Multiple Roots of Estimating Functions and Applications

by

Zejiang Yang

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2000

The University of Waterloo requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Abstract

An estimating function may give multiple solutions. This creates a certain amount of confusion as to which of these roots is the most appropriate choice as an estimate of parameter. This thesis discusses the existing methods, and proposes two new approaches to choose the best root among multiple roots. One approach is based on the root intensity, which is an extension of the probability density function of an estimator to the multiple root case. This method is also applied to some practical examples such as logistic regression models with measurement error and bivariate normal mixture models. Another one is the shifted information method, which can be used in transformation models and, in particular, the location models. Though multiple roots of estimating functions arise in some cases, it can be shown that under some regularity conditions, there is a unique root with high probability in a given bounded closed set which includes the true value.

Acknowledgements

First of all, I wish to express my deepest gratitude to my supervisor, Professor Christopher G. Small, for his invaluable supervision, generous contributions of time and financial support during the completion of this thesis. I have benefited greatly from his stimulating discussions, thoughtful advice and profound knowledge. I am also very thankful to my thesis committee, Professor Mary E. Thompson, Professor Don L. McLeish, Professor Jiahua Cheng, Professor Adam W. Kolkiewicz, Professor Tony S. Wirjanto and my thesis examiner Professor Christopher C. Heyde for their many very helpful suggestions and comments. I also obtained an extensive knowledge for preparation of my thesis from the classes with Professor Jerry F. Lawless, Professor Phelim P. Boyle, Professor K. Steve Brown, Professor Bovas Abraham, Professor David E. Matthews and Professor Richard J. Cook.

My thanks are also due to Professor William J. Welch and Mrs. Linda Lingard for their work during the preparation of this thesis.

Finally, I also thank my wife, Dongqiong Yang, and my son, Yuange Yang, for their support and understanding.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Preliminaries

The expression *estimating function* is a very general statistical term. By *estimating function* we mean a function of both the parameters and the data, on which the inference for parameters in a statistical model is based. It is used in many aspect of theoretical and applied statistics, ranging from statistical inference, stochastic processes, time series, survey sampling, biostatistics, to finance. The theory on estimating functions has been developed since the publication of Godambe's (1960) work on this theory. Heyde (1997) surveyed the major works on both the theoretical and applied aspects of this theory.

In classical statistical theory, least square (LS) and maximum likelihood (ML) methods for estimation are widely used. These two methods have been merged into a single method of estimation from the perspective of estimating function. In fact, almost all methods of estimation correspond to a set of estimating functions. Obviously, method of moment, least square and maximum likelihood are

based on a set of simple estimating functions. In the case of nuisance parameters, many approaches have been proposed to eliminate or reduce their impact on inference. These approaches, including *marginal likelihood*, *integrated likelihood*, *conditional-likelihood*, *profile log-likelihood* and *partial likelihood* also give rise to a set of estimating functions. In time series, biostatistics and finance, there are many statistical models developed from the theory of stochastic processes. In this case, the estimating functions typically involve more complex stochastic process settings.

In section 1.2, we will review some basic theory of estimating functions. First, the concept of *optimal estimating function*, also known as *quasi-score estimating function* is introduced. The initial quasi-score function, first considered by Wedderburn (1974), is applied in a statistical model that only specifies a relationship between the mean and variance. One should note that the *quasi-score estimating function* discussed in this thesis is considered more generally than that in Wedderburn's original definition.

Most of the literature is about the estimating functions themselves. Since inference for parameters is based on estimating functions, from which the estimated parameters are obtained, it is important to explore the properties of estimates based on the estimating functions. In particular, when there are several roots of estimating functions, it is important to rule out extraneous roots. As Hanfelt and Liang (1995) pointed out, "Modifying the method to discriminate among roots is an area of future research". Studying this problem constitutes the main part of this thesis. In section 1.3, we will briefly discuss this issue and related methodology on this issue. A more comprehensive discussion will be given in Chapter 2. In this thesis, we will propose two approaches to choose the best root as a parameter estimator for the true value. One method is based on the *root intensity*, which is an extension of the probability density function of an estimator to the multiple root case. The

concept of root intensity will be discussed in Chapter 3. Another method is based on *shifted information*, which will be discussed in Chapter 5.

1.2 Theory of Estimating Functions

In this subsection, the basic concept of estimating functions will be stated. The related results are outlined in Heyde (1997).

Definition 1.1 An *unbiased estimating function* $G(\boldsymbol{\theta})$ is a function of the data and unknown parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset R^k$, such that an estimator of the unknown parameter can be obtained as its root, with the property $E(G(\boldsymbol{\theta})) = 0$.

The following estimating function derived from the estimating function $G(\boldsymbol{\theta})$:

$$G^{(s)} = -(E\dot{G})'(EGG')^{-1}G \tag{1.1}$$

is called a *standardized estimating function*, where $\dot{G} = (\partial G_i/\partial\theta_j)$, and $()'$ denotes transpose.

The *information criterion* is

$$\mathcal{E}(G) = E(G^{(s)}G^{(s)'}) = (E\dot{G})'(EGG')^{-1}(E\dot{G}) \tag{1.2}$$

which is the Fisher information for the score estimating function.

Definition 1.2 G^* is the *optimal estimating function* within a class of estimating function $\mathcal{H}$ if $G^* \in \mathcal{H}$ and

$$\mathcal{E}(G^*) - \mathcal{E}(G) \tag{1.3}$$

is nonnegative definite for all $G \in \mathcal{H}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

In addition to the above definition, there are several equivalent conditions, such as trace criterion, determinant criterion, smallest eigenvalue criterion and average variance criterion (Heyde, 1997).

The following is a useful and practical criterion for optimal estimating function, which is due to Heyde (1988).

Theorem 1.1 G^* is an optimal estimating function within $\mathcal{H}$ if

$$E(G^{*(s)}G^{(s)'}) = E(G^{(s)}G^{*(s)'}) = E(G^{(s)}G^{(s)'}) \tag{1.4}$$

or equivalently

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

is a constant matrix for all $G \in \mathcal{H}$. Conversely, if $\mathcal{H}$ is convex and $G^* \in \mathcal{H}$ is an optimal estimating function, then (1.4) holds.

There is a similar criterion for optimality in the asymptotic sense (Heyde, 1997, p28), which can be applied to martingales estimating functions. In that case, the optimal estimating function maximizes the martingale information. Estimating functions that are optimal in either sense will be referred to as *quasi-score estimating function* and the corresponding estimator as *quasi-likelihood estimator.*

Wedderburn (1974) proposed the original quasi-score estimating functions as a basis for analyzing a generalized linear regression. He also discussed the existence and uniqueness of the maximum likelihood estimates for some generalized linear

models (see Wedderburn, 1976). The usual likelihood approach needs to specify the form of the distribution of the observations, but the quasi-score function is based only on the first two moments of observations, and provides an analogue to fully parametric likelihood functions for inference. Let $\boldsymbol{x}^t = (x_1, \ldots, x_n)$ be random observations with the joint distribution $p(\boldsymbol{\theta})$ for some k-dimensional parameter $\boldsymbol{\theta}$ in the parameter space $\boldsymbol{\Theta}$, and the first two moments:

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = E\boldsymbol{x} = (\mu_1(\boldsymbol{\theta}), \ldots, \mu_n(\boldsymbol{\theta}))'$$
$$V(\boldsymbol{\theta}) = (cov(x_i, x_j)) = (V_{ij}(\boldsymbol{\theta})) \qquad i, j = 1, \ldots, n.$$

The Wedderburn's quasi-score estimating equation (see McCullagh and Nelder, 1989) is defined as

$$\boldsymbol{q}(\boldsymbol{\theta}, \boldsymbol{x}) = (\dot{\boldsymbol{\mu}}(\boldsymbol{\theta}))'(V(\boldsymbol{\theta}))^{-1}(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta})) = 0 \tag{1.5}$$

where $\dot{\boldsymbol{\mu}}(\boldsymbol{\theta})$ is a $n \times k$-dimensional derivative matrix. The function on the left hand side of (1.5) can be shown to be the projection of the score function into the class $\mathcal{G}$ of linear unbiased estimating functions of the form $a(\boldsymbol{\theta})(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta}))$. It is also an optimal estimating function in $\mathcal{G}$ in Godambe's sense (1960). Based on the results of orthogonal estimating functions, Godambe and Thompson (1989) extended the Wedderburn's quasi-score estimating function by incorporating possible knowledge of the skewness, kurtosis and higher moments of the underlying distribution.

Furthermore, the *generalized estimating equation* (GEE) approach (Liang and Zeger, 1986; Diggle, Liang and Zeger, 1994) was formulated to deal with the problems of longitudinal data analysis where one typically has a series of repeated measurements of a response variable, together with a set of covariates. This approach deals with a longitudinal data set consisting of responses y_{it}, $t = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, K$, where i indexes the individuals, and t indexes the repeated observations per individual. Observations on different individuals would be expected to

be independent, while those on the same individual would be correlated over time. Let the covariance matrix of $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})$ be $V_i = V_i(\boldsymbol{\mu}_i, \boldsymbol{\alpha}_i)$, $i = 1, 2, \ldots, K$ where $\boldsymbol{\mu}_i = \boldsymbol{\mu}_i(\boldsymbol{\theta})$ is the vector of means for the ith individual and $\boldsymbol{\alpha}_i$ is the vector of the parameters that includes variance and correlation components. Then the *generalized estimating equation* is:

$$\boldsymbol{q}(\boldsymbol{\theta}, \boldsymbol{y}) = \sum_{i=1}^{K} (\dot{\boldsymbol{\mu}}_i(\boldsymbol{\theta}))' V_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}(\boldsymbol{\theta})) = 0. \tag{1.6}$$

In practice, the unknown V_i is replaced by "working" or approximate covariance matrices.

Another important class of estimating functions – *martingale estimating function* – arises in time series and stochastic process models. Such process $(X_t, \mathcal{F}_t)$ admits a continuous semimartingale representation:

$$X_t = X_0 + M_t + A_t$$

where X_0 is a random variable, M_t is a continuous local martingale and A_t is a continuous process with a locally bounded variation. In the continuous case, the widely applicable class $\mathcal{H}$ of martingale estimating functions is:

$$G_t(\boldsymbol{\theta}) = \int_0^t \alpha_s(\boldsymbol{\theta}) dM_s(\boldsymbol{\theta}) \tag{1.7}$$

where $\alpha_t(\boldsymbol{\theta})$ is predictable with respect to filtration $\{\mathcal{F}_t\}$, and $M_t(\boldsymbol{\theta})$ is the local martingale due to the semimartingale representation of the process. The continuous martingale estimating function has be used in studying some finance models. In the discrete time case, the martingale estimating function is:

$$G(\boldsymbol{\theta}, \boldsymbol{x}) = \sum_{i=1}^{n} g_i(\boldsymbol{\theta}, \boldsymbol{x})$$

with the martingale property

$$E[g_i(\boldsymbol{\theta}) | \mathcal{F}_{i-1}] = g_{i-1}(\boldsymbol{\theta})$$

where $\mathcal{F}_{i-1} = \sigma(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{i-1})$. A typical example of this is the autoregression model in time series.

In the past several decades, many methods have been developed for estimating functions. These methods are used in many different fields of theoretical and applied statistics such as biostatistics and finance. Apart from Godambe's (1960, 1985) optimality idea which underlies estimating function theory, the optimal estimating function can be looked upon as a projection on a class of unbiased estimating functions (McLeish & Small 1988, 1992; Small & McLeish 1988, 1989, 1994). This idea was developed into a useful way to search for an approximation to estimating functions. In particular, it provides a potential way to choose the best root among multiple roots (see Section 1.3 and Chapter 2).

1.3 Multiple Roots of Estimating Functions

As discussed above, the estimating function method is widely used in theoretical and applied statistics such as biostatistics and finance. However, this method is plagued by the problem of multiple roots. A basic example of multiple solutions is the score estimating function for the Cauchy location model. Multiple roots can also occur in the likelihood equations for mixture models and nonlinear regression models. More examples will be provided in Chapter 2.

For maximum likelihood estimation, the accepted practice is to use the likelihood itself to discriminate between multiple roots. As the roots to the score estimating equation represent various local maxima and minima of likelihood function, the root corresponding to the global maximum of the likelihood is usually chosen as the estimator of parameter. In many cases, the maximum likelihood estimator is consistent. However, there do exist examples where the global maximum

of the likelihood is an inconsistent estimator while at the same time another local maximum of the likelihood is consistent. See Lehmann(1983, p420). Therefore, one may not just use the maximum of the likelihood to discriminate between roots.

For estimating functions other than the score estimating function, the roots cannot easily be interpreted as local maxima and minima of some real valued function. Finding an analogue of the likelihood for more general estimating functions has motivated the search for potential functions. McLeish and Small (1992) and Small and McLeish (1994) proposed the *projected likelihood function* method. B. Li (1993), Hanfelt and Liang (1995) later developed the *approximate likelihood ratio* method, which provides a possible way to pick up the best root for the quasi-likelihood setting. Heyde (1997) proposed three simple direct methods to choose the correct root for estimating functions. Singh and Mantel (1999) also suggested an alternative criterion which minimizes the square of a supplementary estimating function. From another viewpoint, Small and Yang (1999) considered the distribution of the roots as a random set and proposed the concept of *root intensity*. Based on root intensity, this thesis proposes a new method to choose the best root among multiple roots, since it can be shown that under some regularity conditions, the root intensity at the true value tends to infinity as the sample size approaches infinity (see Chapter 4). For transformation distributions, the *shifted information* method is also suggested to solve multiple root problem. Simulation results demonstrate that the root intensity and the shifted information methods are reasonable. These methods are also applied to some practical examples such as logistic regression models with measurement error and bivariate normal mixture models.

Chapter 2

Multiple Roots of Estimating Functions

2.1 Preliminaries

Estimating functions are widely used in statistics. However, in some cases, the estimating equation can produce several roots. This can create a certain amount of confusion as to which of these roots is the most appropriate choice for estimating the parameter. A basic example is the score estimating function for the Cauchy location model with density $f(x;\theta) = \{\pi[1+(x-\theta)^2]\}^{-1}$ (see Section 2.2.1, Section 3.4). Another classical example constructed by LeCam (1979) indicates that the conditions of Theorem 2.3 of Chapter 6 in Lehmann's book (1983) are not enough to ensure the consistency of the maximum likelihood estimator. Multiple roots often appear in mixture models (see Section 2.2.6) and nonlinear regression models. This problem also arises in the Littlewood model (see Section 2.2.4) in software reliability. In addition, the score estimating function in the Tobit regression model seems to

have multiple roots since the Hessian is not necessarily negative definite; However, Olsen (1978) indicated that it has only one root.

An interesting problem is how to discriminate among multiple roots in the estimating equations. There have been several ways proposed in the literature to attack this problem. The common goal of all these methods is to find an analogue of the likelihood function whose local maxima and minima occur at the roots of the estimating function. By maximizing this objective function, the hope is to be able to distinguish among the roots. However, in practice it is difficult to find a complete satisfactory objective function with the required properties. The first method is the approximate estimation function method. McLeish and Small (1992) and Small and McLeish (1994) proposed the *projected likelihood function*, which projected the likelihood ratios for independent observations into a class of estimating functions consisted of the tensor products. B. Li (1993) suggested *linear projected likelihood ratio* as an approximate likelihood ratio for the Wedderburn's quasi-score estimating functions. Hanfelt and Liang (1995) later developed the *approximate likelihood ratio* method to general estimating functions.

Furthermore, B. Li (1997) also extended this idea and the minimax approach to the generalized estimating equations. It also provide a possible way to pick up the best root. Recently, Heyde and Morton (1998) proposed three simple direct methods to choose the correct root for estimating functions. The methods involves (1) examining the asymptotics to see which root provides a consistent result; (2) picking the root for which $\dot{\boldsymbol{G}}(\boldsymbol{\theta})$ behaves asymptotically as its expected value $E_{\theta}\{\dot{\boldsymbol{G}}(\boldsymbol{\theta})\}$; and (3) using a least square or goodness of fit criterion to select the best root. Singh and Mantel (1999) developed this least square criterion by providing a general guideline to ensure the suitability of the criterion, and proposing a criterion for general estimating functions. This criterion chooses the root which minimizes

the square of the supplementary estimating function. Wang and Small (1998) put forward a general method to test consistency of roots to an estimating function by bootstrapping the quadratic local likelihood ratio. This method also provides an alternative way to choose the best root among multiple roots. In Section 2.3, we will describe these methods briefly. From another viewpoint, Small and Yang (1999) considered the distribution of the roots as a random set and proposed the concept of *root intensity*, which is an extension of the probability density function of an estimator to the multiple root case. Since it can be shown that under some regularity conditions (see Chapter 4), the first order root intensity at the true value tends to infinity as the sample size increases. As a result, we propose a new approach to choose the best root based on root intensity. For the transformation models, we develop an information criterion to select the best root.

2.2 Examples

2.2.1 Cauchy Location Model and Normal Stratified Sampling

The Cauchy location model is a classical example of multiple roots. Suppose a sample is taken from a Cauchy location model with density function $f(x;\theta) = 1/\{\pi[1+(x-\theta)^2]\}$. The score estimating equation is

$$\sum_{i=1}^{n} \frac{2(x_i-\theta)}{1+(x_i-\theta)^2} = 0$$

which is equivalent to a polynomial equation of degree $2n-1$. That is, it may admit $2n-1$ distinct roots. Reeds (1985) showed that if the true value $\theta = 0$, r_n

is the number of roots, then for each k,

$$\lim_{n\to\infty} P\{\frac{r_n - 1}{2} = k\} = \frac{e^{-1/\pi}}{\pi^k k!}. \tag{2.1}$$

That is, the number of false local maxima of the Cauchy score estimating function is asymptotically Poisson with parameter $1/\pi$. This is consistent with the computer experiment results reported in Barnett (1966).

When estimating the normal mean in a stratified sampling, an estimating function similar to the above score estimating equation for the Cauchy location model can be derived. Suppose $x_{i1}, \cdots, x_{in_i}$ are taken from $N(\mu, \sigma_i^2)$ $(i = 1, 2, \cdots, m)$, let $\bar{x}_i = \sum x_{ij}/n_i$ and $s_i^2 = \sum(x_{ij} - \bar{x}_i)^2/n_i$. We are interested in estimating the common mean μ based on x_{ij} $(i = 1, \cdots, m, j = 1, \cdots, n_i)$. Neyman & Scott (1948) considered an estimating function of the form

$$\sum_{i=1}^{m} \frac{w_i(\bar{x}_i - \mu)}{s_i^2 + (\bar{x}_i - \mu)^2} = 0 \tag{2.2}$$

with general weights w_i. When $w_i = n_i - 1$, it is the estimating function advocated by Kalbfleish & Sprott (1970) based on sufficiency and ancillarity arguments. The profile likelihood results in an estimating equation with $w_i = n_i$ in (2.2). Neyman & Scott (1948) found that the estimator with $w_i = n_i - 2$ is asymptotically more efficient than the maximum likelihood estimator.

Similar to the Cauchy location model, (2.2) corresponds to a polynomial equation of degree $2m - 1$, which may have $2m - 1$ roots. However the Cauchy location model and stratified normal models are quite different, so the probabilities of multiple roots arising in the two models are quite different. That is, the distribution of these roots is not similar.

2.2.2 Estimation of the Correlation Coefficient

Let us consider a sample (x_i, y_i) from a bivariate normal distribution (X, Y) which is standardized to have mean $\mu_x = \mu_y = 0$ and variances $\sigma_x^2 = \sigma_y^2 = 1$. We assume that there is an unknown correlation coefficient ρ between X and Y. The estimating equation which is equivalent to the score function for ρ is:

$$S(\rho) = \rho(1-\rho^2) + \frac{(1+\rho^2)\sum_{i=1}^n x_i y_i}{n} - \frac{\rho\sum_{i=1}^n (x_i^2 + y_i^2)}{n} = 0 \qquad (2.3)$$

which can have as many as three real roots in the interval $(-1, 1)$. Small, Wang and Yang (1999) discussed all cases. When the discriminant of the quadratic equation $S'(\rho) = 0$

$$D = 4[\frac{\sum x_i y_i}{n}]^2 + 12[1 - \frac{\sum (x_i^2 + y_i^2)}{n}] \qquad (2.4)$$

is zero or strictly negative, $S'(\rho) = 0$ has at most one real solution. Thus $S'(\rho)$ is nonnegative or nonpositive since $S'(\rho)$ is a quadratic function. Then the cubic function $S(\rho)$ will be monotone, thus having a unique real root. From the law of large numbers, $\sum x_i y_i/n \to E(XY) = \rho$, $\sum x_i^2/n \to E(X^2) = 1$, and $\sum y_i^2/n \to E(Y^2) = 1$, so D converges to $4\rho^2 - 12$ as $n \to \infty$. Therefore, with probability converging to one, the estimating equation (2.3) will have a unique root for large samples. This is consistent with our general result in Chapter 5 (Theorem 5.3). Figure 3.1 in Chapter 3 shows the root intensity function for this estimating equation.

2.2.3 Censored Regression Models

The Tobit regression model (Greene, 1989) is:

$$y_i^* = \boldsymbol{\beta}' \boldsymbol{x}_i + \epsilon_i \qquad y_i = \max(0, y_i^*) \qquad (2.5)$$

where $\boldsymbol{x}_i$ is the explanatory vector observed for observation i, y_i is the observed counterpart to the latent dependent variable y_i^*, and $\epsilon_i \sim N(0, \sigma^2)$. We use $\sum_0$ to indicate summation over all observations where $y_i = 0$, $\sum_1$ is the summation over all observations where $y_i > 0$, and N_1 is the number of observations where $y_i > 0$. Apart from a constant, the log-likelihood is:

$$\begin{aligned} \log L &= (-N_1/2)\log\sigma^2 - (1/2\sigma^2)\sum_1(y_i - \boldsymbol{\beta}'\boldsymbol{x}_i)^2 \\ &\quad + \sum_0 \log[1 - \Phi(\boldsymbol{\beta}'\boldsymbol{x}_i/\sigma)]. \end{aligned} \tag{2.6}$$

Since the Hessian of the corresponding score estimating functions in the Tobit model is not necessarily negative definite, it suggests that the score estimating equation could have multiple roots. However, when the model is reparameterized in terms of $\boldsymbol{\gamma} = \boldsymbol{\beta}/\sigma$ and $\theta = 1/\sigma$ (Olsen, 1978), (2.6) becomes

$$\begin{aligned} \log L &= (N_1/2)\log\theta^2 - \sum_1(\theta y_i - \boldsymbol{\gamma}'\boldsymbol{x}_i)^2 \\ &\quad + \sum_0 \log[1 - \Phi(\boldsymbol{\gamma}'\boldsymbol{x}_i)] \end{aligned}$$

The Hessian matrix in terms of $\boldsymbol{\gamma}$ and θ is negative definite. So there is a unique root. That is, nonnegative definiteness does not necessarily imply multiple solutions.

