A_T^{-1}\})$$

$$\le \nu([0,T], |x| > \varepsilon A_T^{-1}) = \int_0^T \int_{\{|x|>\varepsilon A_T^{-1}\}} \nu(ds, dx)$$

$$\le \frac{1}{\varepsilon^2} \int_0^T \int_{\{|x|>\varepsilon A_T^{-1}\}} A_T^2 x^2 \nu(ds, dx). \tag{10}$$

Because of Assumption 4, $\forall \delta > 0$, there are $T_1 > 0$, $\lambda_1 > 0$ such that, if $T > T_1$ and $\lambda > \lambda_1$,

$$P(\int_0^T \int_{\{|x|>\lambda\}} A_T^2 x^2 \nu(ds, dx) > \varepsilon^2 \delta) < \delta/2.$$

Using Assumption 1, for $\delta > 0$, there is $E \subset \Omega$ and $T_2 \ge T_1$, such that, if $T > T_2$

$$|A_T^{-1}\varepsilon| > \lambda_1 \qquad \omega \in E,$$

and

$$P(E^c) < \delta.$$

Then, combining these results we obtain, if $T > T_2$,

$$P(\sup_{s\le 1} \nu^T(\{s\} \times \{|x| > \varepsilon\}) > \delta)$$

$$\le P(\frac{1}{\varepsilon^2} \int_0^T \int_{\{|x|>\varepsilon A_T^{-1}\}} A_T^2 x^2 \nu(ds, dx) > \delta)$$

$$\le P((\frac{1}{\varepsilon^2} \int_0^T \int_{\{|x|>\varepsilon A_T^{-1}\}} A_T^2 x^2 \nu(ds, dx) > \delta) \bigcap E) + P(E^c)$$

$$\leq \frac{\delta}{2} + \frac{\delta}{2} = \delta.$$

i.e.

$$\sup_{s \leq 1} \nu^T(\{s\} \times \{|x| > \varepsilon\}) \xrightarrow{p} 0, \qquad \text{as } T \to \infty.$$

Secondly, we prove (iv). For any $g \in C_1(R^d)$, there are c and a such that $|g(x)| \leq c|x|^2\chi_{(|x|>a)}$(Jacod and Shiryaev(1987), p376). Therefore

$$g * \nu_1^T = \int_0^T \int_R g(A_T x)\nu(ds, dx)$$

$$\leq \int_0^T \int_R c|A_T x|^2 \chi_{(|A_T x|>a)}\nu(ds, dx)$$

$$= \int_0^T \int_{\{|x| \leq \lambda\}} c|A_T x|^2 \chi_{(|A_T x|>a)}\nu(ds, dx)$$
$$+ \int_0^T \int_{\{|x| > \lambda\}} c|A_T x|^2 \chi_{(|A_T x|>a)}\nu(ds, dx).$$

In the above inequality,

$$\int_0^T \int_{\{|x| > \lambda\}} c|A_T x|^2 \chi_{(|A_T x|>a)}\nu(ds, dx)$$

$$\leq \int_0^T \int_{\{|x| > \lambda\}} c|A_T x|^2 \nu(ds, dx) \xrightarrow{p} 0, \qquad \text{as } T \to \infty. \qquad (11)$$

By virtue of Assumption 1, for $\forall \varepsilon > 0$, there is $E \subset \Omega$ such that $P(E^c) < \varepsilon/2$, and A_T converges uniformly to 0 on E. Therefore there is a $T_\circ > 0$ such that, if $T > T_\circ$,

$$A_T \lambda \leq a.$$

and, if $T \geq T_\circ$,

$$P(\int_0^T \int_{\{|x| \leq \lambda\}} c|A_T x|^2 \chi_{(|A_T x|>a)}\nu(ds, dx) > \varepsilon)$$

$$\leq P(E^c) + P((\int_0^T \int_{\{|x| \leq \lambda\}} c|A_T x|^2 \chi_{(|A_T x|>a)}\nu(ds, dx) > \varepsilon)) \cap E)$$
$$\leq P(E^c) = \frac{\varepsilon}{2} \quad .$$

i.e. $$\int_0^T \int_{\{|x| \leq \lambda\}} c|A_T x|^2 \chi_{(|A_T x|>a)}\nu(ds, dx) \xrightarrow{p} 0, \qquad \text{as } T \to \infty. \qquad (12)$$

Combining (11) and (12), the result (iv) follows.