2.2.4 Counting Processes

The multiple root problem also arises in some counting processes in reliability and survival analysis. We assume that a computer program contains a finite but unknown number of N faults initially. Let $n(t)$ denote the number of faults detected at time t. If testing could go on indefinitely, all the N faults would be detected with probability one. Let T_i, $i = 1, 2, \cdots, N$ be the failure times. In the model,

introduced by Littlewood (1980), it is assumed that at any time, the failure rate is proportional to the number of remaining errors. For the Littlewood model with intensity

$$\lambda(t) = \frac{\alpha(N - n(t-))}{1 + \epsilon t} \qquad t \in [0, \tau]$$

the loglikelihood function is:

$$\log L_\tau(N, \alpha, \epsilon) = n(\tau)\log(\alpha) - \alpha(N - n(\tau))\frac{\log(1+\epsilon\eta)}{\epsilon} + \sum_{i=1}^{n(\tau)} \log(N - i + 1) - (\alpha + \epsilon)\sum_{i=1}^{n(\tau)} \frac{\log(1 + \epsilon T_i)}{\epsilon}. \tag{2.7}$$

Barendregt and Van Pul (1995) showed that for parameter $\boldsymbol{\theta} = (N, \alpha, \epsilon)$, the corresponding system of the score estimating equations may have more than one solution. In their paper, they presented the following data set: $\tau = 709.5$, $n(\tau) = 3$, $T_1 = 1$, $T_2 = 399.9$, and $T_3 = 400.1$, and found that for this data set, $\log L_\tau(N, \alpha, \epsilon)$ has a global maximum at the boundary ($N = n(\tau)$ and $\epsilon = 0$) and both a local maximum and saddle-point in the interior of the parameter set.

2.2.5 Functional Normal Regression Models

Stefanski and Carroll (1987) considered functional normal and logistic regression models. Assume that

$$y_i \sim N(\alpha + \boldsymbol{\beta}'\boldsymbol{u}_i, \sigma^2)$$

and $\boldsymbol{u}_i$ cannot be observed but independent measurements $\boldsymbol{x}_i$ is available and $\text{var}(\boldsymbol{x}|\boldsymbol{u}) = \Omega\sigma^2$ where Ω is known. Stefanski and Carroll (1987) derived the estimating function for $\boldsymbol{\beta}$ as

$$G(\boldsymbol{\beta}) = -\boldsymbol{\beta}' S_{xy}\Omega\boldsymbol{\beta} + (S_{yy}\Omega - S_{xx})\boldsymbol{\beta} + S_{xy} \tag{2.8}$$

where

$$S_{xx} = \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^2$$

$$S_{xy} = \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(y_i - \bar{y})$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

This is a quadratic equation, so there are two roots. The corresponding estimate of α is given by $\bar{y} - \hat{\boldsymbol{\beta}}' \bar{\boldsymbol{x}}$. Multiple roots also appear in functional logistical regression models, which will be investigated in Chapter 4.

2.2.6 Mixture Models

Mixture models are widely used in biostatistics. Consider a practical example about plasma glucose distribution in south Pacific populations, which was studied by Raper, L.R. *et. al.* (1983). Surveys were carried out in August and September, 1978, and January, 1982 in Western Samoa and Nauru respectively. In both surveys, subjects were asked to fast and then present themselves to the Survey Center by 8.00 a.m. A fasting blood sample was taken and a 75g oral glucose load was administered. Two hours after the glucose load, a further blood sample was taken. The following group data presents the observed values for log two-hours plasma glucose (PG) concentration (mg/100 ml) in a sample of 89 Western Samoa females, aged 45-54.

The distribution for the logarithms of 2-h PG concentration is assumed to be a bimodal normal with the following mixture model:

$$f(x) = \frac{\alpha}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + \frac{1-\alpha}{\sigma_2\sqrt{2\pi}} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}. \tag{2.9}$$

Table 2.1: Log 2-hours Plasma Glucose Concentration

Log PG	Number	Log PG	Number
(50, 59.46)	1	(168.18,200)	5
(59.46,70.71)	4	(200,237.84)	2
(70.71,84.09)	9	(237.84,282.84)	3
(84.09,100)	18	(282.84.336.36)	3
(100, 118.92)	26	(336.36,400)	1
(118.92,141.42)	15	(400,475.68)	1
(141.42,168.18)	1		

In other words, the distribution function is:

$$F(x) = \alpha\Phi(\frac{x-\mu_1}{\sigma_1}) + (1-\alpha)\Phi(\frac{x-\mu_2}{\sigma_2}) \tag{2.10}$$

where $\Phi(x)$ is the standard normal distribution function. For group data, the likelihood function is:

$$L(\boldsymbol{\theta}, \boldsymbol{x}) \propto \prod_{i=1}^{k}(F(x_i) - F(x_{i-1}))^{n_i} \tag{2.11}$$

where n_i is the number of observations in $(x_{i-1}, x_i]$. For the plasma glucose data, we obtained the estimates of parameters for the above unrestricted model by using a gradient search in Matlab:

$$\hat{\alpha} = 0.8003$$
$$\hat{\mu}_1 = 102.0206 \quad \hat{\mu}_2 = 234.6601$$
$$\hat{\sigma}_1 = 19.3099 \quad \hat{\sigma}_2 = 84.6602.$$

If α in (2.9) is known, and is $\alpha = 0.80$, say, two local maxima can be found. That is, the two roots of the corresponding score estimating function are:

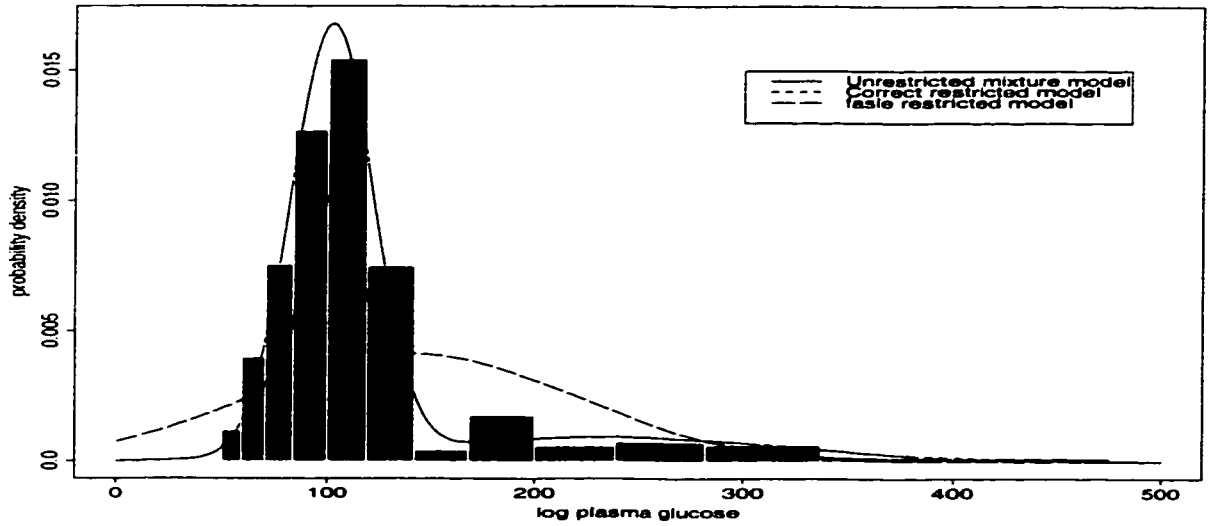

Figure 2.1: Mixture Model for Plasma Glucose Data

$$\hat{\boldsymbol{\theta}}_1 = (\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2) = (102.0196, 234.5956, 19.3082, 84.6888)$$
$$\hat{\boldsymbol{\theta}}_2 = (\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2) = (141.8957, 101.6242, 77.4806, 12.0523).$$

Since the loglikelihood for $\hat{\boldsymbol{\theta}}_1$ is -184.6139, the loglikelihood for $\hat{\boldsymbol{\theta}}_2$ is -212.8364, $\hat{\boldsymbol{\theta}}_1$ is the better estimator. Figure 2.1 also shows that $\hat{\boldsymbol{\theta}}_1$ (corresponding to the correct restricted model) is much better than $\hat{\boldsymbol{\theta}}_2$ (corresponding to the false restricted model).

In Chapter 4, we will investigate a more complex practical example which involves in a bivariate normal mixture. In this example, we will consider a real data set taken from Habbema, Hermans and van den Broek (1974).

2.2.7 An Application in Finance

Wedderburn's quasi-score estimating function is only based on the assumption concerning a relationship between the first two moments of the distribution. Godambe

and Thompson (1989) extended it by using the knowledge of the skewness and kurtosis of the distribution. From the first two moment conditions, we have two basic estimating functions

$$h_1 = X - \mu$$
$$h_2 = (X-\mu)^2 - \sigma^2$$

where $\mu = E(X)$ and $\sigma^2 = Var(X)$, then we can construct an orthogonal estimating function to h_1:

$$h_3 = (X-\mu)^2 - \sigma^2 - \gamma_1\sigma(X-\mu) \tag{2.12}$$

where $\gamma_1 = E(X-\mu)^3/\sigma^3$. Following Godambe and Thompson (1989), for an unknown parameter θ, we may obtain the following optimal linear combination of estimating functions h_1 and h_3,

$$l^* = \alpha^* h_1 + \beta^* h_3 \tag{2.13}$$

where

$$\alpha^* = \frac{E(\frac{\partial h_1}{\partial \theta})}{E(h_1^2)}$$
$$\beta^* = \frac{E(\frac{\partial h_3}{\partial \theta})}{E(h_3^2)}$$

Since it is a quadratic estimating function in μ, it has multiple roots. In particular, when $\theta = \mu$,

$$\alpha^* = \frac{E(\frac{\partial h_1}{\partial \mu})}{E(h_1^2)} = -\frac{1}{\sigma^2} \tag{2.14}$$
$$\beta^* = \frac{E(\frac{\partial h_3}{\partial \mu})}{E(h_3^2)} = \frac{\gamma_1}{\sigma^3(\gamma_2 + 2 - \gamma_1^2)}$$

where $\gamma_2 = E(X-\mu)^4/\sigma^4 - 3$.

X. Li (1999) applied this approach to calculating Value at Risk in which skewness, kurtosis and volatility are explicitly used. Nowadays, Value at Risk(VaR) is a popular risk management method. For a given time horizon t and confidence level p, the value at risk is the loss in the market value over the time horizon t that is exceeded with probability $1-p$. Thus, it is our interest to estimate the mean of the market value and construct an approximate interval for it based on estimating function (2.13). Many observed financial return series have tails that are "fatter" than those implied by a normal distribution, the VaR calculated under the normal assumption underestimates the actual risk. Without the normality assumption, some parametric and nonparametric approaches were proposed, such as a mixture of two normal distributions (Hull and White, 1998), order statistics method or Monte Carlo simulation. Since $l^*/\sqrt{Var(l^*)}$ has a standard normal distribution approximately, then we can construct an approximate confidence interval. This method is applied to the daily exchange rates for 12 major currencies between February 17, 1989 and February 8, 1999. In this case, the financial return series are the daily logarithm change $X_t = \ln(S_t/S_{t-1})$, where S_t is the spot exchange rate at time t. This study shows that the estimating function approach captures the extreme tail much better than the standard VaR calculation method such as RiskMetrics approach.

2.3 Methodologies for Root Selection

In this section, I will review the existing methods, describe some potential methods, and propose our new methods in discriminating among multiple roots of estimating functions. These methods include *projected likelihood ratios*, *one-root estimating functions*, *statistical information methods* based on consistence, chi-square

criterion, and the methods based on the root intensity function and the shifted information.

2.3.1 Projected Likelihood Ratios

When an estimating function has multiple roots, an approximate estimating function can be constructed with a certain property, then the best root is chosen among several roots of the estimating function based on these properties.

Let us begin with the Wedderburn's quasi-score estimating function

$$\boldsymbol{q}(\boldsymbol{\theta}, \boldsymbol{x}) = (\dot{\boldsymbol{\mu}}(\boldsymbol{\theta}))'(V(\boldsymbol{\theta}))^{-1}(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta})) = 0 \tag{2.15}$$

where $\boldsymbol{\mu}$ is the $n \times 1$ vector of mean responses, $\boldsymbol{\theta}$ is a $k \times 1$ vector of regression parameters relating to explanatory variables to $\boldsymbol{\mu}$, and $V(\boldsymbol{\theta})$ denotes the $n \times n$ variance matrix of the response. By analogy with the relationship between the score function and the log-likelihood function, a line integral can be performed to obtain the quasi-likelihood ratio,

$$Q(\boldsymbol{\theta}, \boldsymbol{\eta}) = \int_{\boldsymbol{\eta}}^{\boldsymbol{\theta}} \boldsymbol{q}(\boldsymbol{t}, \boldsymbol{x}) \cdot d\boldsymbol{t}. \tag{2.16}$$

However, it should be noted that this integral is path-dependent when $k > 1$, unless there is a potential function U such that $\partial U/\partial \theta_s = q_s$, $s = 1, 2, \ldots, k$. That is, the derivative matrix for $\boldsymbol{q}(\boldsymbol{\theta}, \boldsymbol{x})$ is symmetric provided $\boldsymbol{q}(\boldsymbol{\theta}, \boldsymbol{x})$ is continuously differentiable. To overcome this difficulty, McLeish and Small (1992) proposed the *projected likelihood ratio*, which projects $\lambda(\boldsymbol{\theta}, \boldsymbol{\eta}) = L(\boldsymbol{\eta})/L(\boldsymbol{\theta})$ for independent observations with mean $\mu_i(\boldsymbol{\theta})$ and variance $\sigma_i^2(\boldsymbol{\theta})$ $(i = 1, 2, \ldots, n)$ into a class of estimating functions. This class is the largest class of estimating functions which can be determined by the mean and variance of x_i. By a direct calculation, they

obtained the projected likelihood ratio

$$\hat{\lambda}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \prod_{i=1}^{n}[1 + \sigma_i^{-2}(\boldsymbol{\theta})\{\mu_i(\boldsymbol{\eta}) - \mu_i(\boldsymbol{\theta})\}\{x_i - \mu_i(\boldsymbol{\theta})\}] \tag{2.17}$$

which is tangent to the quasi-likelihood ratio function at $\boldsymbol{\theta}$. Later, B. Li (1993) suggested the following *linear projected likelihood ratio* $R(\boldsymbol{\eta}, \boldsymbol{\theta})$ as an alternative:

$$\frac{1}{2}(\boldsymbol{\mu}(\boldsymbol{\eta}) - \boldsymbol{\mu}(\boldsymbol{\theta}))^t\{(V(\boldsymbol{\theta}))^{-1}(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta})) + (V(\boldsymbol{\eta}))^{-1}(\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\eta}))\} \tag{2.18}$$

which is anti-symmetric and linear in observation. As the *projected likelihood ratio*, it is also the first order approximation of the quasi-likelihood ratio. Furthermore, under certain conditions, as $n \to \infty$, for all $\boldsymbol{\eta}$ in $\boldsymbol{\Theta}$ and $\boldsymbol{\eta} \neq \boldsymbol{\theta}_0$,

$$\begin{aligned} &P\{R(\boldsymbol{\eta}, \boldsymbol{\theta}_0, \boldsymbol{x}) < R(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0, \boldsymbol{x}); \boldsymbol{\theta}_0\} \to 1 \\ &P\{R(\boldsymbol{\theta}_0, \boldsymbol{\eta}, \boldsymbol{x}) > R(\boldsymbol{\eta}, \boldsymbol{\eta}, \boldsymbol{x}); \boldsymbol{\theta}_0\} \to 1. \end{aligned} \tag{2.19}$$

This result indicates that for any inference point, no matter whether the reference point is at the true parameter value $\boldsymbol{\theta}_0$ or the wrong value,

$$P\{R(\boldsymbol{\theta}_0, \text{inference points}, \boldsymbol{x}) > R(\boldsymbol{\eta}, \text{inference points}, \boldsymbol{x}); \boldsymbol{\theta}_0\} \to 1. \tag{2.20}$$

Since the linear projected likelihood ratio is a good approximation of quasi-likehood functions, which may be path-dependent, it is reasonable to use the linear projected likelihood ratio to distinguish the roots of the Wedderburn's quasi-score estimating functions. B. Li (1993) discussed the circle model of McCullagh (1990), which assumed that x_{1i}, x_{2i} $(i = 1, \ldots, n)$ are independent random variables satisfying $\mu_\theta = (\cos\theta, \sin\theta)$ and $V_\theta = n^{-1}I$. The corresponding Wedderburn's quasi-score function is

$$q(\theta, \boldsymbol{x}) = -n\sin\theta(\bar{x}_1 - \cos\theta) + n\cos\theta(\bar{x}_2 - \sin\theta). \tag{2.21}$$

There are two roots $\hat{\theta}_0 = \tan^{-1}(\bar{x}_2/\bar{x}_1)$, $\hat{\theta}_1 = \hat{\theta}_0 + \pi$. In this case, both functions $R(\eta, \hat{\theta}_0, \boldsymbol{x})$ and $R(\eta, \hat{\theta}_1, \boldsymbol{x})$ take the maximum at $\eta = \hat{\theta}_0$ and the minimum at $\eta = \hat{\theta}_1$. Naturally, $\hat{\theta}_0$ is picked as the better estimator. In general, when there is a root $\hat{\boldsymbol{\theta}}_0$ (for example, the one close to the real parameter) such that for all other roots $\hat{\boldsymbol{\theta}}_i$ $(i = 1, 2, \cdots, k)$, $R(\hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_i, \boldsymbol{x}) > R(\hat{\boldsymbol{\theta}}_j, \hat{\boldsymbol{\theta}}_i, \boldsymbol{x})$ $(i = 0, 1, \cdots, k,\ j = 1, 2, \cdots, k)$ hold with probability close to 1 as the sample size n is large, this method can work. However, when two roots $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ of an estimating function are found with properties: $R(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2, \boldsymbol{x}) > R(\hat{\boldsymbol{\theta}}_2, \hat{\boldsymbol{\theta}}_2, \boldsymbol{x}) = 0$ and $R(\hat{\boldsymbol{\theta}}_2, \hat{\boldsymbol{\theta}}_1, \boldsymbol{x}) > R(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_1, \boldsymbol{x}) = 0$, The question of which root to choose arises.

Hanfelt and Liang (1995) considered the approximate likelihood ratios in the more general case. The elementary estimating function is assumed to be $\boldsymbol{g}(\boldsymbol{\psi}, \boldsymbol{x})$, where $\boldsymbol{g}$ is a vector of length k such that $\mathrm{E}_{\boldsymbol{\psi}}\{\boldsymbol{g}(\boldsymbol{\psi}, \boldsymbol{x})\} = 0$ for all $\boldsymbol{\psi}$. The optimal estimating function in the class of the weighted estimating functions $(\boldsymbol{a}(\boldsymbol{\psi}))'\boldsymbol{g}(\boldsymbol{\psi}, \boldsymbol{x})$ is

$$\boldsymbol{q}(\boldsymbol{\psi}, \boldsymbol{x}) = D_{\boldsymbol{\psi}}^t V_{\boldsymbol{\psi}}^{-1} \boldsymbol{g}(\boldsymbol{\psi}, \boldsymbol{x})$$

where $D_{\boldsymbol{\psi}} = \mathrm{E}_{\boldsymbol{\psi}}(-\partial \boldsymbol{g}/\partial \boldsymbol{\psi})$ and $V_{\boldsymbol{\psi}} = \mathrm{var}_{\boldsymbol{\psi}}(\boldsymbol{g})$. When the elementary function is $\boldsymbol{x} - \boldsymbol{\mu}(\boldsymbol{\theta})$, it is the Wedderburn's quasi-score estimating function. The *general linear projected likelihood ratio* $R(\boldsymbol{\theta}, \boldsymbol{\eta})$ is defined as

$$\frac{1}{2} C(\boldsymbol{\theta}, \boldsymbol{\eta})^t (V(\boldsymbol{\eta}))^{-1} \boldsymbol{g}(\boldsymbol{\psi}(\boldsymbol{\eta}), \boldsymbol{x}) - \frac{1}{2} C(\boldsymbol{\eta}, \boldsymbol{\theta})^t (V(\boldsymbol{\theta}))^{-1} \boldsymbol{g}(\boldsymbol{\psi}(\boldsymbol{\theta}), \boldsymbol{x}) \qquad (2.22)$$

where $C(\boldsymbol{\theta}, \boldsymbol{\eta}) = \mathrm{E}_{\boldsymbol{\theta}}\{\boldsymbol{g}(\boldsymbol{\psi}(\boldsymbol{\eta}), \boldsymbol{x})\}$, $V(\boldsymbol{\eta}) = \mathrm{var}_{\boldsymbol{\eta}}\{\boldsymbol{g}(\boldsymbol{\psi}(\boldsymbol{\eta}), \boldsymbol{x})\}$. Hanfelt and Liang (1995) further considered the case with a nuisance parameter. They also discussed the generalized quasi-likelihood ratio which is path-dependent and which can also be used to discriminate the multiple roots. The *general linear projected likelihood ratio* has similar properties as the *linear projected likelihood ratio*, so it can used in discriminating among multiple roots as well.