Thirdly, we prove (i). From the relation between B_1^T and B_T, we have

$$B_1^T(h) = A_T B_T(h) + A_T(X - h) * \nu_T - (A_T X - h(A_T X)) * \nu_T.$$

By assumption,

$$A_T B_T(h) + A_T(X - h) * \nu_T \xrightarrow{p} 0, \qquad \text{as } T \to \infty.$$

The only thing remaining to prove is

$$(A_T X - h(A_T X)) * \nu_T \xrightarrow{p} 0, \qquad \text{as } T \to \infty.$$

Since $h \in b_t$, there exists $a_1 > 0$, such that

$$h(x) = x \qquad |x| \leq a_1.$$

Therefore

$$|(A_T X - h(A_T X)) * \nu| = |\int_0^T \int_R (A_T x - h(A_T x))\nu(ds, dx)|$$

$$\leq |\int_0^T \int_{\{|A_T x| \leq a_1\}} (A_T x - h(A_T x))\nu(ds, dx)| + |\int_0^T \int_{\{|A_T x| > a_1\}} (A_T x - h(A_T x))\nu(ds, dx)|$$

$$\leq \int_0^T \int_{\{|A_T x| > a_1\}} |A_T x| \nu(ds, dx) + \int_0^T \int_{\{|A_T x| > a_1\}} |h(A_T x)| \nu(ds, dx)$$

$$\leq \frac{1}{a_1} A_T^2 \int_0^T \int_{\{|x| > A_T^{-1} a_1\}} x^2 \nu(ds, dx) + \frac{c}{a_1^2} \int_0^T \int_{\{|x| > A_T^{-1} a_1\}} x^2 \nu(ds, dx)$$

for some constant $c > 0$. Using the same method of proof as in the second part of the proof, we obtain

$$(A_T X - h(A_T X)) * \nu_T \xrightarrow{p} 0, \qquad \text{as } T \to \infty.$$

Finally, we prove (ii). Following the definition of $\tilde{C}$,

$$\tilde{C}_1^T(h) = A_T^2 \tilde{C}(h_T') = A_T^2 (C_T + (h')^2 * \nu_T)$$

$$= A_T^2 C_T + h^2(A_T X) * \nu_T$$

$$= A_T^2 C_T + (A_T X)^2 * \nu_T + h^2(A_T X) * \nu_T - (A_T X)^2 * \nu_T.$$

Now Assumption 3 gives

$$A_T^2 C_T + (A_T X)^2 * \nu_T \xrightarrow{p} 1 \qquad \text{as } T \to \infty,$$

and we can obtain

$$|h^2(A_T X) * \nu_T - (A_T X)^2 * \nu_T| \xrightarrow{p} 0. \qquad \text{as } T \to \infty,$$

because of

$$\begin{aligned}
&|h^2(A_T X) * \nu_T - (A_T X)^2 * \nu_T| \\
&= |\textstyle\int_0^T \int_R (h^2(A_T x) - (A_T x)^2)\nu(ds, dx)| \\
&= |\textstyle\int_0^T \int_{\{|A_T x| \leq a_1\}} (h^2(A_T x) - (A_T x)^2)\nu(ds, dx) \\
&+ \textstyle\int_0^T \int_{\{A_T x| > a_1\}} (h^2(A_T x) - (A_T x)^2)\nu(ds, dx)| \\
\leq\ & c^2 \textstyle\int_0^T \int_{\{A_T x| > a_1\}} \nu(ds, dx) + A_T^2 \int_0^T \int_{\{|x| > a_1 A_T^{-1}\}} x^2 \nu(ds, dx) \\
\leq\ & c^2/a_1^2 \textstyle\int_0^T \int_{\{|x| > a_1 A_T^{-1}\}} A_T^2 x^2 \nu(ds, dx) + A_T^2 \int_0^T \int_{\{|x| > a_1 A_T^{-1}\}} x^2 \nu(ds, dx).
\end{aligned}$$

So the theorem is proved.

<u>Corollary</u> **2** *If*

(1) $\qquad [X]_T^{-1/2} \xrightarrow{a.s.} 0, \qquad$ *as* $T \to \infty$,

(2) $\qquad [X]_T^{-1/2}(B(h) + (X - h) * \nu_T) \xrightarrow{a.s.} 0, \qquad$ *as* $T \to \infty$,

(3) for any fixed $\varepsilon > 0$

$$[X]_T^{-1}\Big(\sum_{s \leq T} \Delta X_s{}^2 - \int_0^T \int_{\{|x| \leq \varepsilon\}} x^2 \nu(ds, dx) + \sum_{s \leq T}\Big(\int_{\{|x| \leq \varepsilon\}} x\nu(\{s\}, dx)\Big)^2\Big) \xrightarrow{p} 0,$$

as $T \to \infty$,

(4) $\qquad [X]_T^{-1} \int_0^T \int_{\{|x| > \lambda\}} x^2 \nu(ds, dx) \xrightarrow{p} 0, \qquad$ *as* $T \to \infty,\ \lambda \to \infty$,

then

$$[X]_T^{-1/2} X_T \xrightarrow{d} N(0, 1).$$

Proof: Knowing the proof of Theorem 1, we see that the only thing which remains to be proved is that $[X]_T^{-1} \tilde{C}_T(h') \xrightarrow{p} 1$, *as* $T \to \infty$.

By definition we have

$$\tilde{C}_T^T(h) = [X]_T^{-1}\tilde{C}_T(h')$$

$$= [X]_T^{-1}[C_T + (h'_T)^2 * \nu_T - \sum_{s\leq T}(\int_R h'_T(x)\nu(\{s\}\times dx))^2]$$

$$= [X]_T^{-1}[\langle X^c\rangle_T + (h'_T)^2 * \nu_T - \sum_{s\leq T}(\int_R h'_T(x)\nu(\{s\}\times dx))^2]$$

$$= 1 - [X]_T^{-1}[\sum_{s\leq T}\triangle X_s{}^2 - (h'_T)^2 * \nu_T + \sum_{s\leq T}(\int_R h'_T(x)\nu(\{s\}\times dx))^2].$$

The last term can be rewritten as

$$[X]_T^{-1}[\sum_{s\leq T}\triangle X_s{}^2 - (h'_T)^2 * \nu_T + \sum_{s\leq T}(\int_R h'_T(x)\nu(\{s\}\times dx))^2]$$

$$= [X]_T^{-1}\sum_{s\leq T}\triangle X_s{}^2 - h^2([X]_t^{-1/2}X) * \nu_T + \sum_{s\leq T}(\int_R h([X]_T^{-1/2}x)\nu(\{s\},dx))^2$$

$$= [X]_T^{-1}\sum_{s\leq T}\triangle X_s{}^2 - \int_0^T\int_R h^2([X]_T^{-1/2}x) * \nu(ds,dx) + \sum_{s\leq T}(\int_R h([X]_T^{-1/2}x)\nu(\{s\},dx))^2.$$

Due to definition of h, there exists an $a > 0$, such that

$$h(x) = x, \qquad\qquad |x| \leq a$$

and h is bounded, i.e. $|h(x)| \leq c$ for some $c > 0$; thus

$$[X]_T^{-1}[\sum_{s\leq T}\triangle X_s{}^2 - (h'_T)_2 * \nu_T + (\int_R h'_T(x)\nu(\{s\},dx))^2]$$

$$= [X]_T^{-1}\sum_{s\leq T}\triangle X_s{}^2 - \int_0^T\int_{\{|[X]_T^{-1/2}x|\leq a\}}[X]_T^{-1}x^2\nu(ds,dx)$$

$$+\sum_{s\leq T}(\int_{\{|[X]_T^{-1/2}x|\leq a\}}[X]_T^{-1/2}x\nu(\{s\},dx))^2$$

$$-\int_0^T\int_{\{|x|>a[X]_T^{-1/2}\}}h^2([X]_T^{-1/2}x)\nu(ds,dx)$$

$$+\sum_{s\le T}\Big(\int_{\{|x|>a[X]_T^{1/2}\}} h([X]_T^{-1/2}x)\nu(\{s\},dx)\Big)^2$$

$$+2\sum_{s\le T}\int_{\{|[X]_T^{-1/2}x|\le a\}} h([X]_T^{-1/2}x)\nu(\{s\},dx)\int_{\{|x|>a[X]_T^{1/2}\}} h([X]_T^{-1/2}x)\nu(\{s\},dx).$$

But the fourth term on the right hand side of this equation,

$$\Big|\int_0^T\int_{\{|x|>a[X]_T^{-1/2}\}} h^2([X]_T^{-1/2}x)\nu(ds,dx)\Big|$$

$$\le \frac{c^2}{a^2}\int_0^T\int_{\{|x|>a[X]_T^{-1/2}\}}[X]_T^{-1}x^2\nu(ds,dx)\xrightarrow{p}0, \qquad \text{as } T\to\infty.$$

For the same reason, the fifth term and the sixth term tend in probability to zero as $T\to\infty$.