2.3.2 Minimax Approach to Estimating Functions

Doob (1934) and Wald (1949) demonstrated that the global maximum of the likelihood function is consistent under certain regularity conditions. Yet, there is no analogy to the likelihood function for a general estimating function. For the linear projected likelihood ratio $R(\boldsymbol{\theta}, \boldsymbol{\eta})$ defined in (2.18), B. Li (1996) has shown that a consistent estimator $\hat{\boldsymbol{\theta}}$ for the Wedderburn's quasi-score estimating function can be identified as the minimax points of $R(\boldsymbol{\theta}, \boldsymbol{\eta})$ under mild conditions. That is, $\sup_{\eta}\{R(\hat{\boldsymbol{\theta}}, \boldsymbol{\eta})\} = \inf_{\theta}\sup_{\eta}\{R(\boldsymbol{\theta}, \boldsymbol{\eta})\}$. Later, B. Li (1997) extended this result to the generalized estimating equation(GEE), which is widely used for longitudinal data analysis in biostatistics. The longitudinal data set usually comprises an outcome variable, y_{it}, and a $p \times 1$ vector of covariates, $\boldsymbol{x}_{it}$, observed at time $t = 1, \ldots, n_i$ for subjects $i = 1, \ldots, K$. Let $\mu_{it}(\boldsymbol{\beta})$ and $\phi^{-1}V_{it}(\boldsymbol{\beta})$ be the mean and variance of the observation y_{it}. We will assume that $\mu_{it}(\boldsymbol{\beta}) = \mu(\boldsymbol{x}_{it}^t\boldsymbol{\beta})$, and $V_{it}(\boldsymbol{\beta}) = V(\boldsymbol{x}_{it}^t\boldsymbol{\beta})$ for some known functions $\mu(.)$ and $V(.)$ as in the generalized linear models. The dispersion parameter ϕ is always taken to be positive. The dependence within each i is modeled by the correlation matrix $R(\boldsymbol{\alpha}) = \{R_{tt'}(\boldsymbol{\alpha}) : t, t' = 1, \ldots, n_i\}$. Across i, $R_{tt'}(\boldsymbol{\alpha})$ is assumed to remain constant as long as observations y_{it} and $y_{it'}$ are present in the cluster i. That is, $R(\boldsymbol{\alpha})$ is independent of cluster except for its dimension. Let $V_i(\boldsymbol{\beta}) = \mathrm{diag}(V_{i1}(\boldsymbol{\beta}), \ldots, V_{in_i}(\boldsymbol{\beta}))^t$. We refer to $R(\boldsymbol{\alpha})$ as a 'working' correlation matrix. We also call

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

a 'working' covariance matrix, which will be equal to $\mathrm{cov}(\boldsymbol{y}_i)$ if $R(\boldsymbol{\alpha})$ is the true correlation matrix for $\boldsymbol{y}_i$'s, where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^t$. If the working correlation assumption were correct, then for each fixed ϕ and $\boldsymbol{\alpha}$, the optimal linear combination of $\{y_{it} - \mu_{it}(\boldsymbol{\beta})\}$ which yields the highest information about $\boldsymbol{\beta}$ in the sense of

Godambe (1960), is

$$\boldsymbol{q}(\boldsymbol{\beta},\phi,\boldsymbol{\alpha}) = \sum_{i=1}^{K}\{\dot{\boldsymbol{\mu}}_i(\boldsymbol{\beta})\}^t\{W_i(\boldsymbol{\beta},\phi,\boldsymbol{\alpha})\}^{-1}\{\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\} = 0 \tag{2.23}$$

where $\boldsymbol{\mu}_i(\boldsymbol{\beta})$ is the $n_i \times 1$ vector $\{\mu_{it}(\boldsymbol{\beta}) : t = 1,\ldots,n_i\}$, $\dot{\boldsymbol{\mu}}_i(\boldsymbol{\beta})$ is the $n_i \times p$ dimensional gradient matrix of $\boldsymbol{\mu}_i(\boldsymbol{\beta})$. The above equation can be re-written as a function of $\boldsymbol{\beta}$ alone by replacing ϕ with a $\sqrt{K}$-consistent estimate $\hat{\phi}(\boldsymbol{\beta})$, and $\boldsymbol{\alpha}$ with a $\sqrt{K}$-consistent estimate $\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta},\hat{\phi}(\boldsymbol{\beta}))$. That is,

$$\boldsymbol{g}(\boldsymbol{\beta}) \equiv \boldsymbol{q}\{\boldsymbol{\beta},\hat{\phi}(\boldsymbol{\beta}),\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta},\hat{\phi}(\boldsymbol{\beta}))\} = 0 \tag{2.24}$$

which is called the *generalized estimating equation* according to Liang and Zeger (1986). Though a generalized estimating equation has a consistent solution with the probability tending to one under regularity conditions, it remains to show how to identify a specific sequence of solutions which is consistent, since it may either have multiple solutions or have none at all in many applications. B. Li (1997) proposed a minimax procedure which yields a consistent estimate. In order to state this result, we need an analogy to the linear projected likelihood ratio for the generalized estimating equation.

Definition 2.1 Let $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1,\phi_1,\boldsymbol{\alpha}_1^t)^t$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\beta}_2,\phi_2,\boldsymbol{\alpha}_2^t)^t$ be two points in the parameter space $\boldsymbol{\Theta}$. The parameter space for $\boldsymbol{\beta}$ will be written as B. Let $\mathcal{D} : \boldsymbol{\Theta}\times\boldsymbol{\Theta}\times\mathcal{Y} \to R^1$ be the function

$$\begin{aligned}\mathcal{D}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) &= \sum_{i=1}^{K}\{\boldsymbol{\mu}_i(\boldsymbol{\beta}_2) - \boldsymbol{\mu}_i(\boldsymbol{\beta}_1)\}^t \\ &\quad \{\frac{\phi_1}{2}W_i^{-1}(\boldsymbol{\theta}_1)[\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_1)] + \frac{\phi_2}{2}W_i^{-1}(\boldsymbol{\theta}_2)[\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_2)]\}\end{aligned} \tag{2.25}$$

The *deviance function* of the generalized estimating equation (2.24), is a mapping $R : B \times B \times \mathcal{Y} \to R^1$ defined by

$$R(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \mathcal{D}\{\boldsymbol{\beta}_1, \hat{\phi}(\boldsymbol{\beta}_1), \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}_1, \hat{\phi}(\boldsymbol{\beta}_1)); \boldsymbol{\beta}_2, \hat{\phi}(\boldsymbol{\beta}_2), \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}_2, \hat{\phi}(\boldsymbol{\beta}_2))\} \tag{2.26}$$

The *centering function* $J(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ of $R(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ is the mapping $J : B \times B \to R^1$ defined by

$$\mathrm{E}_{\boldsymbol{\beta}_0} \mathcal{D}\{\boldsymbol{\beta}_1, \mathrm{E}\hat{\phi}(\boldsymbol{\beta}_1), \mathrm{E}\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}_1, \hat{\phi}(\boldsymbol{\beta}_1)); \boldsymbol{\beta}_2, \mathrm{E}\hat{\phi}(\boldsymbol{\beta}_2), \mathrm{E}\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}_2, \hat{\phi}(\boldsymbol{\beta}_2))\}. \tag{2.27}$$

Obviously, $J(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = -J(\boldsymbol{\beta}_2, \boldsymbol{\beta}_1)$. Since $E_{\boldsymbol{\beta}_0}(\boldsymbol{y}_i) = \boldsymbol{\mu}_i(\boldsymbol{\beta}_0)$, $J(\boldsymbol{\beta}_0, \boldsymbol{\beta})$ is the negative quadratic form:

$$-\sum_{i=1}^{K} \{\boldsymbol{\mu}_i(\boldsymbol{\beta}) - \boldsymbol{\mu}_i(\boldsymbol{\beta}_0)\}^t W_i^{-1} \{\boldsymbol{\beta}, E\hat{\phi}(\boldsymbol{\beta}), E\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, E\hat{\phi}(\boldsymbol{\beta}))\} \{\boldsymbol{\mu}_i(\boldsymbol{\beta}) - \boldsymbol{\mu}_i(\boldsymbol{\beta}_0)\}$$

When $n = n_i$, $(i = 1, 2, \cdots, K)$, and some regularity conditions hold (B. Li, 1997), then any parameter value $\hat{\boldsymbol{\beta}}$ that satisfies the relation

$$\sup_{\boldsymbol{\beta} \in B} R(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \inf_{\boldsymbol{\beta} \in B} \sup_{\boldsymbol{\beta}' \in B} R(\boldsymbol{\beta}, \boldsymbol{\beta}') \tag{2.28}$$

is a consistent estimate of $\boldsymbol{\beta}_0$. Furthermore, when

$$\inf_{\boldsymbol{\beta} \in B} \sup_{\boldsymbol{\beta}' \in B} R(\boldsymbol{\beta}, \boldsymbol{\beta}') = \sup_{\boldsymbol{\beta}' \in B} \inf_{\boldsymbol{\beta} \in B} R(\boldsymbol{\beta}, \boldsymbol{\beta}') \tag{2.29}$$

then any solution $\tilde{\boldsymbol{\beta}}$ of equation (2.24) that satisfies

$$\sup_{\boldsymbol{\beta} \in \hat{B}} R(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}) = 0 \tag{2.30}$$

is consistent, where $\hat{B}$ is the set of all solutions to the generalized estimating equation (2.24).

The above results suggest a method to choose a consistent solution for the generalized estimating equation, since (2.30) indicates that we need to compare the roots of the generalized estimating equations (2.24) only. However, the conditions (2.28) and (2.29) are not generally true and cannot easily be checked. Improving these conditions and finding alternative conditions required further work.

2.3.3 Approximate One-root Estimating Functions

In the last subsection, we discuss the approximate estimating function which can used to compare the different roots. From other viewpoint, Kolkiewicz (1995) and McLeish and Small (1988) considered the approximate one-root estimating function. In other words, they built a new estimating function which has only one root, as an approximation to the original estimating function in some senses. This root is closely related to the roots of the original estimating function; so it can be taken as an estimator of the parameter.

Projection to monotone functions: Kolkiewicz (1995) considered the projection of score functions into a set of monotone functions, which have only one root. This provides an alternative projection for the estimating functions. He discussed this case in which the score estimating function $s(x)$ is in the class $\mathcal{G}$ with the following property: there exists points m_l and m_r such that on the interval $[m_l, m_r]$ the function is nonincreasing, on the interval $(-\infty, m_l) \cup (m_r, \infty)$ it is nondecreasing and

$$\lim_{x \to -\infty} s(x, \theta) \geq \lim_{x \to \infty} s(x; \theta). \tag{2.31}$$

The subclass $\mathcal{G}_s$ consists of all continuous non-increasing functions on $[x_l, x_r]$ which are constant outside $[x_l, x_r]$. The projection into $\mathcal{G}_s$ from $\mathcal{G}$ is fully determined by

only points x_l and x_r which satisfy $m_l \leq x_l \leq m_l'$ and $m_r' \leq x_r \leq m_r$, where m_l' and m_r' are such that

$$s(m_l') = \lim_{x \to -\infty} s(x)$$
$$s(m_r') = \lim_{x \to \infty} s(x)$$

For a location parameter model with density function $f_\theta(x) = f_0(x - \theta)$, by using Lagrange's method of multipliers, it can be found that x_l and x_r uniquely satisfy the following equations respectively:

$$\int_{-\infty}^{x_l} s_0(x_l) f_0(x) dx = s_0(x_l) \int_{-\infty}^{x_l} f_0(x) dx \tag{2.32}$$
$$\int_{x_r}^{\infty} s_0(x) f_0(x) dx = s_0(x_r) \int_{x_r}^{\infty} f_0(x) dx \tag{2.33}$$

where $s_0(x)$ is the score function when $\theta = 0$. When $f_0(x)$ is symmetric, $x_l = -x_r$. By projecting $s(x_i, \theta) = s_0(x_i - \theta)$ into $\mathcal{G}_s$, the score function $S(x_1, \ldots, x_n; \theta) = \sum_{i=1}^n s(x_i, \theta)$ can be projected onto a function in $\mathcal{G}_s$, which yields one root only. It is easily checked that the score estimating function for the symmetric Student's t-distribution with ν degree of freedom (including Cauchy distribution) with a location parameter is in $\mathcal{G}$. The loss of efficiency of the projected score estimating function is maximal for the Cauchy distribution (less than 13%) and decreases quickly as ν becomes larger.

Reducing the number of roots: In order to rule out the incorrect roots for the score estimating functions, McLeish and Small (1988) suggested a method to reduce the number of roots. Suppose $L(\theta, \boldsymbol{x})$ is a likelihood function defined on $(-\infty, +\infty)$ that is continuously differentiable with a finite number of local maxima and minima, and $L(\theta, \boldsymbol{x})$ vanishes at infinity. Let

$$\psi_0(\theta, \boldsymbol{x}) = S(\theta, \boldsymbol{x}) \tag{2.34}$$
$$\psi_\epsilon(\theta, \boldsymbol{x}) = \frac{L(\theta + \epsilon, \boldsymbol{x}) - L(\theta - \epsilon, \boldsymbol{x})}{2\epsilon L(\theta, \boldsymbol{x})}$$

where $S(\theta, \boldsymbol{x})$ is the score function. It should be noticed that this is a particular case of Daniel's *smoothed likelihood* (1960) . The smoothed likelihood, with kernel $u(y)$ is defined as

$$\bar{l}(\theta) = \int_{-\infty}^{+\infty} l(\theta - y)u(y)dy \tag{2.35}$$

where $l(\theta) = \log L(\theta)$. The parameter value $\bar{\theta}$ which maximizes $\bar{l}(\theta)$ is called a *smoothed maximum likelihood estimator*. Assume that we can interchange derivatives and integrals, $\bar{\theta}$ can be written as the root of the smoothed score:

$$\bar{g}(\theta) = \int_{-\infty}^{+\infty} l'(\theta - y)u(y)dy. \tag{2.36}$$

When $u(y) = 1/2\epsilon$ for $|y| \leq \epsilon$ and $u(y) = 0$ for $|y| > \epsilon$, $\bar{\theta}$ will be a solution to the equation $L(\theta + \epsilon) - L(\theta - \epsilon) = 0$. This is the same as (2.34).

It can be shown (see below) that when ϵ is large enough, ψ_ϵ has only one root. Thus ψ_ϵ provides a straightforward estimate in small samples. In order to extend this method to a more general class, which provides more flexibility to choose coefficients, we consider the following estimating function in the form of:

$$\tilde{G}(\theta, \boldsymbol{x}) = \sum_{i=1}^{m} \frac{\lambda_i[L(\theta + \epsilon_i, \boldsymbol{x}) - L(\theta, \boldsymbol{x})]}{L(\theta, \boldsymbol{x})}. \tag{2.37}$$

Assume that $\sum_{i=1}^{m} \lambda_i = 0$ for unbiasedness of $\tilde{G}(\theta, \boldsymbol{x})$ since $\int L(\theta, \boldsymbol{x})d\boldsymbol{x} = 1$ for all θ. (2.34) is obviously the special case of (2.37) when $\epsilon_1 = -\epsilon_2 = \epsilon$, $\lambda_1 = -\lambda_2 = 1/2\epsilon$. Under the above assumptions about $L(\theta, \boldsymbol{x})$ and the condition $\lambda_i\epsilon_i > 0$ for all $i = 1, \ldots, m$, there is only one root of $\tilde{G}(\theta, \boldsymbol{x}) = 0$ when $|\epsilon_i|$ $(i = 1, \ldots, m)$ is large enough. In fact, since $L(\theta, \boldsymbol{x})$ has only a finite number of local maxima and minima and vanishes at infinity, we can conclude that there is a sufficient large number M such that $L(\theta, \boldsymbol{x})$ decreases in θ on $[M, +\infty]$ and increases on $[-\infty, -M]$. Assume that θ_1 and θ_2 are two roots of $\tilde{G}(\theta, \boldsymbol{x}) = 0$, that is:

$$\sum_{i=1}^{m} \lambda_i[L(\theta_l + \epsilon_i, \boldsymbol{x}) - L(\theta_l, \boldsymbol{x})] = 0 \tag{2.38}$$

for $l = 1, 2$. Without loss of generality, we assume $\theta_1 < \theta_2$. According to the above discussion, when $|\epsilon_i|$ is large enough, we have:

$$L(\theta_1 + \epsilon_i, \boldsymbol{x}) > L(\theta_2 + \epsilon_i, \boldsymbol{x}) \qquad \epsilon_i > 0 \tag{2.39}$$

$$L(\theta_1 + \epsilon_i, \boldsymbol{x}) < L(\theta_2 + \epsilon_i, \boldsymbol{x}) \qquad \epsilon_i < 0 \tag{2.40}$$

Thus, when $\lambda_i \epsilon_i > 0$ for $i = 1, \ldots, m$,

$$\begin{aligned} & \sum_{i=1}^{m} \lambda_i [L(\theta_1 + \epsilon_i, \boldsymbol{x}) - L(\theta_1, \boldsymbol{x})] \\ = & \sum_{i=1}^{m} \lambda_i L(\theta_1 + \epsilon_i, \boldsymbol{x}) \\ > & \sum_{i=1}^{m} \lambda_i L(\theta_2 + \epsilon_i, \boldsymbol{x}) - (\sum_{i=1}^{m} \lambda_i) L(\theta_2, \boldsymbol{x}) \\ = & \sum_{i=1}^{m} \lambda_i [L(\theta_2 + \epsilon_i, \boldsymbol{x}) - L(\theta_2, \boldsymbol{x})] \end{aligned}$$

This contradicts the (2.38). Thus $\tilde{G}(\theta, \boldsymbol{x}) = 0$ has at most one root. Now we will extend the method to more general estimating functions. For an estimating function $G(\theta, \boldsymbol{x})$ which is continuous in θ, we define

$$L(\theta, \boldsymbol{x}) = \int_0^{\theta} G(\tau, \boldsymbol{x}) d\tau \tag{2.41}$$

Then under the above conditions imposed on λ_i and ϵ_i and

$$G(\theta, \boldsymbol{x}) G(-\theta, \boldsymbol{x}) < 0 \quad \text{if } \theta \text{ large} \tag{2.42}$$

which ensures that $L(\theta, \boldsymbol{x})$ is monotone when $|\theta|$ is large. Thus $\tilde{G}(\theta, \boldsymbol{x})$ defined as in (2.37) has at most one root when $|\epsilon_i|$ is large enough. Since $\log(\theta)$ is a strictly monotone function, we can use the following alternative estimating function

$$\tilde{G}^*(\theta, \boldsymbol{x}) = \sum_{i=1}^{m} \lambda_i [\log L(\theta + \epsilon_i, \boldsymbol{x}) - \log L(\theta, \boldsymbol{x})] \tag{2.43}$$

where $\sum \lambda_i = 0$, but it is not an unbiased estimating function.

2.3.4 Four Methods for Choosing a Root

A completely different approach is to find some simple and direct methods for root selection in the multiple root case. That is, we wish to define statistical information based on estimating function as criteria to choose the best root among the multiple roots. Heyde (1997) and Heyde and Morton (1998) suggested three methods to choose among multiple roots. Let us describe the methods and some examples discussed in their paper.

Method 1. *Examining the asymptotics and choosing the root which is consistent.* If an estimating equation has several roots, and one of them is consistent, we will choose it as the estimator of the true parameter. For example, let $x_1, \ldots, x_n$ be i.i.d with mean θ $(-\infty < \theta < \infty)$ and known variance σ^2 and consider the estimating function

$$G(\theta) = \sum_{i=1}^{n} \{a(x_i - \theta) + (x_i - \theta)^2 - \sigma^2\} \tag{2.44}$$

Let

$$\Delta = \frac{1}{4}a^2 + \bar{x}^2 - \frac{1}{n}\sum_{i=1}^{n} x_i^2 + \sigma^2$$

where $\bar{x}$ is the sample mean. Then, on the set $\{\Delta > 0\}$ the roots of $G(\theta) = 0$ are

$$\hat{\theta}_1 = \frac{1}{2}a + \bar{x} - \Delta^{1/2}$$
$$\hat{\theta}_2 = \frac{1}{2}a + \bar{x} + \Delta^{1/2}$$

and the strong law of large numbers gives $\hat{\theta}_1 \to \theta_0$, $\hat{\theta}_2 \to \theta_0 + a$, where θ_0 is the true value. Thus, we should select $\hat{\theta}_1$ as the correct root.

Method 2. *Picking the root for which $G'(\theta)$ behaves asymptotically as its expected value $E_\theta G'(\theta)$* . Let us go back to the estimation of the angle in circular

data (see Section 2.3.1), the quasi-score estimating function is

$$G(\theta) = -n\bar{x}_1 \sin(\theta) + n\bar{x}_2 \cos(\theta). \tag{2.45}$$

Then $G(\theta) = 0$ has two solutions $\hat{\theta}_1$ and $\hat{\theta}_2$ in $(-\pi/2, \pi/2)$: $\hat{\theta}_1 = \tan^{-1}(\bar{x}_2/\bar{x}_1)$, $\hat{\theta}_2 = \hat{\theta}_1 + \pi$. Note that $E_\theta G'(\theta) = -1$, and $G'(\hat{\theta}_1)/E_{\hat{\theta}_1} G'(\hat{\theta}_1) \to 1$, $G'(\hat{\theta}_2)/E_{\hat{\theta}_2} G'(\hat{\theta}_2) \to -1$. Thus we will choose $\hat{\theta}_1$. Method 1 also gives the same choice.