Let us consider the first three terms. We can rewrite them as

$$[X]_T^{-1}\sum_{s\le T}\triangle x_s{}^2-\int_0^T\int_{\{|x|\le a[X]_T^{1/2}\}}[X]_T^{-1}x^2\nu(ds,dx)$$

$$+\sum_{s\le T}\Big(\int_{\{|x|\le a[X]_T^{1/2}\}}[X]_T^{-1/2}x\nu(\{s\},dx)\Big)^2$$

$$=[X]_T^{-1}\sum_{s\le T}\triangle x_s{}^2-\int_0^T\int_{\{|x|\le\lambda\}}[X]_T^{-1}x^2\nu(ds,dx)$$

$$+\sum_{s\le T}\Big(\int_{\{|x|\le\lambda\}}[X]_T^{-1/2}x\nu(\{s\},dx)\Big)^2$$

$$-\int_0^T\int_{\{\lambda<|x|\le a[X]_T^{1/2}\}}[X]_T^{-1}x^2\nu(ds,dx)$$

$$+\sum_{s\le T}\Big(\int_{\{\lambda<|x|\le a[X]_T^{1/2}\}}[X]_T^{-1/2}x\nu(\{s\},dx)\Big)^2$$

$$+2\sum_{s\le T}\int_{\{|x|\le\lambda\}}[X]_T^{-1/2}x\nu(\{s\},dx)\sum_{s\le T}\int_{\{\lambda<|x|\le a[X]_T^{1/2}\}}[X]_T^{-1/2}x\nu(\{s\},dx).$$

So for $\forall\varepsilon>0$,

$$P(|[X]_T^{-1}\sum_{s\le T}\triangle x_s{}^2-\int_0^T\int_{\{|x|\le a[X]_T^{1/2}\}}[X]_T^{-1}x^2\nu(ds,dx)$$

$$+\sum_{s\le T}\Big(\int_{\{|x|\le a[X]_T^{1/2}\}}[X]_T^{-1/2}x\nu(\{s\},dx)\Big)^2| > \varepsilon)$$

$$\le P(|[X]_T^{-1}\sum_{s\le T}\Delta x_s{}^2 - \int_0^T\int_{\{|x|\le\lambda\}}[X]_T^{-1}x^2\nu(ds,dx)$$

$$+\sum_{s\le T}\Big(\int_{\{|x|\ge\lambda\}}[X]_T^{-1/2}x\nu(\{s\},dx)\Big)^2| > \frac{\varepsilon}{4})$$

$$+P(|\int_0^T\int_{\{\lambda<|x|\le a[X]_T^{1/2}\}}[X]_T^{-1}x^2\nu(ds,dx)| > \frac{\varepsilon}{4})+$$

$$+P(|\sum_{s\le T}\Big(\int_{\{\lambda<|x|\le a[X]_T^{1/2}\}}[X]_T^{-1/2}x\nu(\{s\},dx)\Big)^2| > \frac{\varepsilon}{4})$$

$$+P(|2\sum_{s\le T}\int_{\{|x|\le\lambda\}}[X]_T^{-1/2}x\nu(\{s\},dx)\sum_{s\le T}\int_{\{\lambda<|x|\le a[X]_T^{1/2}\}}[X]_T^{-1/2}x\nu(\{s\},dx)| > \frac{\varepsilon}{4}).$$

By Assumptions 1, 3 and 4, and using the same technique as in Theorem 1, it is easy to obtain that

$$[X]_T^{-1}\sum_{s\le T}\Delta x_s{}^2 - \int_0^T\int_{\{|x|\le a[X]_T^{1/2}\}}[X]_T^{-1}x^2\nu(ds,dx)$$

$$+\sum_{s\le T}\Big(\int_{\{|x|\le a[X]_T^{1/2}\}}[X]_T^{-1/2}x\nu(\{s\},dx)\Big)^2 \xrightarrow{p} 0, \qquad \text{as } T\to\infty.$$

So

$$[X]_T^{-1}\tilde{C}_T(h') \xrightarrow{p} 1, \qquad \text{as } T\to\infty$$

and

$$[X]_T^{-1}X_T \xrightarrow{d} N(0,1), \qquad \text{as } T\to\infty$$

<u>**Corollary**</u> **3** *If*

(1) $\langle X^c\rangle_T^{-1/2} = C_T^{-1/2} \xrightarrow{a.s.} 0,$ *as* $T\to\infty,$