Method 3. *Selecting the best root based on least square or goodness of fit criterion.* Assume that $\boldsymbol{x}$ has mean $\boldsymbol{\mu}(\boldsymbol{\theta})$, and consider

$$S(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu}(\boldsymbol{\theta})'(\boldsymbol{x}_i - \boldsymbol{\mu}(\boldsymbol{\theta}))$$

or weighted least squares. By using the law of large numbers, they justified that the $S(\boldsymbol{\theta})$ at the wrong root is far more larger than that at a consistent root. When applying this method to (2.44), we find $S(\hat{\theta}_2) - S(\hat{\theta}_1) = 2a\Delta^{\frac{1}{2}} \sim a^2 n$, so we prefer $\hat{\theta}_1$ to $\hat{\theta}_2$. For (2.45), the same result can be obtained by using Method 1 and Method 2.

Recently, Singh and Mantel (1999) developed a least square criterion and proposed another method based on the theory of minimum chi-square estimation.

Method 4. *Choose the root which minimizes the square of a supplementary estimating function.* Let us consider a general estimating function of the form $G(\boldsymbol{x}, \boldsymbol{\theta}) = \sum w_i(\boldsymbol{\theta}) g_i(\boldsymbol{x}_i, \boldsymbol{\theta})$, where $g_i(\boldsymbol{x}_i, \boldsymbol{\theta})$ are elementary estimating functions based on observation $\boldsymbol{x}_i$. When $G(\boldsymbol{x}, \boldsymbol{\theta}) = 0$ gives multiple roots, we define a supplementary estimating function $G_\xi(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{n} a_i(\boldsymbol{\theta}) g_i(\boldsymbol{x}_i, \boldsymbol{\theta})$ such that $E_\theta\{G_\xi(\boldsymbol{x}, \boldsymbol{\theta})\} = 0$ and $G_\xi(\boldsymbol{x}, \hat{\boldsymbol{\theta}}) \neq 0$ at the roots $\hat{\boldsymbol{\theta}}$ of $G(\boldsymbol{x}, \boldsymbol{\theta})$. The choice of the supplementary estimating function can be based on model testing H_0: $\xi = 0$. This method chooses the root that minimizes $G_\xi^2(\boldsymbol{x}, \boldsymbol{\theta})$.

2.3.5 Proposed New Methods

The root intensity is an extension of the probability density function to the case when an estimating function has multiple roots. It can be shown that under regularity conditions, the root intensity at the true parameter tends to infinity as the sample size n approaches infinity. Utilizing the asymptotic property of root intensity, this thesis proposes a new method to select the best root. This method suggests choosing the root with the largest estimated root intensity among several roots of the estimating functions. Simulation results have shown that this is a useful method. In Chapter 4, we will discuss this method thoroughly from theory to applications.

For transformation models, in particular, location models, we will define the shifted information (see Chapter 5), then choose the root which has the largest estimated shifted information as the estimator of the parameter.

2.4 Discussion

In this chapter, we have discussed several methods to choose the root among the multiple roots of estimating functions. However no one method can be used in a general context. Note that the (extended) projected likelihood ratio is only the first order approximation of the quasi-likelihood ratio function, if we wish to obtain a good approximation, we may use a higher order approximation, which makes the solution more complex. In addition, this method can only be used in a limited case. As we have pointed out in Section 2.3.1, this method may fail when $R(\boldsymbol{\eta}, \hat{\boldsymbol{\theta}}_i, \boldsymbol{x})$ takes the maximum at different points, where $\hat{\boldsymbol{\theta}}_i$ $(i = 1, 2, \ldots, k)$ are different roots. The methods of *projection into monotone function* and *reducing the number of roots*

are used to build an estimating function with a unique root. The former imposes strong conditions on the score estimating function, and is harder to use in a wider class. The latter is based on the fact that $\tilde{G}(\theta, \boldsymbol{x})$ in (2.34) and (2.37) has a unique root when $|\epsilon_i|$ is large. In this case, however, $\tilde{G}(\theta, \boldsymbol{x})$ may be very different from the original estimating function.

Since approximation methods do not give a good solution, many direct methods have been proposed. These include the three criteria (method 1,2,3) proposed by Heyde and Morton(1998) and the chi-square criterion (method 4) recently suggested by Singh and Mantel (1999). In addition, the minimax approach (B. Li, 1996, 1997) and the local likelihood function approach (Wang and Small, 1998) suggested useful ways to choose a consistent root. Since root intensity is a general concept for the estimating functions, the method based on root intensity can be used in a wide class of estimating functions. As a complement, the shifted information is also a quite useful tool for transformation models. Though there is no general method to solve the multiple root problem, the above methods provide useful ways to deal with the problem. The root intensity method can be shown to be a useful method in most examples discussed in Section 2.2.

Chapter 3

Root Intensity of Estimating Functions

3.1 Preliminaries

The estimating function approach is a popular method to estimate the unknown parameter, which includes maximum likelihood and least squares as well as many semiparametric methods. Asymptotic theory shows that for a wide variety of cases, there exists a unique consistent root. Perlman (1983) discussed the limiting behavior of multiple roots of score estimating equations. In general, Crowder (1986) has considered the consistency and inconsistency of estimating equations. However, the estimating equation $G(\boldsymbol{\theta}, \boldsymbol{x}) = 0$ may give multiple solutions. As an extension of probability density function to the multiple root case, Small and Yang (1999) considered the distribution of the roots as a random set and proposed the concept of *root intensity function* for which some important results have been obtained. The form of the root intensity depends upon a complete specification of the model and

its parameters. In practice, of course, the unknown parameters are the quantities we are trying to estimate. So the root intensity cannot be assumed to be known. However, we shall consider some empirical approximations to the true root intensity using sample-based methods such as the bootstrap and the saddlepoint approximation. Based on this approximation to the root intensity, a method is suggested to discriminate among multiple roots of estimating function. In the following sections, the concept of root intensity and its properties will be discussed in detail, and some approximations for root intensity such as the saddlepoint approximation will be suggested, also the root intensity for Cauchy score estimating function will be investigated extensively.

3.2 Root Intensity

By a *random zero set* of an estimating function G we mean a subset of the parameter space $\boldsymbol{\Theta} \in R^k$ given by

$$\mathcal{Z} = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : G(\boldsymbol{\theta}) = 0\} \tag{3.1}$$

Let the *count statistic* N_A be the cardinality of the set $\mathcal{Z} \bigcap A$, and $N_A^{(r)} = N_A(N_A - 1)(N_A - 2)\ldots(N_A - r + 1)$. The following Definition 3.1, and Theorems 3.2 and 3.3 are due to Small (Small & Yang, 1999).

Definition 3.1 We define the r-th order root intensity function Δ_r for $G(\boldsymbol{\theta})$ to be a function of the form

$$\Delta_r : \boldsymbol{\Theta}^r \rightarrow R^+ \tag{3.2}$$

where $\Theta^r = \Theta \times \ldots \times \Theta (r \text{ times})$, satisfying

$$E(\prod_{i=1}^{r} N_{A_i}) = \int_{A_r} \cdots \int_{A_1} \Delta_r(\theta_1, \ldots, \theta_r) d\theta_1 \ldots \theta_r \tag{3.3}$$

for every sequence of disjoint measurable subsets $A_1, \ldots, A_r$ of Θ.

It should be pointed out that the first order root intensity is the same as the probability density function of the estimator, when the corresponding estimating equation has only one root. When an estimating function has two roots $\hat{\theta}_1$ and $\hat{\theta}_2$ with probability one, and $\hat{\theta}_1$ and $\hat{\theta}_2$ have density functions $f_1(\theta)$ and $f_2(\theta)$ respectively, then the first order root intensity function is

$$\Delta(\theta) = f_1(\theta) + f_2(\theta).$$

Example 3.1: Let us consider the circle model of McCullagh (1990) again. In this case, $\bar{x}_1 \sim N(\cos(\theta_0), 1/n)$, and $\bar{x}_2 \sim N(\sin(\theta_0), 1/n)$, where $\theta_0 \in (-\pi/2, \pi/2)$, and they are independent. A routine calculation gives the density function of $\hat{\theta}_1 = \tan^{-1}(\bar{x}_2/\bar{x}_1)$:

$$\begin{aligned} & f(\theta) = \sec^2(\theta) \exp[-\frac{n}{2} \cos^2(\theta)(\sin(\theta_0) - \cos(\theta_0) tan(\theta))] \\ & \{\frac{\cos^2(\theta)}{\pi} \exp[-\frac{n}{2}(\cos(\theta_0) + \sin(\theta_0) \tan(\theta))^2 \cos^2(\theta)] \\ & + \sqrt{\frac{n}{2\pi}}(\cos(\theta_0) + \sin(\theta_0) \tan(\theta)) \cos^3(\theta) \\ & [2\Phi(\sqrt{n} \cos(\theta)(\cos(\theta_0) + \sin(\theta_0) \tan(\theta))) - 1]\} \qquad -\frac{\pi}{2} < \theta < \frac{\pi}{2}. \end{aligned} \tag{3.4}$$

We have known that another root is $\hat{\theta}_2 = \hat{\theta}_1 + \pi$, thus the first order root intensity function $\Delta(\theta) = f(\theta) + f(\theta - \pi)$. When $\theta_0 = 0$, $f(\theta) = f(\theta - \pi)$, the first order

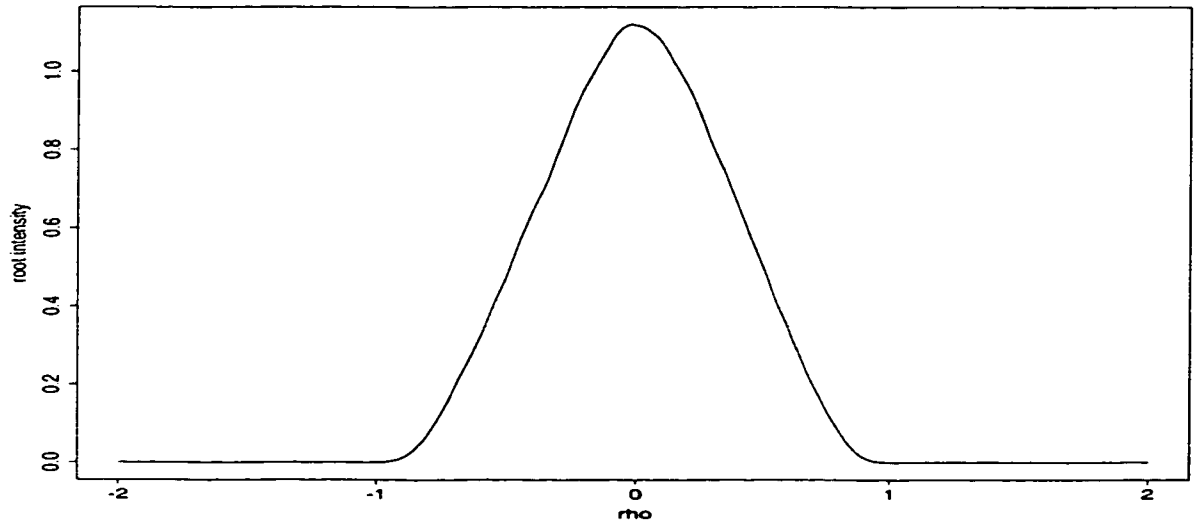

Figure 3.1: Root Intensity for Normal Correlation Coefficient ($\rho = 0$)

root intensity is

$$\Delta(\theta) = 2\sec^2(\theta)\exp[\frac{n\sin(2\theta)}{4}] \tag{3.5}$$
$$\{\frac{\cos^2(\theta)}{\pi}\exp[-\frac{n\cos^2(\theta)}{2}] - \sqrt{\frac{n}{2\pi}}\cos^3(\theta)[2\Phi(-\sqrt{n}\cos(\theta)) - 1]\}.$$

Example 3.2. When estimating the correlation coefficient for a bivariate normal distribution, we obtain the estimating function given by (2.3). Based on Theorem 3.3, using simulation and kernel approximation method (see Section 3.4), we obtain an approximation to the first order root intensity function with sample size 10.

The following theorem (Small & Yang, 1999) states that the distribution of N_A can be expressed in terms of root intensities.

Theorem 3.2 Suppose that there exists a positive integer m such that $N_A \leq m$ with probability one. Furthermore, suppose that the r-th order root intensities Δ_r

exist and are continuous for all orders $1 \leq r \leq m$. Then the probability function of N_A is given by

$$P\{N_A = r\} = \frac{1}{r!} \sum_{i=0}^{m-r} \frac{(-1)^i}{i!} \int_A \cdots \int_A \Delta_{i+r}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i+r}) d\boldsymbol{\theta}_1 \ldots d\boldsymbol{\theta}_{i+r} \tag{3.6}$$

where the zero-fold integral of Δ_0 is defined to be equal to one, and the higher order root intensities Δ_r, $r > m$ vanish everywhere.

In order to state the following results, some notations have to be introduced. Let $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$ be random vectors. We shall write $\boldsymbol{y} = o_p(\boldsymbol{x})$ if there exists a positive function $k(t)$ with $\lim_{t \to 0+} t^{-1}k(t) = 0$, such that

$$\lim_{t \to 0+} P(\| \boldsymbol{y} \| > k(t) \mid \| \boldsymbol{x} \| \leq t) = 0 \tag{3.7}$$

A combination of the two conditions $\boldsymbol{z} = o_p(\boldsymbol{x})$ and $\boldsymbol{z} = o_p(\boldsymbol{y})$ will be denoted as $\boldsymbol{z} = o_p(\boldsymbol{x}, \boldsymbol{y})$.

Let $\tilde{\boldsymbol{\theta}}(\boldsymbol{\theta})$ be the one-step estimator defined as

$$\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta} - G(\boldsymbol{\theta})(\partial G(\boldsymbol{\theta}))^{-1} \tag{3.8}$$

and $\hat{\boldsymbol{\theta}}(\boldsymbol{\theta})$ be the root of $G(\boldsymbol{\theta}) = 0$ which is closest to $\boldsymbol{\theta}$. For most estimating functions, it is reasonable to suppose that

$$\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}} = o_p(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}, \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}). \tag{3.9}$$

Jensen and Wood (1999) have shown that the above assumption is true under some regularity conditions on the model and estimating function (also see Skovgaard, 1990). However, we find some examples such that (3.9) does not hold. For example, consider the estimating function $(\bar{x} - \theta)^3 = 0$, then $\tilde{\theta} = (2\theta + \bar{x})/3$, $\hat{\theta} = \bar{x}$, then $\tilde{\theta} - \hat{\theta} = 2(\theta - \bar{x})/3$, and $\tilde{\theta} - \theta = (\bar{x} - \theta)/3$, thus (3.9) is not true.

The following theorem (Small and Yang, 1999) provides a method to evaluate the root intensity functions and investigate the properties of roots based on the estimating function itself.

Theorem 3.3 Let $f(x_1, \ldots, x_r; \theta_1, \ldots, \theta_r)$ be the density function of the continuous random vector

$$(-G(\theta_1)[\partial G(\theta_1)]^{-1}, \ldots, -G(\theta_r)[\partial G(\theta_r)]^{-1}). \tag{3.10}$$

Then under (3.9), we have

$$\Delta_r(\theta_1, \ldots, \theta_r) = f(0, \ldots, 0; \theta_1, \ldots, \theta_r) \tag{3.11}$$

for all $\theta_1, \ldots, \theta_r \in \Theta$.

In particular, the first order root intensity of the estimating function $G(\theta)$ is

$$\Delta(\theta) = f(0; \theta) \tag{3.12}$$

where $f(x; \theta)$ is the probability density function of the continuous random variable $-G(\theta)[\partial G(\theta)]^{-1}$.

3.3 Approximation to Root Intensity

3.3.1 Approximation Formula

In the last section, the formula for the root intensity is obtained. However it is not easy to calculate the root intensity in most situations, so some approximation methods are needed. In the following, we will use a saddlepoint approximation to find the approximations of some root intensity functions. It will be seen that the

saddlepoint approximation provides a very good approximation to the exponential family. For simplicity, assume that the parameter space is a subset of the real line. However, all discussions can be extended to higher dimensions. Let us consider the following estimating function:

$$G(\theta; x_1, x_2, \cdots, x_n) = \sum_{i=1}^{n} g(\theta; x_i). \tag{3.13}$$

Then

$$-\frac{\partial G(\theta; x_1, x_2, \cdots, x_n)}{\partial \theta} = -\sum_{i=1}^{n} \frac{\partial g(\theta; x_i)}{\partial \theta} \tag{3.14}$$

where $x_1, x_2, \ldots, x_n$ are identically and independently distributed as X. Assume that the characteristic function of $(g(\theta, X), -g'(\theta, X))$ is $C(t, s; \theta)$. Then the characteristic function of $(G(\theta), -G'(\theta))$ is $C^n(t, s; \theta)$. Thus, by inverse Fourier transformation, their joint probability density function can be expressed as:

$$\begin{aligned} f(x, y; \theta) &= \frac{1}{(2\pi)^2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} C^n(t, s; \theta) exp(-itx - isy) dt ds \\ &= \frac{1}{(2\pi i)^2} \int_{-i\infty}^{+i\infty} \int_{-i\infty}^{+i\infty} exp(nK(t, s; \theta) - tx - sy) dt ds \end{aligned} \tag{3.15}$$

where $K(t, s; \theta)$ is the cumulant generating function of $(g(\theta, X), -g'(\theta, X))$. By the saddlepoint approximation (McCullagh, 1987, p179 and Daniels, 1954), we have the following approximation

$$f(x, y; \theta) \sim \frac{1}{2\pi\sqrt{n}} |K^*_{rs}(\frac{x}{n}, \frac{y}{n}; \theta)|^{1/2} exp(-nK^*(\frac{x}{n}, \frac{y}{n}; \theta)) \tag{3.16}$$

where $K^*(x, y)$ is the conjugate function of $K(t, s; \theta)$ which is defined as

$$K^*(x, y) = sup_{s,t}\{tx + sy - K(s, t)\}$$

and $K^*_{rs}(x, y)$ is the second derivative matrix. Here we also use the fact that the conjugate function of $nK(t, s)$ is $nK^*(x/n, y/n)$. Since the probability density

function of $-G(\theta)/G'(\theta)$ is given by

$$\int_{-\infty}^{+\infty} f(xy, y; \theta)|y|dy$$

then the first order root intensity becomes

$$\begin{aligned}\Delta(\theta) &= \int_{-\infty}^{+\infty} f(0, y; \theta)|y|dy \\ &\sim \int_{-\infty}^{+\infty} \frac{1}{2\pi\sqrt{n}}|\hat{K}^{**}_{rs}(0, \frac{y}{n}; \theta)|^{1/2} exp(-n\hat{K}^{*}(0, \frac{y}{n}; \theta))|y|dy \\ &= \frac{n\sqrt{n}}{2\pi}\int_{-\infty}^{+\infty} |\hat{K}^{**}_{rs}(0, z; \theta)|^{1/2} exp(-n\hat{K}^{*}(0, z; \theta))|z|dz.\end{aligned}$$

Assume that $\hat{K}^{*}(0, z; \theta)$ has a smooth absolute minimum at the interior point z_0, that is, $\partial\hat{K}^{*}(z_0)/\partial z = 0$, and $\partial^2\hat{K}^{*}(z_0)/\partial z^2 > 0$. It is also assumed that for all $z \neq z_0$, $\partial\hat{K}^{*}(z)/\partial z \neq 0$. Then when n is large, using the Laplace's formula, we have

$$\Delta(\theta) \sim \frac{n}{\sqrt{2\pi\hat{K}^{*''}(0, z_0; \theta)}}|\hat{K}^{**}_{rs}(0, z_0; \theta)|^{1/2} exp(-n\hat{K}^{*}(0, z_0; \theta))|z_0| \qquad (3.17)$$

Note: In some cases, $g'(\theta, X)$ is independent of X. That is, $g'(\theta) = c(\theta)$, which does not depend on the data. Using the saddlepoint approximation, we may obtain:

$$\Delta(\theta) \sim \sqrt{\frac{n|\hat{K}^{*''}(0; \theta)|}{2\pi}}c(\theta)exp(-n\hat{K}^{*}(0; \theta)) \qquad (3.18)$$

where $\hat{K}^{*}(x)$ is the conjugate function of the cumulant generating function $\hat{K}(t)$ of $g(\theta, X)$.

3.3.2 Examples

Example 3.3. Assume that $(g(\theta, X), -g'(\theta, X))$ has a bivariate normal distribution with mean $(\mu_1(\theta), \mu_2(\theta))$ and covariance matrix:

$$\Sigma(\theta) = \begin{pmatrix} \sigma_1^2(\theta) & \rho(\theta)\sigma_1(\theta)\sigma_2(\theta) \\ \rho(\theta)\sigma_1(\theta)\sigma_2(\theta) & \sigma_2^2(\theta) \end{pmatrix}$$

then the cumulant generating function of $(g(\theta,X), -g'(\theta,X))$ is:

$$K(t,s) = \mu_1(\theta)t + \mu_2(\theta)s + \frac{1}{2}\sigma_1^2(\theta)t^2 + \rho(\theta)\sigma_1(\theta)\sigma_2(\theta)ts + \frac{1}{2}\sigma_2^2(\theta)s^2$$

and the conjugate function $K^*(x,y)$ of $K(t,s;\theta)$ is:

$$\frac{\sigma_2^2(\theta)(x-\mu_1(\theta))^2 - 2\rho(\theta)\sigma_1(\theta)\sigma_2(\theta)(x-\mu_1(\theta))(y-\mu_2(\theta)) + \sigma_1^2(\theta)(y-\mu_2(\theta))^2}{2(1-\rho^2(\theta))\sigma_1^2(\theta)\sigma_2^2(\theta)}.$$

In this case,

$$|K^*_{rs}(x,y)| = \frac{1}{(1-\rho^2(\theta))\sigma_1^2(\theta)\sigma_2^2(\theta)}$$
$$z_0 = \mu_2(\theta) - \frac{\rho(\theta)\sigma_2(\theta)\mu_1(\theta)}{\sigma_1(\theta)}$$
$$K^*(0,z_0;\theta) = \frac{\mu_1^2(\theta)}{2\sigma_1^2(\theta)}.$$

Then (3.17) gives the approximation to the first order root intensity:

$$\Delta(\theta) \sim \frac{n}{\sqrt{2\pi}\sigma_1(\theta)} \exp(-\frac{n\mu_1^2(\theta)}{2\sigma_1^2(\theta)})|\mu_2(\theta) - \frac{\rho(\theta)\sigma_2(\theta)\mu_1(\theta)}{\sigma_1(\theta)}|. \tag{3.19}$$

On the other hand, since $(G(\theta), -G'(\theta))$ has a bivariate normal distribution with mean $(n\mu_1(\theta), n\mu_2(\theta))$ and covariance matrix:

$$\Sigma_n(\theta) = \begin{pmatrix} n\sigma_1^2(\theta) & n\rho(\theta)\sigma_1(\theta)\sigma_2(\theta) \\ n\rho(\theta)\sigma_1(\theta)\sigma_2(\theta) & n\sigma_2^2(\theta) \end{pmatrix}$$

by direct calculation (see the more general case below), we may obtain the first order root intensity $\Delta(\theta)$:

$$\frac{\sigma_2(\theta)\sqrt{1-\rho^2(\theta)}}{\sigma_1(\theta)}[\frac{1}{\pi}\exp(-\frac{n}{2}(\frac{\mu_1^2(\theta)}{\sigma_1^2(\theta)} + \gamma^2(\theta)) + \frac{\sqrt{n}\gamma(\theta)}{\sqrt{2\pi}}(2\Phi(\sqrt{n}\gamma(\theta)) - 1)]$$

where

$$\gamma(\theta) = \frac{\mu_2(\theta)\sigma_1(\theta) - \mu_1(\theta)\rho(\theta)\sigma_2(\theta)}{\sqrt{1-\rho^2(\theta)}\sigma_1(\theta)\sigma_2(\theta)}.$$

This is a special case of formula (3.20) (see below).