(2) $\langle X^c\rangle_T^{-1/2}(B(h) + (X\cdot h) * \nu_T) \xrightarrow{p} 0,$ *as* $T\to\infty,$

(3) for any fixed $\varepsilon > 0$

$$\langle X^c\rangle_T^{-1}\int_0^T\int_{\{|x|<\varepsilon\}}x^2\nu(ds,dx) - \langle X^c\rangle_T^{-1}\sum_{s\le T}\Big(\int_{\{|x|\le\varepsilon\}}x\nu(\{s\},dx)\Big)^2 \xrightarrow{p} 0,$$

as $T \to \infty$,

(4) $\langle X^c\rangle_T^{-1} \int_0^T \int_{\{|x|>\epsilon\}} x^2 \nu(ds, dx) \xrightarrow{P} 0,$ *as* $T \to \infty$, *as* $\lambda \to \infty$.

then

$$\langle X^c\rangle_T^{-1/2} X_T \xrightarrow{d} N(0,1). \quad \text{as } T \to \infty$$

The proof is the same as that of Corollary 1.

The following Corollary can be seen as a special application of the Theorem 1.

Corollary 4 *Suppose that* $S_n = \sum_1^n X_i$ *and* $\{X_i\}_{i\geq 1}$ *are i.i.d., in which* $EX_1 = 0$ *and* $E{X_1}^2 < \infty$. *Then*

$$\frac{1}{\sqrt{nE{X_1}^2}} \sum_1^n X_i \xrightarrow{d} N(0,1).$$

Proof: Following Theorem 1, one only needs to prove that

(i) $\frac{1}{\sqrt{nE{X_1}^2}}(B_n(h) - (X-h) * \nu_n) \xrightarrow{P} 0,$ as $n \to \infty$,

(ii) $\frac{1}{nE{X_1}^2} C_n + (\frac{1}{\sqrt{nE{X_1}^2}} X)^2 * \nu_n \xrightarrow{P} 1,$ as $n \to \infty$,

(iii) $\frac{1}{\sqrt{nE{X_1}^2}}(X^2 \chi_{\{|X|>\lambda\}}\nu_n \xrightarrow{P} 0,$ as $n \to \infty$, $\lambda \to \infty$,

where $(B_n(h),\ C_n,\ \nu_n)$ are the characteristics of $S_n = \sum_1^n X_i$. Since $\{X_i\}_{i\geq 1}$ are i.i.d., following p375 and p371 in Jacod and Shiryaev(1987), (i) is valid. Following p97 in Jacod and Shiryaev(1987), C_n=0 because $\{X_i\}_{i\geq 1}$ are i.i.d. and

$$\left(\frac{1}{\sqrt{nE{X_1}^2}} X\right)^2 * \nu_n = X^2 * \nu_1^n$$

$$= nE\left(\frac{X_1}{\sqrt{nE{X_1}^2}}\right)^2 = 1.$$

So (ii) is true.

Finally, by calculating

$$\frac{1}{nE{X_1}^2}(X^2 \chi_{\{|X|>\lambda\}}) * \nu_n$$

$$= \frac{n}{nE{X_1}^2} E(X_1^2 \chi_{\{|X_1|>\lambda\}} \xrightarrow{P} 0 \qquad \text{as } \lambda \to \infty,$$

(iii) is proved and we obtain the Corollary.

References

[1] P.K.Andersen, O.Borgan, R.D.Gill and N.Keiding (1988), Censoring, truncation and filtering in statistical models based on counting process, Contemporary Math., Vol.8, 19-60.

[2] T.W.Anderson and J.B.Taylor (1974), Some theorems concerning strong consistency of least squares estimates in linear regression models, Working paper No.47 Inst. Math. Studies Social Sci. Stanford Univ.

[3] B.D.O.Anderson and J.B.Moore (1976), A matrix Kronecker lemma, Linear Algebra Appl., 15, 227-234.

[4] V.V.Anh (1988), Nonlinear least squares and maximum likelihood estimation of a heteroscedastic regression model, Stoch. Process. Appl., 29, 317-333.

[5] E.Arjas and P.Haara (1984), A marked point process approach to censored failure data with complicated covariates, Scand. J. Statist., 11, 193-209.

[6] O.E.Barndorff-Nielsen and M.Sørensen(1989), Information quantities in non-classical settings, Research Report No. 192. Department of Theoretical Statistics, Aarhus University, Denmark.