In general, when $(G(\theta; x_1, \cdots, x_n), -G'(\theta; x_1, \cdots, x_n))$ has a bivariate normal distribution with mean $(\nu_1(\theta), \nu_2(\theta))$ and covariance matrix:

$$\Omega(\theta) = \begin{pmatrix} \delta_1^2(\theta) & r(\theta)\delta_1(\theta)\delta_2(\theta) \\ r(\theta)\delta_1(\theta)\delta_2(\theta) & \delta_2^2(\theta) \end{pmatrix}.$$

The typical examples appear in linear and nonlinear regressions. In this case, the corresponding first order root density $\Delta(\theta)$ is:

$$\frac{\delta_2(\theta)\sqrt{1-\rho^2(\theta)}}{\delta_1(\theta)}[\frac{1}{\pi}\exp(-(\frac{\nu_1^2(\theta)}{2\delta_1^2(\theta)} + \frac{s^2(\theta)}{2})) + \frac{s(\theta)}{\sqrt{2\pi}}(2\Phi(s(\theta)) - 1)] \qquad (3.20)$$

where

$$s(\theta) = \frac{\nu_2(\theta)\delta_1(\theta) - \rho(\theta)\nu_1(\theta)\delta_2(\theta)}{\sqrt{1-\rho^2(\theta)}\delta_1(\theta)\delta_2(\theta)}.$$

The above formula can be derived by assuming that the joint density function of $(G(\theta), -G'(\theta))$ is $f(x, y; \theta)$, then

$$\begin{aligned} \Delta(\theta) &= \int_{-\infty}^{\infty} f(0, y; \theta)|y|dy \\ &= c\int_{-\infty}^{\infty} |y|\exp\{-\frac{1}{2(1-r^2)}[\frac{\nu_1^2}{\delta_1^2}dy + \frac{2r\nu_1(y-\nu_2)}{\delta_1\delta_2} + \frac{(y-\nu_2)^2}{\delta_2^2}] \\ &= c\exp(-\frac{\nu_2^2}{2\delta_1^2})\int_{-\infty}^{\infty} |y|\exp(-\frac{(y-\mu)^2}{2a^2})dy \end{aligned}$$

where $c = 1/(2\pi\delta_1\delta_2\sqrt{1-r^2})$, $a = \sqrt{1-r^2}\delta_2$ and $\mu = \nu_2 - r\nu_1\delta_2/\delta_1$. Note that

$$\begin{aligned} &\int_{-\infty}^{\infty} |y|\exp(-\frac{(y-\mu)^2}{2a^2})dy \\ =\ & a\int_{-\frac{\mu}{a}}^{\infty}(at+\mu)e^{-\frac{t^2}{2}}dt + a\int_{\frac{\mu}{a}}^{\infty}(at-\mu)e^{-\frac{t^2}{2}}dt \\ =\ & a^2e^{-\frac{\mu^2}{2a^2}} + a\mu\sqrt{2\pi}(\Phi(\frac{\mu}{a}) - \Phi(\frac{-\mu}{a})) \end{aligned}$$

so (3.20) can be easily obtained from the above equalities.

An interesting example is the compartmental model:

$$y_i = e^{-\theta t_i} + \epsilon_i \tag{3.21}$$

where $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, 2, \cdots, n$, σ is fixed. In this case, the estimating function for θ is :

$$G(\theta) = \sum_{i=1}^{n} t_i (y_i - e^{-\theta t_i}) e^{-\theta t_i}. \tag{3.22}$$

Then $(G(\theta), -G'(\theta))$ has a bivariate normal distribution.

Example 3.4. Let us consider the exponential family of distributions parameterized by θ in the form:

$$f_X(x;\theta) = exp\{\theta x - K(\theta)\} f_X(x) \tag{3.23}$$

where $K(.)$ is the cumulant generating function of the distribution $f_X(x)$. Then the cumulant generating function of $f_X(x;\theta)$ is $K(t+\theta) - K(\theta)$. In this case, the first order root intensity, which is the probability density function of the estimator, can be found directly. The following example will demonstrate that the saddlepoint approximation is a very good approximation. Note that $g(\theta; X) = X - K'(\theta)$, and $g'(\theta; X) = -K''(\theta)$. Under the hypothesis that $\theta = 0$, the cumulant generating function of $g(\theta; X) = X - K'(\theta)$ is $G(t;\theta) = K(t) - K'(\theta)t$, and the corresponding conjugate function $G^*(x;\theta)$ is $K^*(x + K'(\theta))$, where $K^*(x)$ is the conjugate function of $K(\theta)$. By (3.18), we have:

$$\Delta(\theta) \sim \sqrt{\frac{n|K^{*\prime\prime}(K'(\theta))|}{2\pi}} K''(\theta) \exp[-nK^*(K'(\theta))]. \tag{3.24}$$

Now we apply (3.24) to Gamma distribution which has the density function:

$$f_X(x;\beta) = \frac{\beta^{\alpha-1} x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} = exp(\theta x - K(\theta)) f_X(x) \tag{3.25}$$

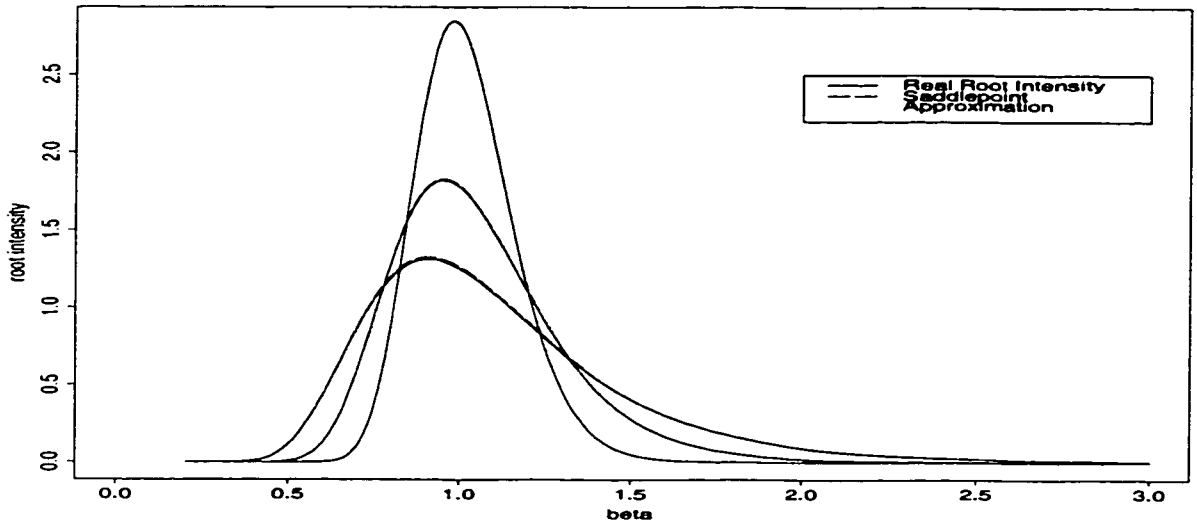

Figure 3.2: Saddlepoint Approximation vs Real Root Intensity for Gamma Distribution with $\alpha = 1$ and $\beta = 1$ for $n = 10, 20, 50$

where α is fixed, $\theta = 1 - \beta$, $K(t) = -\alpha \log(1-t)$ and $f_X(x) = x^{\alpha-1}e^{-x}/\Gamma(\alpha)$. Then $K^*(x) = x - \alpha + \alpha \log(\alpha/x)$, the saddlepoint approximation of the first order root intensity for θ is:

$$\Delta(\theta) \sim \frac{1}{1-\theta}\sqrt{\frac{n\alpha}{2\pi}} \exp\{-n\alpha(\frac{\theta}{1-\theta} + \log(1-\theta))\}.$$

Then the saddlepoint approximation of the first order root intensity for β is:

$$\Delta(\beta) \sim \frac{1}{\beta}\sqrt{\frac{n\alpha}{2\pi}} \exp\{-n\alpha(\frac{1-\beta}{\beta} + \log(\beta))\}. \tag{3.26}$$

Under the hypothesis that $\theta = 1$ ($\beta = 0$), the maximum likelihood estimator $\hat{\beta} = n\alpha / \sum X_i$ has the following density function:

$$h(x;\beta) = \frac{1}{\beta\Gamma(n\alpha)} \exp\{-n\alpha(\frac{1}{\beta} - \log n\alpha + \log\beta)\}. \tag{3.27}$$

In Figure 3.2, the saddlepoint approximations and the first order root intensities for Gamma distribution when sample sizes $n = 10, 20, 50$ are plotted. The plots

indicate that the saddlepoint approximation is a perfect approximation to the first order root intensity of the scale parameter in the Gamma distribution.

3.4 The Root Intensity for Cauchy Distribution

The Cauchy location parameter family is an important example involving multiple roots. The Cauchy score estimating equation is:

$$\begin{aligned} g(\theta, \boldsymbol{x}) &= \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log \frac{1}{\pi(1+(x_i-\theta)^2)} \\ &= \sum_{i=1}^{n} \frac{2(x_i-\theta)}{1+(x_i-\theta)^2} = 0. \end{aligned} \tag{3.28}$$

Reeds (1985) established a remarkable result on the number of its roots: the number of false relative maxima has asymptotically a Poisson distribution. The properties of the Cauchy score estimating equation will be investigated from the viewpoint of the root intensity. To begin, 5000 simulations, each with sample of size 5, were taken from the standard Cauchy distribution X (i.e., $\theta = 0$), then the 'sturm' command in Maple V is used to compute the number of roots. The 'sturm' is based on the Sturm's Theorem for polynomials. The Sturm's Theorem provides a way for finding the number of the roots of a polynomial in a given interval. Suppose that we wish to find the number of roots of a real polynomial $f(x)$ over an interval $I = [a, b]$, the Sturm chain for $f(x)$ is a sequence of polynomials f_0, f_1, f_2, ..., where $f_0 = f$, $f_1 = f'$, and for $j \geq 0$, $f_j = q_j f_{j+1} - f_{j+2}$, that is, $-f_{j+2}$ is the remainder when f_j is divided by f_{j+1}. Finally, we will obtain f_s, which is a nonzero constant. If the number of sign changes of the sequence $(f_0(a), f_1(a), \cdots, f_s(a))$ is p, and the number of sign changes of the sequence $(f_0(b), f_1(b), \cdots, f_s(b))$ is q, the Sturm's Theorem states that the number of real roots over (a,b) is $p - q$. After

Table 3.1: Numbers of Roots Based on 5000 Simulations

Interval	Number	Interval	Number	Interval	Number
(-20.0,-15.0)	228	(-4.5,-3.5)	79	(4.5, 5.5)	72
(-15.0,-12.5)	177	(-3.5,-2.5)	103	(5.5,6.5)	66
(-12.5,-10.5)	206	(-2.5,-1.5)	205	(6.5,7.5)	74
(-10.5,-9.5)	120	(-1.5,-0.5)	1018	(7.5,8.5)	121
(-9.5,-8.5)	119	(-0.5,0.5)	2853	(8.5,9.5)	147
(-8.5,-7.5)	118	(0.5,1.5)	979	(9.5,10.5)	131
(-7.5,-6.5)	70	(1.5,2.5)	217	(10.5,12.5)	216
(-6.5,-5.5)	51	(2.5,3.5)	101	(12.5,15.0)	172
(-5.5,-4.5)	65	(3.5,4.5)	62	(15.0,20.0)	241

dividing $(-20.0, 20.0)$ into 27 intervals, we use the 'sturm' command in Maple V to calculate the number of roots on each interval. The total number of roots in these intervals for 5000 simulations are shown in the following table.

The corresponding histogram is shown in Figure 3.3. From this histogram, secondary local maxima can been seen that are close to ± 9. From Theorem 3.3, the r-th order root intensity $\Delta_r(\theta_1, \ldots, \theta_r) = f(0, \ldots, 0; \theta_1, \ldots, \theta_r)$, where $f(x_1, \ldots, x_r; \theta_1, \ldots, \theta_r)$ is the density function of the random vector

$$(-g(\theta_1, \boldsymbol{x})[g'(\theta_1, \boldsymbol{x})]^{-1}, \ldots, -g(\theta_r, \boldsymbol{x})[g'(\theta_r, \boldsymbol{x})]^{-1})$$

where

$$g'(\theta, \boldsymbol{x}) = \sum_{i=1}^{n} \frac{2(x_i - \theta)^2 - 2}{(1 + (x_i - \theta)^2)^2} \tag{3.29}$$

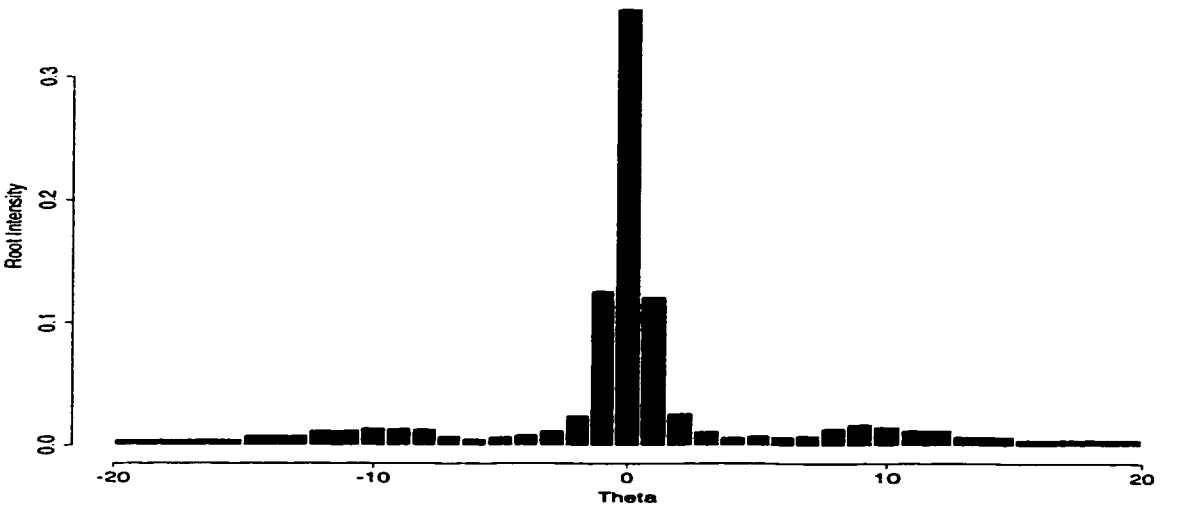

Figure 3.3: Histogram of Root Intensity for a Sample of Size 5 from the Standard Cauchy Distribution

It is easy to check that

$$E[g(\theta, \boldsymbol{x})] = \frac{-2n\theta}{\theta^2 + 4} \tag{3.30}$$

$$E[g'(\theta, \boldsymbol{x})] = \frac{2n(\theta^2 - 4)}{(\theta^2 + 4)^2}. \tag{3.31}$$

Let $r(\theta, \boldsymbol{x}) = -g(\theta, \boldsymbol{x})/g'(\theta, \boldsymbol{x})$, then the first order root intensity is:

$$\begin{aligned} \Delta(\theta) &= f(0;\theta) = \lim_{h \to 0} \frac{P(-h < r(\theta, \boldsymbol{x}) \leq h)}{2h} \\ &= \lim_{h \to 0} \frac{1}{2h} \int_{|r(\theta, \boldsymbol{x})| < h} r(\theta, \boldsymbol{x}) d\boldsymbol{x}. \end{aligned} \tag{3.32}$$

Since it is not easy to find the value of the above integration, the Monte-Carlo method can be used to find the first order root intensity. In order to increase the efficiency of calculation, a C program was written to implement this method. Based on 50000 simulations with sample size 5, a rough graph was produced for the first order root intensity. Since it contained many irregular parts, the kernel method

was used to smooth it. In other words, an approximation was made by using the kernel estimator:

$$\hat{f}(0;\theta) = \frac{1}{md}\sum_{k=1}^{m} k(\frac{r(\theta, \boldsymbol{x}_k)}{d}) \tag{3.33}$$

where the kernel function is

$$k(x) = \begin{cases} \frac{1}{4}x^4 - \frac{1}{2}|x|^3 + \frac{1}{2} & |x| \leq 1 \\ \frac{1}{4}x(2-|x|)^3 & 1 \leq |x| \leq 2 \\ 0 & |x| \geq 2 \end{cases} \tag{3.34}$$

as suggested by Silverman(1978). By taking $m = 50000$, and width $d = 0.50$, a kernel approximation of the first order root intensity for the Cauchy distribution can be made (see Figure 3.4). By comparing Figure 3.3 with Figure 3.4, we find that they have the same shape, which suggests that the method is reasonable. While the Sturm Theorem is only valid for polynomials, the Monte-Carlo and kernel methods can be used in more general situations.

In order to investigate why the root intensity of the Cauchy distribution appears as in Figure 3.3 and 3.4, especially the appearance of the bumps, let us decompose the first root intensity into two parts by Polar representation: upcrossing intensity and downcrossing intensity (see Figure 3.5 and 3.6), which represent the intensity for local minima and maxima of the likelihood respectively. From these graphs, the upcrossing intensity contributes mainly to the bumps. Similarly, a kernel approximation of the second order root intensity was obtained using the standard bivariate normal density function $h(x, y)$ as the kernel function (see Figure 3.7). That is, the kernel approximation is:

$$\hat{f}(0,0;\theta_1,\theta_2) = \frac{1}{md}\sum_{i=1}^{m} h(\frac{r(\theta_1, \boldsymbol{x})}{d}, \frac{r(\theta_2, \boldsymbol{x})}{d}) \tag{3.35}$$

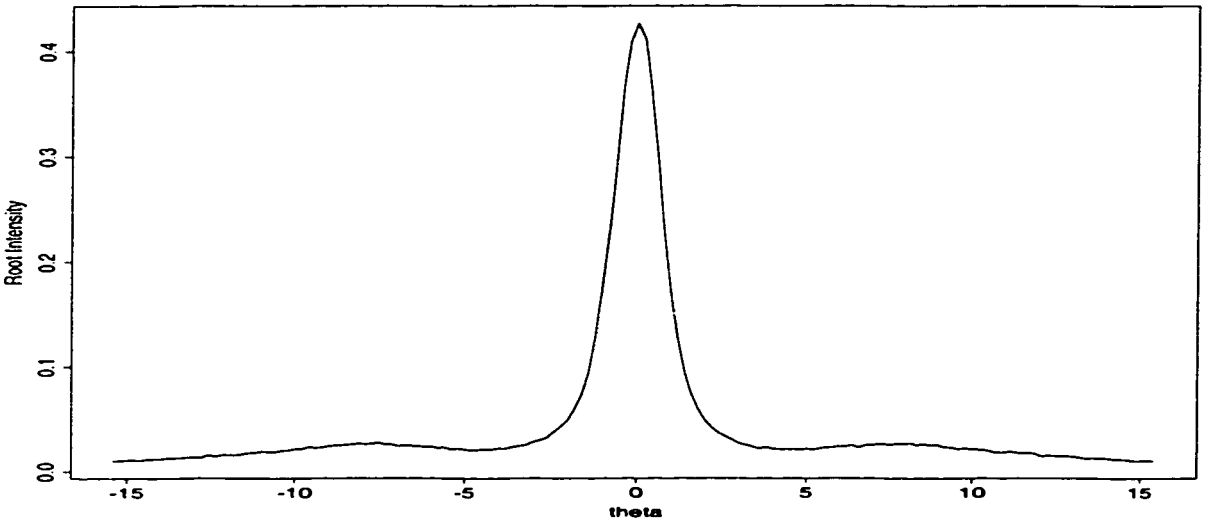

Figure 3.4: Kernel Approximation of First Order Root Intensity for a Sample of Size 5 from the Standard Cauchy Distribution

In order to check the dependence between roots, the value $\Delta_2(\theta_1, \theta_2) - \Delta_1(\theta_1)\Delta_1(\theta_2)$ was plotted as in Figure 3.8. In this case, the function is nonpositive and discernably strictly negative for a number of values of θ_1 and θ_2. There is a tendency for the roots to be isolated from each other, which provides a guarantee that for large samples the roots closest to any $\sqrt{n}$–consistent estimator will be uniquely determined and consistent.

3.5 Discussion

In this chapter, the concept of root intensity and some related results have been discussed. Theorem 3.3 provides a method to evaluate the root intensity, which is the foundation of further study on root intensity. It has been seen that the saddlepoint approximation provides a very good approximation. Laplace formula

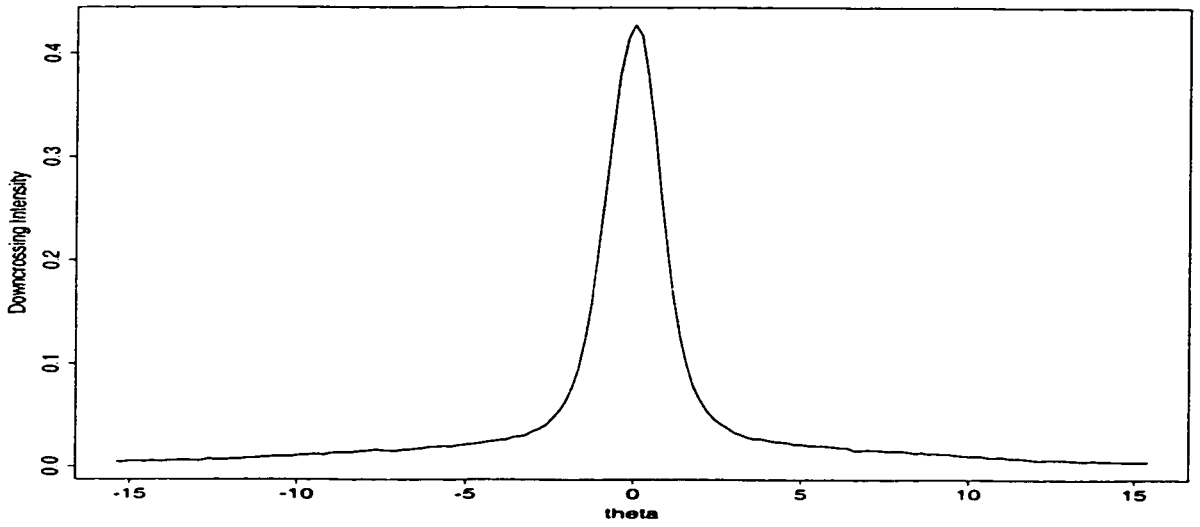

Figure 3.5: Kernel Approximation of Downcrossing Intensity for a Sample of Size 5 from the Standard Cauchy Distribution

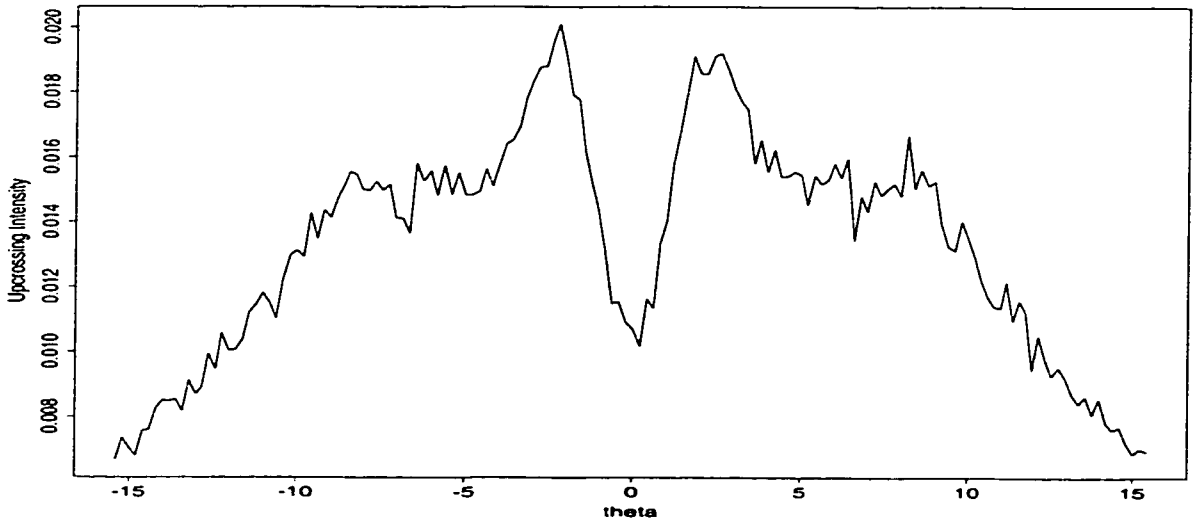

Figure 3.6: Kernel Approximation of Upcrossing Intensity for a Sample of Size 5 from the Standard Cauchy Distribution

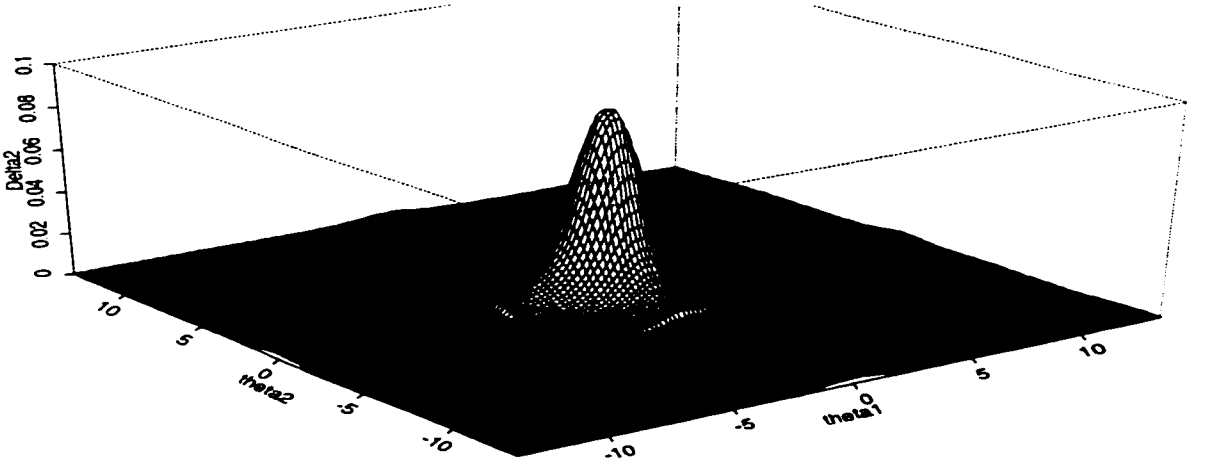

Figure 3.7: Kernel Approximation of Second Order Root Intensity for a Sample of Size 5 from the Standard Cauchy Distribution

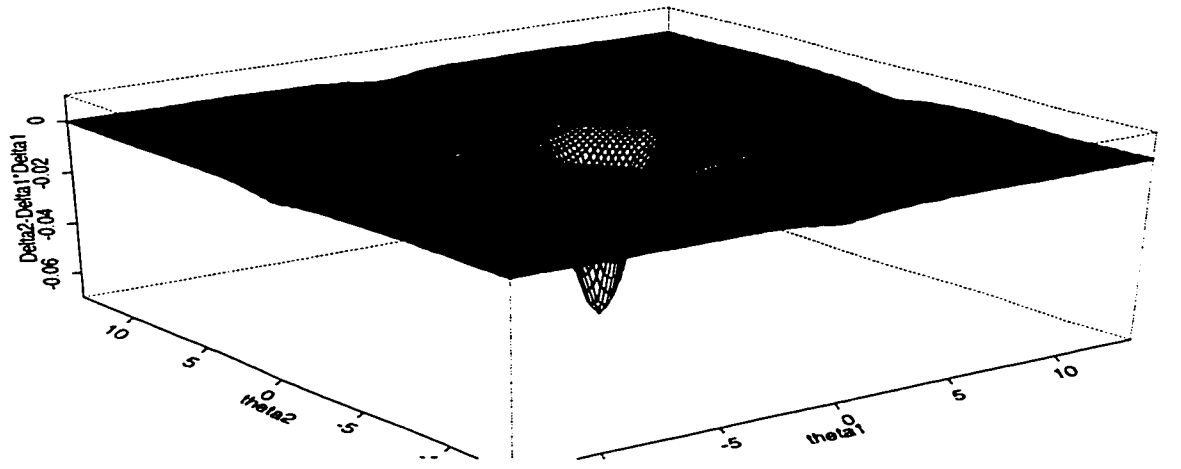

Figure 3.8: $\Delta_2(\theta_1, \theta_2) - \Delta_1(\theta_1)\Delta_1(\theta_2)$ for Cauchy Score

is also a very useful method to approximate the first order root intensity. Root intensity of Cauchy location distribution is studied in detail, which describes the root distribution for Cauchy location models.

Chapter 4

Root Selection Based on Root Intensity

4.1 Preliminaries

In this chapter, we consider estimating functions in the form of

$$G(\boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{\theta}, X_i)$$

where $X_1, \cdots, X_n$ are an i.i.d. sample from $f(x, \boldsymbol{\theta}_0)$, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset R^m$. This class of estimating functions includes all score estimating functions with i.i.d. observations. We have introduced the concept of root intensity for estimating functions in Chapter 3. It will be seen that the root intensity can provide a new approach to choose the best root among multiple roots.

We assume that $E(\boldsymbol{g}(\boldsymbol{\theta}, X_i)) \neq 0$ for all $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. We will show that under certain regularity conditions, the first order root intensity at the true value for a sample with size n, $\Delta^{(n)}(\boldsymbol{\theta}_0)$, tends to infinity in the order of $n^{m/2}$, while $\Delta^{(n)}(\boldsymbol{\theta})$ at any

other point tends to 0 exponentially as n approaches infinity. This result underlies an approach to choose the best root among multiple roots.

However, an immediate question is how to estimate the first order root intensity based on observations. A simple method is based on the central limit theorem. According to Small and Yang (1999), the first order root intensity

$$\Delta(\boldsymbol{\theta}) = f(0;\boldsymbol{\theta}) \tag{4.1}$$

where $f(\boldsymbol{x};\boldsymbol{\theta})$ is the density function of $-G(\boldsymbol{\theta})/G'(\boldsymbol{\theta})$. In this case, each component in both $G(\boldsymbol{\theta})$ and $G'(\boldsymbol{\theta})$ is the sum of independently and identically distributed random variables. The central limit theorem and related limiting distribution theory imply (Serfling, 1980)

$$\sqrt{n}(\frac{G(\boldsymbol{\theta})}{G'(\boldsymbol{\theta})} - \frac{\mu(\boldsymbol{\theta})}{\nu(\boldsymbol{\theta})}) \rightarrow N(0, \Omega(\boldsymbol{\theta}))$$

where $\Omega(\boldsymbol{\theta})$ depends on the mean vector $(\mu(\boldsymbol{\theta}), \nu(\boldsymbol{\theta}))$ and covariance matrix $\Sigma(\boldsymbol{\theta})$ of $(G(\boldsymbol{\theta}), G'(\boldsymbol{\theta}))$. Thus $\Omega(\boldsymbol{\theta})$ can be estimated by the corresponding sample mean and sample variance. Based on the above result, it is reasonable to use a normal density function to approximate the density function of $-G(\boldsymbol{\theta})/G'(\boldsymbol{\theta})$, so we may obtain an approximation of the root intensity.

Alternative methods to estimate the first order root intensity include the Edgeworth approximation, the general saddlepoint approximation, the bootstrap method with a kernel approximation. The trimmed method is also used to diminish the effect of outliers. Once we obtain an estimate of the root intensity, this approach suggests picking the root with the maximum value of the estimated root intensity as an estimate of parameter. Simulation results indicate that this is a reasonable method to choose the best estimator among multiple roots.

4.2 Asymptotic Properties of Root Intensity

In this section, we will discuss the asymptotic properties of the root intensity of the estimating function in the form of

$$G(\boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{\theta}, X_i) \tag{4.2}$$

where $X_1, \ldots, X_n$ are a sample from X. We will show that under some regularity conditions, the root intensity at the true value tends to infinity with rate $n^{m/2}$ as the sample size n become large, where m is the dimension of the parameter space.

Based on Theorem 3.3, in order to find the first order root intensity function, we need the probability density function of $-G(\boldsymbol{\theta})[\partial G(\boldsymbol{\theta})]^{-1}$. For estimating function (4.2), it is the ratio of two sums of independent and identical random variables in one dimensional case. In the next subsection, we will discuss this case. In 4.2.2, the result in the single parameter case will be extended to the multiparameter case.

4.2.1 Single Parameter Case

Geary (1944) has proved the following result when Y is nonnegative. Here we rewrite it for a more general case, that is, $Y \neq 0$.

Proposition 4.1. If (X, Y) has a bivariate density function $f(x, y)$ and characteristic function $\phi(t, s)$, let the characteristic function of $(X\text{sign}(Y), |Y|)$ be $\tilde{\phi}(t, s)$. We assume that $\lim_{s \to \pm\infty} \tilde{\phi}(t, s) = 0$. Then the density function for $R = X/Y$ is:

$$f_R(r) = \frac{1}{2\pi i} \int_{-\infty}^{\infty} [\frac{\partial \tilde{\phi}(t, s)}{\partial s}]_{s=-rt} dt \tag{4.3}$$

provided the above integral is absolutely convergent.

Proof: The density function of R is:

$$f_R(r) = \int_{-\infty}^{\infty} f(ry, y)|y|dy \tag{4.4}$$
$$= \int_0^{\infty} [f(ry, y) + f(-ry, -y)]ydy$$

Let

$$\tilde{f}(x, y) = \begin{cases} f(x, y) + f(-x, -y) & \text{if } y > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

which is the density function of $(X\text{sign}(Y), |Y|)$. Thus

$$f_R(r) = \int_0^{\infty} \tilde{f}(ry, y)ydy \tag{4.6}$$
$$= \frac{1}{4\pi^2} \int_0^{\infty} [\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \tilde{\phi}(t, s)e^{-i(tr+s)y}dtds]ydy$$
$$= \frac{1}{4\pi^2} \int_0^{\infty} \int_{-\infty}^{\infty} e^{-itry}[\int_{-\infty}^{\infty} \tilde{\phi}(t, s)ye^{-isy}ds]dtdy.$$

Since $\lim_{s\to\pm\infty} \tilde{\phi}(t, s) = 0$, the partial integration of integral gives:

$$g(t, y) \equiv \frac{i}{2\pi} \int_{-\infty}^{\infty} \tilde{\phi}(t, s)ye^{-isy}ds = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\partial\tilde{\phi}(t, s)}{\partial s} e^{-iys}ds. \tag{4.7}$$

Note that

$$\tilde{\phi}(t, s) = \int_{-\infty}^{+\infty} e^{iys}\tilde{\phi}_X(t, y)dy$$

where

$$\tilde{\phi}_X(t, y) = \int_{-\infty}^{+\infty} e^{ixt}\tilde{f}(x, y)dx$$

we have

$$\tilde{\phi}_X(t, y) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \tilde{\phi}(t, s)e^{-iys}ds$$

when $y < 0$, $\tilde{f}(x,y) = 0$, then $\tilde{\phi}_X(t,y) = 0$, therefore $g(t,y) = iy\tilde{\phi}_X(t,y) = 0$. From (4.7),

$$\frac{\partial\tilde{\phi}(t,s)}{\partial s} = \int_{-\infty}^{+\infty} g(t,y)e^{iys}dy = \int_{0}^{+\infty} g(t,y)e^{iys}dy \tag{4.8}$$

Then, from (4.6)

$$\begin{aligned} f_R(r) &= \frac{1}{2\pi i}\int_{-\infty}^{+\infty}\int_{0}^{+\infty} g(t,y)e^{-itry}dydt \\ &= \frac{1}{2\pi i}\int_{-\infty}^{+\infty}[\frac{\partial\tilde{\phi}(t,s)}{\partial s}]_{s=-rt}dt. \end{aligned} \tag{4.9}$$

Remark 4.1. In order to discuss the asymptotic properties of the first order root intensity, we hope to find the relationship between the distribution of the ratio and joint distribution of two random variables. Based on Daniels (1954), consider the following density function:

$$h(x,y) = \frac{1}{\eta}y\tilde{f}(x,y) \tag{4.10}$$

where $\eta = E_{\tilde{f}}(Y)$ and $\tilde{f}(x,y)$ is defined by (4.5). The corresponding characteristic function:

$$\psi(t,s) = \frac{1}{\eta}\frac{\partial\tilde{\phi}(t,s)}{i\partial s}. \tag{4.11}$$

Assume that (U,V) has the density function $h(x,y)$, let $W = U - rV$, then the density function of W is:

$$f_W(w) = \frac{1}{\eta}\int_{0}^{+\infty}\tilde{f}(w+ry,y)ydy. \tag{4.12}$$

So from (4.4) and (4.5), $f_R(r) = \eta f_W(0)$. Obviously, the characteristic function of W:

$$\phi_W(t) = \psi(t,-rt) = \frac{1}{\eta}[\frac{\partial\tilde{\phi}(t,s)}{i\partial s}]_{s=-rt} \tag{4.13}$$

Remark 4.2. If Y is positive, then $\tilde{f}(x,y) = f(x,y)$, $\tilde{\phi}(t,s) = \phi(t,s)$, so the density function for $R = X/Y$ is:

$$f_R(r) = \frac{1}{2\pi i}\int_{-\infty}^{\infty}[\frac{\partial\phi(t,s)}{\partial s}]_{s=-rt}dt. \tag{4.14}$$

Now we apply Proposition 4.1 to inquire into the properties of the first order root intensity. When $E(Y) \neq 0$, without loss of generality, we can assume that $E(Y) > 0$, by the strong law of large numbers, $T_n = \sum_{i=1}^n Y_i > 0$ if n is large enough. Let $\phi(t,s)$ be the characteristic function of (X,Y), and (X_i,Y_i) $(i = 1,2,\ldots,n)$ be a sample from this distribution. We wish to find an approximation to the distribution of

$$R_n = \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n Y_i} = \frac{\bar{X}_n}{\bar{Y}_n} \tag{4.15}$$

where $\bar{X}_n$ and $\bar{Y}_n$ are sample means. Based on the above discussion and the saddlepoint approximation, we have the following result:

Proposition 4.2. Let the cumulant generating function of (X_n,Y_n) be $K(t,s)$, then there is t_0 such that the density function of R_n is:

$$\begin{aligned} f_{R_n}(r) &= [\frac{n}{2\pi K''(t_0,-rt_0)}]^{1/2}\exp[nK(t_0,-rt_0)] \\ &\quad\times[\frac{\partial K(t_0,-rt_0)}{\partial u}][1+O(\frac{1}{n})] \end{aligned} \tag{4.16}$$

where $\partial K/\partial u$ is the derivative with respect to the second variable, and t_0 satisfies

$$\begin{aligned} &[\frac{\partial K(t,-rt)}{\partial t}]_{t=t_0} = 0 \\ &[\frac{\partial^2 K(t,-rt)}{\partial t^2}]_{t=t_0} > 0. \end{aligned} \tag{4.17}$$

Proof: Since the characteristic function of $(\bar{X}_n, \bar{Y}_n)$ is

$$\phi_n(t,s) = [\phi(\frac{t}{n}, \frac{s}{n})]^n = \exp[nK(\frac{it}{n}, \frac{is}{n})]$$

then

$$\frac{\partial \phi_n(t,s)}{\partial s} = i \exp(nK(\frac{it}{n}, \frac{is}{n})][\frac{\partial K(\tau,u)}{\partial u}]_{\tau=\frac{it}{n}, u=\frac{is}{n}}.$$

As with Remark 4.1, we assume that (U_n, V_n) has a density function given by $yf_n(x,y)/\eta$, where $f_n(x,y)$ is the probability density function of $(\bar{X}_n, \bar{Y}_n)$, and the corresponding characteristic function is $\psi_n(t,s)$. For any given r, let $W_n = U_n - rV_n$, then the characteristic function of W_n is

$$\phi_{W_n}(t) = \psi_n(t,-rt) = \frac{1}{\eta}[\frac{\partial \phi_n(t,s)}{i\partial s}]_{s=-rt}$$

where $\eta = E(\bar{Y}_n) = E(Y)$. The density function of R_n is:

$$\begin{aligned} f_{R_n}(r) &= \eta f_{W_n}(0) = \frac{\eta}{2\pi}\int_{-\infty}^{+\infty} \phi_{W_n}(t)dt \qquad (4.18)\\ &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} \exp[nK(\frac{it}{n}, \frac{-irt}{n})][\frac{\partial K(\tau,u)}{\partial u}]_{\tau=\frac{it}{n}, u=\frac{-irt}{n}}dt \\ &= \frac{n}{2\pi i}\int_{\alpha-i\infty}^{\alpha+i\infty} \exp(nK(t,-rt))[\frac{\partial K(\tau,u)}{\partial u}]_{\tau=t, u=-rt}dt \\ &= [\frac{n}{2\pi K''(t_0,-rt_0)}]^{1/2}\exp[nK(t_0,-rt_0)][\frac{\partial K(t_0,-rt_0)}{\partial u}][1+O(\frac{1}{n})] \end{aligned}$$

where t_0 satisfies (4.17). In the last equality, we used the saddlepoint approximation (Daniels, 1954).

Remark 4.3. We may assume that $K(t,-rt)$ has only a local minimum. From (4.17), it has been seen that

$$K(t_0,-rt_0) = \inf_t(K(t,-rt)).$$

When $r = 0$, $K(t_0, -rt_0) = \inf_t(K(t,0)) = \inf_t K_X(t)$. Note that we have the following useful fact:

Lemma 4.1. If $E(X) = 0$, then $\inf_t K_X(t) = 0$; if $E(X) \neq 0$, then $\inf_t K_X(t) < 0$.

Proof: When $E(X) = 0$, then $K_X(0) = 0$ and

$$K_X(t) = \log E(e^{tX}) \geq E(\log e^{tX}) = E(tX) = 0$$

thus $\inf_t K_X(t) = 0$. On the other hand, since $K'_X(0) = E(X)$, when $E(X) \neq 0$, $K_X(t)$ is strictly increasing or decreasing locally at 0, there is a point t^* in the neighborhood of 0 such that $K_X(t^*) < 0$, thus $\inf_t K_X(t) < 0$.