[7] O.E.Barndorff-Nielson and M.Sørensen (1989), Asymptotic likelihood theory for stochastic process, A Research Report No. 169. Dept. of Theoretical Statistics, Aarhus University.

[8] I.V.Basawa and B.L.S.Prakasa Rao (1980), Asymptotic inference for stochastic processes, Stochastic. Process. Appl., 10, 221-254.

[9] I.V.Basawa and D.J.Scott (1983), Asymptotic Optimal Inference for Non-ergodic Models, Lecture Notes in Statistics, 17, Springer-Verlag, New York.

[10] I.V.Basawa and P.J.Brockwell (1984), Asymptotic conditional inference for regular non-ergodic models with an application to autoregressive processes, Ann. Statist., 12, 161-171.

[11] I.V.Basawa (1991), Generalized score test for composite hypothese, Estimating Functions, Clarendon Press, Oxford.

[12] D.M.Bates and D.G.Watts (1988), Nonlinear Regression Analysis and Its Applications, Wiley, New York.

[13] S.K.Berberian (1961), Introduction to Hilbert Space, Oxford Uni. Press, New York.

[14] P.Billingsley (1961), Statistical Inference for Markov Processes, Chicago: University of Chicago Press.

[15] E.Bodewig (1956), Matrix Calculus, Amsterdam, North-Holland.

[16] E.L.Bradley (1973), The equivalence of maximum likelihood and weighted least squares estimates in the exponential family, J. Amer. Statist. Assoc, 68, 199-200.

[17] I.S.Chang and C.A.Hsiung (1988), Counting process approach to time-sequential and sequential point estimation with censored data, Sequential Anal., 7, 127-148.

[18] I.S.Chang and C.A.Hsiung (1990), Time-sequential point estimation through estimating equations, Ann. Statist., 18, 1378-1388.

[19] P. Chen (1980), σ-integrability and its application (in Chinese), Acta Math. Sinica, 23, 183-191.

[20] D.R.Cox and D.V.Hinkley (1974), Theoretical Statistics, Chapman and Hall, London.

[21] M.Davidian and R.J.Carroll (1988), A note on extended quasi-likelihood , J. R. Statist. Soc., B, 50, 74-82.

[22] C.Dellacherie and P.A.Meyer (1975), Probabilités et Potentiels, 2e êdition, Chapitres, V-VIII, Hermann, Paris.

[23] A.F.Desmond (1989), The theory of estimating equations, Encyclopedia of Statistical Sciences, supplement volume, Kotz, S., Johnson, N.L. and Read, C.B., Eds. Wiley, New York.

[24] A.F.Desmond (1991), Quasi-likelihood, stochastic processes, and optimal estimating functions, Oxford Statistical Science Series 7, Estimating Functions, 133-146.

[25] A.F.Desmond (1991), Optimal estimating functions quasi-likelihood statistical modelling (reprint).

[26] B.Efron and D.V.Hinkley (1978), Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information, Biometrika 65, 457-487.

[27] P.D.Feigin (1985), Stable convergence of semimartingales, Stoch. Process. Appl., 19, 125-134.

[28] D.Firth (1987), On the efficiency of quasi-likelihood estimation, Biometrika, 74, 233-245.

[29] R.A.Fisher (1922), On the mathematical foundation of theoretical statistics, Philos. Trans. R. Soc. (Lond.) Ser., A. 222, 309-368.

[30] R.A.Fisher (1925), Theory of statistical estimation, Proc. Camb. Philos. Soc., 22, 700-725.

[31] R.Fürth (1918), Statistik und Wahrscheinlichkeitnachwirkung, Physik. Zeitschr, 19, 421-426.

[32] V.P.Godambe (1960), An optimum property of regular maximum likelihood estimation, Ann. Math. Statist., 31, 1208-1212.

[33] V.P.Godambe (1976), Conditional likelihood and unconditional optimum estimating equations, Biometrika, 63, 277-284.

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

[35] V.P.Godambe and M.E.Thompson (1974), Estimating equations in the presence of a nuisance parameter, Ann. Statist., 2, 568-571.

[36] V.P.Godambe and M.E.Thompson (1985), Logic of least squares revisited(preprint).

[37] V.P.Godambe and M.E.Thompson (1989), An extension of quasi-likelihood estimation, J. Statist. Planning Inference, 22, 137-152.

[38] V.P.Godambe and B.K.Kale (1991), Estimating functions: an overview, Oxford Statistical Science Series 7, Estimating Functions, 3-20.

[39] P.Hall and C.C.Heyde (1980), Martingale Limit Theory and its Application, Academic Press, New York.

[40] P.R.Halmos (1956), Measure Theory, Van Nostrand, Princeton, New Jersey.

[41] C.R.Heathcote (1965), A branching process allowing immigration, J.R.Statist. Soc., B27, 138-143.

[42] C.R.Heathcote (1966), Corrections and comments on the paper " a branching process allowing immigration", J.R.Statist. Soc., B28, 213-217.

[43] C.C.Heyde (1988), Fixed sample and asymptotic optimality for classes of estimating function, Contemporary Math., 80, 241-247.

[44] C.C.Heyde and R.Gay (1989), On asymptotic quasi-likelihood estimation, Stoch. Process. Appl., 31, 223-236.

[45] C.C.Heyde (1987), On combining quasi-likelihood estimating functions, Stoch. Process. Appl., 25, 281-287.

[46] C.C.Heyde (1989,a), On efficiency for quasi-likelihood and composite quasi-likelihood methods, Statistical Data Analysis and Inference, Elsevier Science Publishers B.V. North-Holland.

[47] C.C.Heyde (1989,b), Quasi-likelihood and optimality for estimating functions : some current unifying themes, Bull. Int. Statist. Inst., 53, Book 1, 19-29.

[48] C.C.Heyde (1992), On best asymptotic confidence intervals for parameters of stochastic process, Ann. Statist. (to appear).

[49] C.C.Heyde and E.Seneta (1972), Estimation theory for growth and immigration rates in a multiplicative process, J. Appl. Prob., 9, 235-256,

[50] C.C.Heyde and Y.-X. Lin (1991), Approximate confidence zones in an estimating function context, Estimating Functions, Clarendon Press, Oxford.

[51] C.C.Heyde and Y.-X. Lin (1992), On quasi-likelihood methods and estimation for Branching processes and Heteroscedastic regression models, Austral. J. Statist. (to appear).

[52] J.E.Hutton and P.I.Nelson (1984), A mixing and stable central limit theorem for continuous time martingales, Technical Report No.42, Dept. of Statistics, Kausas State University.

[53] J.E.Hutton and P.I.Nelson (1986), Quasi-likelihood estimation for semimartingales, Stoch. Process. Appl., 22, 245-257.

[54] M.Jacobsen (1989), Right censoring and martingale methods for failure time data, Ann. StatistC. 17, 1133-1156.

[55] J.Jacod and A.N.Shiryaev (1987), Limit Theorems for Stochastic Processes, Springer-Verlag, New York.

[56] B.Jørgensen (1983), Maximum likelihood estimation and large sample inference for generalized linear and non-linear regression models, Biometrika, 70, 19-28.

[57] N.C.Kakwami and D.B. Gupta (1967), Note on the bias of the Prais and Aitchison and Fisher's iterative estimators in regression analysis with heteroscedastic errors, Internat. Statist. Rev., 35, 291-295.

[58] H.Kaufmann (1987), On the strong law of large numbers for multivariate martingales, Stoch. Process. Appl., 26, 73-85.

[59] H.Kaufmann (1989), On weak and strong consistency of maximum likelihood estimator in stochastic processes, preprint.

[60] N.Keiding and R.D.Gill (1987), Random truncation models and Markov processes, Research Report 87/3, Statistical Research Unit, University of Copenhagen, Also issued as report MS-R8702, Centre for Mathematics and Computer Science, Amsterdam.

[61] U.Küchler and M.Sørensen (1989), Exponential families of stochastic processes: a unifying semimartingale approach, Inter. Statist. Rev., 57, 123-144.

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

[63] K.Y.Liang (1990), Quasi-likelihood and estimating functions: an overviews, paper presented at the Workship on Statistical Estimating Quations and Their applications, La Trobo University, Australia.

[64] D.Lepinglé (1977), Sur le comportment asymptotique des martingales locals, Springer-Verlag Lecture Notes in Math., 649, 148-161.

[65] R.Sh.Liptser and A.N.Shiryayev (1986), Theory of Martingales, Kluwer Academic Publishers, London.

[66] R.Sh.Liptser and A.N.Shiryaev (1977), Statistics of Random Processes, New York, Springer.

[67] R.SH.Lipster and A.N.Shiryaev (1980), A functional central limit theorem for semimartingales, Theor. Probab. Appl., 25, 667-688.