From Proposition 4.2, Remark 4.3 and Theorem 3.3, the first order root intensity $\Delta^{(n)}(\theta)$ of the estimating function $G(\theta) = \sum g(\theta, X_i)$ is the density function of $-G(\theta)/G'(\theta)$ evaluated at 0. Since $-G(\theta)/G'(\theta)$ is the ratio of the sums of random variables, Proposition 4.2 can be applied to this case. That is, the first order root intensity of $G(\theta)$ can be expressed as

$$\Delta^{(n)}(\theta) = f_{R_n}(0;\theta) = c\sqrt{n}\exp[n \inf_t(K_g(t,\theta))][1 + O(\frac{1}{n})]$$

where $K_g(t,\theta)$ is the cumulant generating function of $g(\theta, X)$, and c is a constant independent of n. Based on Lemma 4.1, when $E_{\theta_0}[g(X;\theta_0)] = 0$, $\inf_t[K_g(t,\theta_0)] = 0$, and when $E_{\theta_0}[g(X;\theta)] \neq 0$, $\inf_t[K_g(t,\theta)] < 0$. Thus we have the following result:

Theorem 4.1. Let $X_1, \cdots, X_n$ be a sample from X. Assume that

(i) $E_{\theta_0}[g(X;\theta)] \neq 0$ if $\theta \neq \theta_0$; and $E_{\theta_0}[g(X;\theta_0)] = 0$.

(ii) $E_{\theta_0}[g'(X;\theta)] \neq 0$ for all $\theta \in \Theta$.

Then the first order root intensity $\Delta^{(n)}(\theta)$ of the estimating function $G(\theta)$ has the following properties:

(i)

$$\lim_{n\to\infty} \frac{\Delta^{(n)}(\theta_0)}{\sqrt{n}} = c > 0.$$

(ii) For any $\theta \neq \theta_0$, there is a $\alpha = \alpha(\theta) > 0$ such that

$$\lim_{n\to\infty} \Delta^{(n)}(\theta) e^{n\alpha} = 0.$$

In other words, $\Delta^{(n)}(\theta)$ tends to 0 exponentially as n.

4.2.2 Multiparameter Case

In order to obtain an estimate of the first order root intensity, based on theorem 3.3, we hope to find an estimate for $-G(\boldsymbol{\theta})[\partial G(\boldsymbol{\theta})]^{-1}$. To this end, we need to extend Geary's (1944) result to a higher dimension case. To the best of my knowledge, there is no literature about this. In the following, I will use the properties of Fourier transformation for a higher dimension to develop a similar result.

In order to state the following result, we need some notations:

(i) If $S = (s_{kl})$ and $Y = (y_{kl})$ are two $m \times m$ matrix, we define $S'Y = \sum_{k=1}^{m} \sum_{l=1}^{m} s_{kl} y_{kl}$.

(ii) Let $\boldsymbol{t} = (t_1, \cdots, t_m)^t$, and $\boldsymbol{r} = (r_1, \cdots, r_m)^t$,

$$\boldsymbol{t} \otimes \boldsymbol{r} = \boldsymbol{t}\boldsymbol{r}' = \begin{pmatrix} t_1 r_1 & \cdots & t_1 r_m \\ \cdots & \cdots & \cdots \\ t_m r_1 & \cdots & t_m r_m \end{pmatrix}$$

(iii) Let the characteristic function of $(\boldsymbol{x}, Y)$ be $\phi(\boldsymbol{t}, S)$, that is,

$$\phi(\boldsymbol{t}, S) = E(e^{i(\boldsymbol{t}'\boldsymbol{x}+S'Y)}) = E(e^{i(\sum_{k=1}^{m} t_k x_k + \sum_{k=1}^{m}\sum_{l=1}^{m} s_{kl} y_{kl})})$$

(iv) Denote an operator is defined as

$$\Lambda = \begin{vmatrix} \frac{\partial}{\partial s_{11}} & \frac{\partial}{\partial s_{12}} & \cdots & \frac{\partial}{\partial s_{1m}} \\ \frac{\partial}{\partial s_{21}} & \frac{\partial}{\partial s_{22}} & \cdots & \frac{\partial}{\partial s_{2m}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial}{\partial s_{m1}} & \frac{\partial}{\partial s_{m2}} & \cdots & \frac{\partial}{\partial s_{mm}} \end{vmatrix}$$

That is,

$$\Lambda\phi(\boldsymbol{t}, S) = \sum_{j_1 j_2 \cdots j_m} (-1)^{\tau(j_1 j_2 \cdots j_m)} \frac{\partial^m \phi(\boldsymbol{t}, S)}{\partial s_{1j_1} \partial s_{2j_2} \cdots \partial s_{mj_m}}$$

where $(j_1 j_2 \cdots j_m)$ is the permutation of $\{1, 2, \cdots, n\}$ and $\tau(j_1 j_2 \cdots j_m)$ is the number of inverse orders in this permutation. Fox example, $\tau(21) = 1$, $\tau(231) = 2$. When $m = 2$,

$$\Lambda\phi(\boldsymbol{t}, S) = \frac{\partial^2 \phi(\boldsymbol{t}, S)}{\partial s_{11} \partial s_{22}} - \frac{\partial^2 \phi(\boldsymbol{t}, S)}{\partial s_{12} \partial s_{21}}$$

In the following, we will extend Geary's result to $R = Y^{-1}\boldsymbol{x}$ case, where $\boldsymbol{x}$ is the m dimension random vector, $Y = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_m)$ is $m \times m$ random matrix. This matrix is also regarded as a $m \times m$ dimension vector. In order that Y is invertible, without loss of generality, we assume that $\det(Y) > 0$.

Proposition 4.3. Assume that for the characteristic function $\phi(\boldsymbol{t}, S)$

$$\lim_{s_{kl} \to \pm\infty} \phi(\boldsymbol{t}, S) = 0 \quad k, l = 1, 2, \cdots m$$

and all integrals, derivatives and limits are exchangeable, then the joint density function for $R = Y^{-1}\boldsymbol{x}$ is:

$$f_R(\boldsymbol{r}) = \frac{1}{(2\pi i)^m} \int_{R^m} [\Lambda\phi(\boldsymbol{t}, S)]_{S=-\boldsymbol{t}\otimes\boldsymbol{r}} dt \tag{4.19}$$

Proof: Assume that the density function of $(\boldsymbol{x}, Y)$ is $f(\boldsymbol{x}, Y)$. Then the density function of $R = Y^{-1}\boldsymbol{x}$:

$$\begin{aligned} f_R(\boldsymbol{r}) &= \int_{R^{m^2}} f(Y\boldsymbol{r}, Y)\det(Y)dY \\ &= (\frac{1}{2\pi})^{m(m+1)} \int_{R^{m^2}} [\int_{R^{m^2}} \int_{R^m} \phi(\boldsymbol{t}, S)e^{-i(\boldsymbol{t}'Y\boldsymbol{r}+S'Y)}d\boldsymbol{t}dS]\det(Y)dY \\ &= (\frac{1}{2\pi})^{m(m+1)} \int_{\Omega} \int_{R^m} e^{-i\boldsymbol{t}'Y\boldsymbol{r}} [\int_{R^{m^2}} \phi(\boldsymbol{t}, S)e^{-iS'Y}\det(Y)dS]d\boldsymbol{t}dY \end{aligned} \tag{4.20}$$

where $\Omega = \{Y \in R^{m^2}, \det(Y) > 0\}$. Let

$$\begin{aligned} g(\boldsymbol{t}, Y) &= \int_{R^{m^2}} \phi(\boldsymbol{t}, S)e^{-iS'Y}\det(Y)dS \\ &= \int_{R^{m^2}} \phi(\boldsymbol{t}, S)e^{-iS'Y} \sum_{\{j_1 j_2 \cdots j_m\}} (-1)^{\tau(j_1 j_2 \cdots j_m)} y_{1j_1} y_{2j_2} \cdots y_{mj_m} dS \\ &= \sum_{\{j_1 j_2 \cdots j_m\}} (-1)^{\tau(j_1 j_2 \cdots j_m)} \int_{R^{m^2-m}} e^{-i(S'Y - \sum_{k=1}^m s_{kj_k} y_{kj_k})} \\ &\quad [\int_{R^m} \phi(\boldsymbol{t}, S)e^{-i\sum_{k=1}^m s_{kj_k} y_{kj_k}} y_{1j_1} \cdots y_{mj_m} ds_{1j_1} \cdots ds_{mj_m}]dS^* \end{aligned} \tag{4.21}$$

where $S^* = S - \{s_{1j_1} \cdots s_{mj_m}\}$. For $l = 1, 2, \cdots, m$, denote

$$\psi_l(\boldsymbol{t}, S_l) = \int_{R^l} \phi(\boldsymbol{t}, S)e^{-i\sum_{k=1}^l s_{kj_k} y_{kj_k}} y_{1j_1} \cdots y_{lj_l} ds_{1j_1} \cdots ds_{lj_l}$$

where $S_l = S - \{y_{1j_1}, \cdots, y_{lj_l}\}$. By assumptions, we have $\lim_{s_{kj_k} \to \pm\infty} \psi_l(\boldsymbol{t}, S_l) = 0$ when $k > l$. Using partial integration, we have:

$$\begin{aligned} \psi_1(\boldsymbol{t}, S_1) &= \int_{-\infty}^{\infty} \phi(\boldsymbol{t}, S)e^{-is_{1j_1} y_{1j_1}} y_{1j_1} ds_{1j_1} \\ &= -i \int_{-\infty}^{\infty} \frac{\partial \phi(\boldsymbol{t}, S)}{\partial s_{1j_1}} e^{-is_{1j_1} y_{1j_1}} ds_{1j_1}. \end{aligned}$$

We also note that

$$\psi_2(\boldsymbol{t}, S_2) = \int_{-\infty}^{\infty} \psi_1(\boldsymbol{t}, S_1)e^{-is_{2j_2} y_{2j_2}} y_{2j_2} ds_{2j_2}$$

$$= -i\int_{-\infty}^{\infty}\frac{\partial\psi_1(\boldsymbol{t},S_1)}{\partial s_{2j_2}}e^{-is_{2j_2}y_{2j_2}}ds_{2j_2}$$
$$= (-i)^2\int_{R^2}\frac{\partial\phi^2(\boldsymbol{t},S)}{\partial s_{1j_1}\partial s_{2j_2}}e^{-i(s_{1j_1}y_{1j_1}+s_{2j_2}y_{2j_2})}ds_{1j_1}ds_{2j_2}.$$

Repeat this step. Finally, we obtain:

$$\psi_m(\boldsymbol{t},S_m)=(-i)^m\int_{R^m}\frac{\partial^m\phi(\boldsymbol{t},S)}{\partial s_{1j_1}\cdots\partial s_{mj_m}}e^{-i\sum_{k=1}^m s_{kj_k}y_{kj_k}}ds_{1j_1}\cdots ds_{mj_m}. \tag{4.22}$$

Therefore,

$$\begin{aligned} & g(\boldsymbol{t},Y) \qquad (4.23)\\ = & \sum_{\{j_1j_2\cdots j_m\}}(-1)^{\tau(j_1j_2\cdots j_m)}\int_{R^{m^2-m}}e^{-i(S'Y-\sum_{k=1}^m s_{kj_k}y_{kj_k})}\psi_m(\boldsymbol{t},S_m)dS_m\\ = & (-i)^m\int_{R^{m^2}}\sum_{\{j_1j_2\cdots j_m\}}(-1)^{\tau(j_1j_2\cdots j_m)}\frac{\partial^m\phi(\boldsymbol{t},S)}{\partial s_{1j_1}\cdots\partial s_{mj_m}}e^{-iS'Y}dS\\ = & (-i)^m\int_{R^{m^2}}\Lambda\phi(\boldsymbol{t},S)e^{-iS'Y}dS. \end{aligned}$$

Since $\phi(\boldsymbol{t},S)$ can be written as

$$\phi(\boldsymbol{t},S)=\int_{R^{m^2}}e^{iS'Y}\phi_{\boldsymbol{x}}(\boldsymbol{t},Y)dY \tag{4.24}$$

where

$$\phi_{\boldsymbol{x}}(\boldsymbol{t},Y)=\int_{R^m}e^{i\boldsymbol{t}'\boldsymbol{x}}f(\boldsymbol{x},Y)d\boldsymbol{x}.$$

From (4.21) and (4.24), we have

$$\phi_{\boldsymbol{x}}(\boldsymbol{t},Y)=(\frac{1}{2\pi})^{m^2}\int_{R^{m^2}}\phi(\boldsymbol{t},S)e^{-iS'Y}dS=\frac{g(\boldsymbol{t},Y)}{(2\pi)^{m^2}\det(Y)}.$$

Since it is assumed that $P\{\det(Y)>0\}=1$, when $Y\in\Omega^c$, $f(\boldsymbol{x},Y)=0$, then $\phi_{\boldsymbol{x}}(\boldsymbol{t},Y)=0$, therefore $g(\boldsymbol{t},Y)=(2\pi)^{m^2}\det(Y)\phi_{\boldsymbol{x}}(\boldsymbol{t},Y)=0$. From (4.23),

$$\Lambda\phi(\boldsymbol{t},S)=\frac{i^m}{(2\pi)^{m^2}}\int_{R^{m^2}}g(\boldsymbol{t},Y)e^{iS'Y}dY=\frac{i^m}{(2\pi)^{m^2}}\int_{\Omega}g(\boldsymbol{t},Y)e^{iS'Y}dY. \tag{4.25}$$

Then,

$$\begin{aligned} f_R(\boldsymbol{r}) &= (\frac{1}{2\pi})^{m(m+1)} \int_{R^m} \int_{\Omega} g(\boldsymbol{t}, Y) e^{-i\boldsymbol{t}' Y \boldsymbol{r}} dY d\boldsymbol{t} \\ &= (\frac{1}{2\pi})^{m(m+1)} \int_{R^m} \int_{\Omega} g(\boldsymbol{t}, Y) e^{-i(\boldsymbol{t} \otimes \boldsymbol{r})' Y} dY d\boldsymbol{t} \\ &= \frac{1}{(2\pi i)^m} \int_{R^m} [\Lambda \phi(\boldsymbol{t}, S)]_{S=-\boldsymbol{t} \otimes \boldsymbol{x}} d\boldsymbol{t}. \end{aligned} \tag{4.26}$$

This completes the proof of Proposition 4.2.

Lemma 4.2. Assume that the density function and characteristic function of $(\boldsymbol{x}, Y)$ are $f(\boldsymbol{x}, Y)$ and $\phi(\boldsymbol{t}, S)$ respectively, and $(\boldsymbol{u}, V)$ has density function:

$$h(\boldsymbol{x}, Y) = \begin{cases} \frac{1}{\eta} \det(Y) f(\boldsymbol{x}, Y) & \text{if } \det(Y) > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\eta = E_f(\det(Y))$. Then the characteristic function of $(\boldsymbol{u}, V)$ is:

$$\psi(\boldsymbol{t}, S) = \frac{(-i)^m}{\eta} \Lambda \phi(\boldsymbol{t}, S) \tag{4.27}$$

For any given $\boldsymbol{r}$, let $\boldsymbol{w} = \boldsymbol{u} - V\boldsymbol{r}$, then the density function of $R = Y^{-1}\boldsymbol{x}$, $f_R(\boldsymbol{r}) = \eta f_W(0)$, where $f_W(\boldsymbol{w})$ is the density function of $\boldsymbol{w}$.

Proof: It is obvious that

$$\begin{aligned} \Lambda \phi(\boldsymbol{t}, S) &= \Lambda E_f[e^{i(\boldsymbol{t}' \boldsymbol{x} + S' Y)}] \\ &= \Lambda \int_{R^{m(m+1)}} f(\boldsymbol{x}, Y) e^{i(\boldsymbol{t}' \boldsymbol{x} + S' Y)} d\boldsymbol{x} dY \\ &= i^m \int_{R^{m(m+1)}} f(\boldsymbol{x}, Y) \det(Y) e^{i(\boldsymbol{t}' \boldsymbol{x} + S' Y)} d\boldsymbol{x} dY \\ &= i^m \eta E_h[e^{i(\boldsymbol{t}' \boldsymbol{x} + S' Y)}] \\ &= i^m \eta \psi(\boldsymbol{t}, S). \end{aligned} \tag{4.28}$$

Thus (4.27) is obtained. Since the density function of $\boldsymbol{w} = \boldsymbol{u} - V\boldsymbol{r}$ is:

$$f_W(\boldsymbol{w}) = \int_{R^{m^2}} h(\boldsymbol{w} + V\boldsymbol{r}, V) dV = \frac{1}{\eta} \int_{R^{m^2}} f(\boldsymbol{w} + Y\boldsymbol{r}, Y) \det(Y) dY \tag{4.29}$$

Therefore from (4.20), $f_R(\boldsymbol{r}) = \eta f_W(0)$.

Now, we consider a sample $(\boldsymbol{x}_n, Y_n)$ from $(\boldsymbol{x}, Y)$, for which $E(\det(Y)) > 0$. Let

$$R_n = (\sum_{i=1}^{n} Y_i)^{-1} \sum_{i=1}^{n} \boldsymbol{x}_n. \tag{4.30}$$

Since $E(\det(Y)) > 0$, when n is large enough, $\det(\sum_{i=1}^{n} Y_n) > 0$.

Proposition 4.4. Let the cumulant generating function of $(\boldsymbol{x}, Y)$ be $K(\boldsymbol{t}, S)$, then the density function of R_n has the following approximation:

$$f_{R_n}(\boldsymbol{r}) = (\frac{n}{2\pi})^{\frac{m}{2}} \frac{\exp[nK(\boldsymbol{t}_0, -\boldsymbol{t}_0 \otimes \boldsymbol{r})]\Lambda_0 K(\boldsymbol{t}_0, -\boldsymbol{t}_0 \otimes \boldsymbol{r})}{|\det(K_{t_i t_j}(\boldsymbol{t}_0, -\boldsymbol{t}_0 \otimes \boldsymbol{r}))|^{\frac{1}{2}}}[1 + O(\frac{1}{n})] \tag{4.31}$$

where $\boldsymbol{t}_0$ satisfies

$$\nabla K(\boldsymbol{t}, -\boldsymbol{t} \otimes \boldsymbol{r})|_{\boldsymbol{t}_0} = 0 \tag{4.32}$$

and

$$\Lambda_0 K(\boldsymbol{\tau}, U) = \begin{vmatrix} \frac{\partial K}{\partial u_{11}} & \frac{\partial K}{\partial u_{12}} & \cdots & \frac{\partial K}{\partial u_{1m}} \\ \frac{\partial K}{\partial u_{21}} & \frac{\partial K}{\partial u_{22}} & \cdots & \frac{\partial K}{\partial u_{2m}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial K}{\partial u_{m1}} & \frac{\partial K}{\partial u_{m2}} & \cdots & \frac{\partial K}{\partial u_{mm}} \end{vmatrix}.$$

Proof: We use $\bar{\boldsymbol{x}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$, $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ to substitute $\boldsymbol{x}$ and Y in Lemma 4.2 respectively. All corresponding functions and random variables will be denoted by the same notation with subscript n. In this case, the characteristic function of $(\bar{\boldsymbol{x}}_n, \bar{Y}_n)$ is:

$$\phi_n(\boldsymbol{t}, S) = [\phi(\frac{\boldsymbol{t}}{n}, \frac{S}{n})]^n = \exp[nK(\frac{i\boldsymbol{t}}{n}, \frac{iS}{n})].$$

For any given $\boldsymbol{r}$, the $\boldsymbol{w}_n$ corresponding to $\boldsymbol{w}$ in Lemma 4.2 has characteristic function:

$$\phi_{W_n}(\boldsymbol{t}) = E_{h_n}[e^{\boldsymbol{t}'(\boldsymbol{u}_n - V_n \boldsymbol{r}_n)}] = E_{h_n}[e^{\boldsymbol{t}'\boldsymbol{u}_n - (\boldsymbol{t}\otimes\boldsymbol{r})'V_n}] = \frac{(-i)^m}{\eta}\Lambda[\phi_n(\boldsymbol{t}, S)]_{S=-\boldsymbol{t}\otimes\boldsymbol{r}}$$

where $\eta = E(\det(\bar{Y}_n)) = E(\det(Y))$. In fact, let $y_{ij}^{(l)}$ $(l = 1, 2, \ldots, n)$ be a sample from y_{ij}.

$$\begin{aligned} E(\det(\bar{Y}_n)) &= E[\sum_{\{j_1\cdots j_m\}} \frac{(-1)^{\tau(j_1\cdots j_m)}}{n^m}(\sum_{l_1=1}^{n} y_{1j_1}^{(l_1)})(\sum_{l_2=1}^{n} y_{2j_2}^{(l_2)})\cdots(\sum_{l_m=1}^{n} y_{mj_m}^{(l_m)})] \\ &= \sum_{l_1=1,l_2=1,\cdots,l_m=1}^{n} E(\sum_{\{j_1\cdots j_m\}} \frac{(-1)^{\tau(j_1\cdots j_m)}}{n^m} y_{1j_1}^{(l_1)} y_{2j_2}^{(l_2)}\cdots y_{mj_m}^{(l_m)}) \\ &= \sum_{l_1=1,l_2=1,\cdots,l_m=1}^{n} \frac{1}{n^m} E(\det(Y)) = E(\det(Y)). \end{aligned}$$