[68] R.SH.Lipster and A.N.Shiryaev (1981), On necessary and sufficient conditions in the functional central limit theorem for semimartingale, Theor. Probab. Appl., 26, 130-135.

[69] C.W.Lloyd-Smith (1990), Limit theorems for branching processes in a random environment, preprint.

[70] M.Loéve (1963), Probability Theory, Van Nostrand, New York.

[71] P.A.Meyer (1976) , Un cours sur les intégrales stochastiques , Sém. Probab. X. Springer-Verlag Lecture Notes in Math., 511.

[72] P.A.Meyer (1977), Notes sur les intégrales stochastiques, I-VI, Sém, Probab. XI, Springer-Verlag Lecture Notes in Math., 581.

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

[74] P. McCullagh and J. A. Nelder (1989), Generalized Linear Models, 2nd ed., Chapman and Hall, New York.

[75] D.L.McLeish and C.G.Small (1988), The Theory and Application of Statistical Inference Functions, Springer Lecture Notes in Statistics, 44, New York.

[76] C.N.Morris (1982), Natural exponential families with quadratic variance functions, Ann. Statist., 10, 65-80.

[77] C.N.Morris (1983), Natural exponential families with quadratic variance functions: statistical theory, Ann. Statist., 11, 515-529.

[78] R.Morton (1987), A generalized linear model with nested strata of extra-Poisson variation, Biometrika, 74, 247-257.

[79] A.J.Nelder and R.W.Wedderburn (1972), Generalized linear models, J. Roy. Statist. Soc. ser., A, 135, 370-384.

[80] J.A.Nelder and D.Pregibon (1987), An extended quasi-likelihood function, Biometrika, 74, 221-232.

[81] H.T.Nguyen and G.S.Rogers (1989), Fundamentals of Mathematical Statistics, Vol I & Vol II, Springer Texts in Statistics, Springer-Verlag, Berlin.

[82] M.R.Osborne (1992), Fisher's method of scoring, Int. Statist. Rev. (to appear).

[83] S.J.Prais (1953), A note on heteroscedastic errors in regression analysis, Internat. Statist.Rev., 21, 28-29.

[84] S.J.Prais and J. Aitchison (1954), the grouping of observations in regression analysis, Internat. Statist. Rev., 22, 1-22.

[85] S.J.Prais and H.S.Houthakker(1955), The Analysis of Family Budgets, New York, Cambridge University Press.

[86] R.L.Prentice (1988), Correlated binary regression with covariates specific to each binary observation, Biometrics, 44, 1033-1048.

[87] M.P.Quine (1970), The multitype Galton-Watson process with immigration, J. Appl. Prob., 7, 411-422.

[88] C.R.Rao (1965), Linear Statistical Inference and its Applications, Wiley,New York.

[89] P.Révész (1968), The Laws of Large Numbers, Academic Press, New York.

[90] L.C.G.Rogers and D.Williams (1987), Diffusions, Markov Processes and Martingales, Vol. 2, Itô Calculus, Wiley, Chichester.

[91] G.A.F.Seber and C.J.Wild (1989), Nonlinear Regression, Wiley, New York.

[92] M.Sørensen (1990), On quasi likelihood for semimartingales, Stoch. Process. Appl., 35, 331-346.

[93] W.Stout (1974), Almost Sure Convergence, Academic Press, New York.

[94] T.J.Sweeting (1986), Asymptotic conditional inference for the offspring mean of a supercritical Galton-Watson process, Ann. Statist., 14, 925-933.

[95] P.J.Szabłowski (1988), Optimal, recursive procedures of identification, Comput. Math. Applic., 16, 229-246.

[96] H.Theil (1951), Estimates and their sampling variance of parameters of certain heteroscedastic distributions, Internat. Statist. Rev., 19, 141-147.

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

[98] C.Z.Wei and J.Winnicki (1989), Some asymptotic results for branching processes with immigration, Stoch. Process. Appl., 31, 261-282.

[99] J.Winnicki (1988), Estimation theory for branching process with immigration, Contemporary Math., 80, 301-322.

[100] J.A. Yan (1981), The Introduction of Martingales and Stochastic Integration (in Chinese), Science Press, Beijing, China.

[101] N.M.Yanev and S.Tchoukova-Dantcheva (1980), On the statistics of branching processes with immigration, C. R. Acad. Sci. Bulg., 33, 463-471.