So the density function of R_n is:

$$f_{R_n}(\boldsymbol{r}) = \eta f_{W_n}(0) = \frac{\eta}{(2\pi)^m}\int_{R^m} \phi_{W_n}(\boldsymbol{t})d\boldsymbol{t} \tag{4.33}$$

$$\begin{aligned} &= \frac{(-i)^m}{(2\pi)^m}\int_{R^m} \Lambda[\phi_n(\boldsymbol{t}, S)]_{S=-\boldsymbol{t}\otimes\boldsymbol{r}}\, d\boldsymbol{t} \\ &= \frac{1}{(2\pi)^m}\int_{R^m} \exp[nK(\frac{i\boldsymbol{t}}{n}, \frac{-i\boldsymbol{t}\otimes\boldsymbol{r}}{n})]\Lambda^* K(\boldsymbol{\tau}, U)|_{\boldsymbol{\tau}=\frac{i\boldsymbol{t}}{n}, U=-\frac{i\boldsymbol{t}\otimes\boldsymbol{r}}{n}}\, d\boldsymbol{t} \end{aligned}$$

where $\Lambda^* K(\boldsymbol{\tau}, U) = \sum_{\{j_1\cdots j_m\}}(-1)^{\tau(j_1\cdots j_m)} L_{j_1\cdots j_m} K(\boldsymbol{\tau}, U)$, where

$$\begin{aligned} L_{j_1\cdots j_m} K(\boldsymbol{\tau}, U) &= \frac{\partial K}{\partial u_{1j_1}}\frac{\partial K}{\partial u_{2j_2}}\cdots\frac{\partial K}{\partial u_{mj_m}} \\ &+ \frac{1}{n}(\frac{\partial^2 K}{\partial u_{1j_1}\partial u_{2j_2}}\frac{\partial K}{\partial u_{3j_3}}\cdots\frac{\partial K}{\partial u_{mj_m}} + \cdots + \frac{\partial K}{\partial u_{1j_1}}\cdots\frac{\partial K}{\partial u_{(m-2)j_{m-2}}}\frac{\partial^2 K}{\partial u_{(m-1)j_{m-1}}\partial u_{mj_m}}) \\ &+ \cdots + \frac{\partial^m K}{n^{m-1}\partial u_{1j_1}\cdots\partial u_{mj_m}}. \end{aligned}$$

Therefore, $\Lambda^* K(\boldsymbol{\tau}, U)$ can be written as

$$\Lambda^* K(\boldsymbol{\tau}, U) = \Lambda_0 K(\boldsymbol{\tau}, U) + \frac{1}{n}\Lambda_1 K(\boldsymbol{\tau}, U).$$

Thus, from Bruijn (1981) and Bleistein (1975),

$$\begin{aligned}
f_{R_n}(\boldsymbol{r}) &= \frac{n^m}{(2\pi)^m}\int_{C^m}\exp[nK(\boldsymbol{t},-\boldsymbol{t}\otimes\boldsymbol{r}]\Lambda^* K(\boldsymbol{t},-\boldsymbol{t}\otimes\boldsymbol{r})d\boldsymbol{t} \\
&= \frac{n^m}{(2\pi)^m}\int_{C^m}\exp[nK(\boldsymbol{t},-\boldsymbol{t}\otimes\boldsymbol{r}]\Lambda_0 K(\boldsymbol{t},-\boldsymbol{t}\otimes\boldsymbol{r})d\boldsymbol{t} \\
&+ \frac{n^{m-1}}{(2\pi)^m}\int_{C^m}\exp[nK(\boldsymbol{t},-\boldsymbol{t}\otimes\boldsymbol{r}]\Lambda_1 K(\boldsymbol{t},-\boldsymbol{t}\otimes\boldsymbol{r})d\boldsymbol{t} \\
&= \frac{n^m}{(2\pi)^m}\frac{\exp[nK(\boldsymbol{t}_0,-\boldsymbol{t}_0\otimes\boldsymbol{r})]}{|\det(K_{t_it_j}(\boldsymbol{t}_0,-\boldsymbol{t}_0\otimes\boldsymbol{r}))|^{\frac{1}{2}}}(\frac{2\pi}{n})^{\frac{m}{2}}[\Lambda_0 K(\boldsymbol{t}_0,-\boldsymbol{t}_0\otimes\boldsymbol{r}) \\
&\quad +\frac{1}{n}\Lambda_1 K(\boldsymbol{t}_0,-\boldsymbol{t}_0\otimes\boldsymbol{r})+O(\frac{1}{n})] \\
&= (\frac{n}{2\pi})^{\frac{m}{2}}\frac{\exp[nK(\boldsymbol{t}_0,-\boldsymbol{t}_0\otimes\boldsymbol{r})]\Lambda_0 K(\boldsymbol{t}_0,-\boldsymbol{t}_0\otimes\boldsymbol{r})}{|\det(K_{t_it_j}(\boldsymbol{t}_0,-\boldsymbol{t}_0\otimes\boldsymbol{r}))|^{\frac{1}{2}}}[1+O(\frac{1}{n})]
\end{aligned}$$

where $C=\{xi; x\in R\}$. This completes our proof.

Similar to the single parameter case, it has been seen that

$$K(\boldsymbol{t}_0,-\boldsymbol{t}_0\otimes\boldsymbol{r})=\inf_{\boldsymbol{t}}(K(\boldsymbol{t},-\boldsymbol{t}\otimes\boldsymbol{r})).$$

When $\boldsymbol{r}=0$, $K(\boldsymbol{t}_0,-\boldsymbol{t}_0\otimes\boldsymbol{r})=\inf_{\boldsymbol{t}}(K(\boldsymbol{t},0))=\inf_{\boldsymbol{t}}K_{\boldsymbol{x}}(\boldsymbol{t})$. Note that the following important fact:

Lemma 4.4. If $E(\boldsymbol{x})=0$, then $\inf_{\boldsymbol{t}}K_{\boldsymbol{x}}(\boldsymbol{t})=0$; If $E(\boldsymbol{x})\neq 0$, then $\inf_{\boldsymbol{t}}K_{\boldsymbol{x}}(\boldsymbol{t})<0$.

Proof: When $E(\boldsymbol{x})=0$, then $K_{\boldsymbol{x}}(0)=0$ and

$$K_{\boldsymbol{x}}(\boldsymbol{t})=\log E(e^{\boldsymbol{t}'\boldsymbol{x}})\geq E(\log e^{\boldsymbol{t}'\boldsymbol{x}})=E(\boldsymbol{t}'\boldsymbol{x})=0$$

therefore, $\inf_{\boldsymbol{t}}K_{\boldsymbol{x}}(\boldsymbol{t})=0$. On the other hand, since $\partial K_{\boldsymbol{x}}(0)/\partial t_i=E(x_i)$, when $E(\boldsymbol{x})\neq 0$, there is a point $\boldsymbol{t}^*$ in the neighborhood of 0 such that $K_{\boldsymbol{x}}(\boldsymbol{t}^*)<0$, so $\inf_{\boldsymbol{t}}K_{\boldsymbol{x}}(\boldsymbol{t})<0$.

From Proposition 4.2, Lemma 4.4 and Theorem 3.3, based on the similar arguments derived to Theorem 4.1, we can obtain the following results:

Theorem 4.2. Let $X_1, \cdots, X_n$ be the sample from X. Assume that

(i) $E_{\theta_0}[\boldsymbol{g}(\boldsymbol{\theta}, X)] \neq 0$ if $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$; and $E_{\boldsymbol{\theta}_0}[\boldsymbol{g}(\boldsymbol{\theta}_0, X)] = 0$;

(ii) $E_{\boldsymbol{\theta}_0}[\nabla \boldsymbol{g}(X; \boldsymbol{\theta})] \neq 0$ for all $\boldsymbol{\theta}$.

Then the first order root intensity $\Delta^{(n)}(\boldsymbol{\theta})$ of the estimating function $G(\boldsymbol{\theta}) = \sum \boldsymbol{g}(\boldsymbol{\theta}, X_i)$ has the following properties:

(i)

$$\lim_{n\to\infty} \frac{\Delta^{(n)}(\boldsymbol{\theta}_0)}{n^{m/2}} = c > 0.$$

(ii) For any $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, $\Delta^{(n)}(\boldsymbol{\theta})$ tends to 0 exponentially as n.

Example 4.1: We can illustrate this result using the following simple estimating functions:

$$g_1(\mu, \sigma^2) = \sum_{i=1}^{n} X_i - n\mu = 0$$
$$g_2(\mu, \sigma^2) = \sum_{i=1}^{n} X_i^2 - n(\mu^2 + \sigma^2) = 0$$

where $X_1, \cdots, X_n$ is a sample from standard normal distribution. By solving these estimating functions, we obtained $\hat{\mu} = \bar{X}_n$, $\hat{\sigma^2} = \sum_{i=1}^{n}(X_i - \bar{X}_n)^2/n = S_n^2$. Since $\bar{X}_n$ and S_n^2 are independent, and $S_n^2 \sim \chi^2_{n-1}$, we can easily obtain the joint density

function for $(\hat{\mu}, \hat{\sigma^2})$, that is, the first order root intensity for the above estimating functions is:

$$\begin{aligned}\Delta(\hat{\mu}, \hat{\sigma^2}) &= \sqrt{\frac{n}{2\pi}} e^{-\frac{n\hat{\mu}^2}{2}} \frac{n(n\hat{\sigma^2})^{\frac{n-1}{2}-1} e^{-\frac{n\hat{\sigma^2}}{2}}}{\Gamma(\frac{n-1}{2}) 2^{\frac{n-1}{2}}} \\ &\approx \sqrt{\frac{n}{2\pi}} e^{-\frac{n\hat{\mu}^2}{2}} \frac{\sqrt{n}}{2\sqrt{\pi}} (\frac{n}{n-3})^{\frac{n}{2}-1} e^{-\frac{3}{2}} (\hat{\sigma^2})^{\frac{n-3}{2}} e^{\frac{n}{2}(1-\hat{\sigma^2})}\end{aligned}$$

In the above, we have used the Stirling formula for $\Gamma(\frac{n-1}{2})$. Note that

$$\lim_{n\to\infty} (\frac{n}{n-3})^n = e^3$$

then we have

$$\lim_{n\to\infty} \frac{\Delta(0,1)}{n} = \frac{1}{2\sqrt{2\pi}}.$$

Note that $f(x) = xe^{1-x} < 1$ for all positive $x \neq 1$, when $\hat{\sigma^2} \neq 1$, $(\hat{\sigma^2} e^{1-\hat{\sigma^2}})^{n/2}$ tends to 0 exponentially. Therefore, the root intensity at any point except for $(0, 1)$ tends to 0 exponentially as n approaches infinity.

4.3 Estimation Methods for Root Intensity

In this section, we will discuss some practical methods to estimate the root intensity functions. In order to present these methods more clearly, we only describe them in the single parameter case. However, all these methods can be easily extended into the multiparameter case.

4.3.1 Normal Approximation

We have shown in Chapter 3 that the first order root intensity for one parameter θ is:

$$\Delta(\theta) = f(0; \theta) \tag{4.34}$$

where $f(\boldsymbol{x};\theta)$ is the density function of $-G(\theta)/G'(\theta)$. Since

$$G(\theta) = \sum_{i=1}^{n} g(\theta, X_i)$$

and

$$G'(\theta) = \sum_{i=1}^{n} g'(\theta, X_i)$$

$G(\theta)$ and $G'(\theta)$ are the sums of independently and identically distributed random variables. Assume that $E[g(\theta)] = \mu_1(\theta)$ and $E[g'(\theta)] = \mu_2(\theta)$, and

$$\mathrm{Cov}(g(\theta), g'(\theta)) = \Sigma(\theta) = \begin{pmatrix} \sigma_1^2(\theta) & \rho(\theta)\sigma_1(\theta)\sigma_2(\theta) \\ \rho(\theta)\sigma_1(\theta)\sigma_2(\theta) & \sigma_2^2(\theta) \end{pmatrix}.$$

To continue our discussion, we need the following result (Serfling, 1980):

Theorem 4.3. Suppose that X_n is a k-dimensional random vector and $\sqrt{n}(X_n - \boldsymbol{\mu})$ has multinormal limiting distribution with mean 0 and covariance matrix Σ. Let $\boldsymbol{g}(\boldsymbol{x})$ be a differentiable vector valued function, then

$$\sqrt{n}[\boldsymbol{g}(X_n) - \boldsymbol{g}(\boldsymbol{\mu})] \to N(0, D\Sigma D^t) \tag{4.35}$$

in distribution, where $D = (\partial g_i/\partial x_j)_{m\times k}$ at $\boldsymbol{\mu}$.

Thus from the central limit theorem and Theorem 4.3, we have

$$\sqrt{n}[\frac{G(\theta)}{G'(\theta)} - \frac{\mu_1(\theta)}{\mu_2(\theta)}] \to N(0, \omega^2(\theta)) \tag{4.36}$$

in distribution, where

$$\omega^2(\theta) = \frac{\sigma_1^2(\theta)}{\mu_2^2(\theta)} - \frac{2\mu_1(\theta)\rho(\theta)\sigma_1(\theta)\sigma_2(\theta)}{\mu_2^3(\theta)} + \frac{\mu_1^2(\theta)\sigma_2^2(\theta)}{\mu_2^4(\theta)}$$

That is, $-G_n(\theta)/G_n'(\theta)$ are distributed with $N(-\mu_1(\theta)/\mu_2(\theta), \omega^2(\theta))$ asymptotically. In practice, it is reasonable to use the corresponding normal density function to approximate the density function of $-G_n(\theta)/G_n'(\theta)$.

For example, if $X_1, X_2, \cdots, X_n$ has exponential distribution with mean 1, then

$$T_n = \frac{\sum_{i=1}^{n} X_i - n}{\sqrt{n}}$$

has a standard normal distribution asymptotically. By a direct calculation and use of Stirling formula, the probability density function for T_n is:

$$\begin{aligned} f_{T_n}(t) &= \frac{\sqrt{n}[\sqrt{n}(t+\sqrt{n})]e^{-\sqrt{n}(t+\sqrt{n})}}{\Gamma(n)} \\ &\sim \frac{1}{\sqrt{2\pi}}(1+\frac{t}{\sqrt{n}})^n e^{-\sqrt{n}t} \end{aligned}$$

Since

$$\begin{aligned} & n\log(1+\frac{t}{\sqrt{n}}) - \sqrt{n}t \\ = \; & n(\frac{t}{\sqrt{n}} - \frac{t^2}{2n}) + o(\frac{1}{\sqrt{n}}) - \sqrt{n}t \\ = \; & -\frac{t^2}{2} + o(\frac{1}{\sqrt{n}}) \end{aligned}$$

we have

$$f_{T_n}(t) \to \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

which is the probability density function of the standard normal distribution.

According to this idea, the first order root intensity can be approximated by the normal density at zero:

$$f_n(0;\theta) = \frac{\sqrt{n}}{\sqrt{2\pi}\omega(\theta)} \exp[-\frac{n\mu_1^2(\theta)}{2\omega^2(\theta)\mu_2^2(\theta)}].$$

As $\omega(\theta)$ depends only on the means and covariance matrix of $(g(\theta), g'(\theta))$, we may use the corresponding sample means and sample variances to estimate $\mu_1(\theta)$, $\mu_2(\theta)$

and $\omega(\theta)$. For each θ in the parameter space,

$$\hat{\mu}_1(\theta) = \frac{\sum_{i=1}^n g(x_i,\theta)}{n}$$
$$\hat{\mu}_2(\theta) = \frac{\sum_{i=1}^n g'(x_i,\theta)}{n}$$
$$\hat{\sigma}_1^2(\theta) = \frac{\sum_{i=1}^n [g(x_i,\theta)]^2}{n} - [\hat{\mu}_1(\theta)]^2;$$
$$\hat{\sigma}_2^2(\theta) = \frac{\sum_{i=1}^n [g'(x_i,\theta)]^2}{n} - [\hat{\mu}_2(\theta)]^2;$$
$$\hat{\rho}(\theta) = \frac{\sum_{i=1}^n [g(x_i,\theta)g'(x_i,\theta)] - n\hat{\mu}_1(\theta)\hat{\mu}_2(\theta)}{n\hat{\sigma}_1(\theta)\hat{\sigma}_2(\theta)}.$$

When there are outliers, we use the trimmed method described in Section 4.3.5 for the above estimates. Then the corresponding estimate of $\omega^2(\theta)$ is:

$$\hat{\omega}^2(\theta) = \frac{\hat{\sigma}_1^2(\theta)}{\hat{\mu}_2^2(\theta)} - \frac{2\hat{\mu}_1(\theta)\hat{\rho}(\theta)\hat{\sigma}_1(\theta)\hat{\sigma}_2(\theta)}{\hat{\mu}_2^3(\theta)} + \frac{\hat{\mu}_1^2(\theta)\hat{\sigma}_2^2(\theta)}{\hat{\mu}_2^4(\theta)}.$$

Thus we can obtain an estimate of $f_n(0,\theta)$:

$$\hat{f}_n(0,\theta) = \frac{\sqrt{n}}{\sqrt{2\pi}\hat{\omega}(\theta)} \exp\left[-\frac{n\hat{\mu}_1^2(\theta)}{2\hat{\omega}^2(\theta)\hat{\mu}_2^2(\theta)}\right].$$

It is reasonable to use $\hat{f}_n(0;\theta)$ as an estimate of the first order root intensity. The simulations in the next section show that this is a useful and simple method. It is straightforward to extend the above method to a multiparameter case.

4.3.2 Edgeworth Approximations

An alternative approximation is the Edgeworth approximation (McCullagh 1987, Barndorff-Nielsen and Cox, 1979,1989, 1994) which can give a more accurate approximation. The Edgeworth expansion for a density function of random vector X is the normal density, which has the same mean vector and covariance matrix

as X, multiplied by a sum of correction terms whose coefficients are simple combinations of cumulants of X. The normal approximation is a special Edgeworth approximation which only takes the leading term. Let us consider

$$W_n(\theta) = \frac{\sqrt{n}\left(\frac{G_n(\theta)}{G_n'(\theta)} - \frac{\mu_1(\theta)}{\mu_2(\theta)}\right)}{\omega(\theta)} \tag{4.37}$$

which is a nonlinear function of $G_n(\theta)$ and $G_n'(\theta)$. We can also get a form of stochastic expansion of a random variable whose distribution has an Edgeworth expansion with coefficients determined from the standard expansions for the cumulants of $G_n(\theta)$ and $G_n'(\theta)$ (see Barndorff-Nielsen and Cox, 1989). This is similar to *General Saddlepoint Approximations* in the next subsection. Thus we skip the details and leave them to the next subsection. However, the Edgeworth approximation is less accurate than the saddlepoint approximation, particularly in the far tails of the distribution. The saddlepoint approximation is often sufficiently accurate even when the sample size is small (see Section 3.3.2). It also preserves high relative accuracy over all possible value. However, unlike the saddlepoint approximation, the Edgeworth series can be easily computed without knowing the generating function.

4.3.3 General Saddlepoint Approximations

The normal approximation is a simple method, however it may be less accurate than the general saddlepoint approximation. The saddlepoint approximation introduced by Daniels (1954) provides a better approximation than the normal approximation for the distribution of a statistic. Reid (1988) gave a good review of the saddlepoint approximation. However, it was basically proposed as a method of approximating the density of a sum of independently and identically distributed random variables with a known cumulant generating function. Therefore it cannot directly be applied to the ratio of two sums of independently and identically distributed random

variables. Easton and Ronchetti (1986) proposed the *general saddlepoint approximation* which can be used for statistics in multivariate analysis expressed as the smooth functions of a sum of random vectors. This method uses the approximate cumulant generating function based on the first four cumulants of a statistic.

The approximate cumulant generating function of W_n in (4.37) is:

$$\begin{aligned}\tilde{K}_n(t) &= k_1 t + \frac{k_2}{2!}t^2 + \frac{k_3}{3!}t^3 + \frac{k_4}{4!}t^4 \\ &= \frac{1}{2}t^2 + \frac{1}{\sqrt{n}}(a_1 t + \frac{1}{6}a_3 t^3) + \frac{1}{n}(\frac{1}{2}a_2 t^2 + \frac{1}{24}a_4 t^4) + o(n^{-\frac{3}{2}})\end{aligned}$$

where

$$a_1 = -\frac{1}{\mu_{01}^3\omega}(\mu_{01}\mu_{11} - \mu_{10}\mu_{02})$$

$$\begin{aligned}a_2 = \frac{1}{\mu_{01}^6\omega^2}(&5\mu_{01}^2\mu_{11}^2 - 2\mu_{01}\mu_{10}^2\mu_{03} - \mu_{01}^4\mu_{20} + 3\mu_{01}^2\mu_{02}\mu_{20} + 2\mu_{01}^3\mu_{10}\mu_{11} \\ &+8\mu_{10}^2\mu_{02}^2 - \mu_{01}^2\mu_{10}^2\mu_{02} - 2\mu_{01}^3\mu_{21} - 16\mu_{01}\mu_{10}\mu_{02}\mu_{11} + 4\mu_{01}^2\mu_{10}\mu_{12})\end{aligned}$$

$$\begin{aligned}a_3 = \frac{1}{\mu_{01}^9\omega^3}(&3\mu_{01}^4\mu_{10}^2\mu_{12} + \mu_{01}^6\mu_{30} + 12\mu_{01}^4\mu_{10}\mu_{11}^2 - \mu_{01}^3\mu_{10}^3\mu_{03} - 3\mu_{01}^5\mu_{21}\mu_{10} \\ &-18\mu_{01}^3\mu_{10}^2\mu_{02}\mu_{11} - 6\mu_{01}^5\mu_{20}\mu_{11} + 6\mu_{01}^4\mu_{10}\mu_{02}\mu_{20} + 6\mu_{01}^2\mu_{10}^3\mu_{02}^2)\end{aligned}$$

$$\begin{aligned}a_4 = -\frac{1}{\mu_{01}^{12}\omega^4}(&24\mu_{01}^3\mu_{10}^4\mu_{02}\mu_{03} + 288\mu_{01}^3\mu_{10}^3\mu_{11}\mu_{02}^2 - 12\mu_{01}^7\mu_{10}\mu_{11}\mu_{20} \\ &+48\mu_{01}^5\mu_{10}^2\mu_{02}\mu_{21} - 60\mu_{01}^4\mu_{10}^3\mu_{02}\mu_{12} + 84\mu_{01}^5\mu_{10}^2\mu_{11}\mu_{12} - 36\mu_{01}^4\mu_{10}^3\mu_{11}\mu_{03} \\ &-348\mu_{01}^4\mu_{10}^2\mu_{11}^2\mu_{02} + 168\mu_{01}^5\mu_{10}\mu_{11}\mu_{02}\mu_{20} - 60\mu_{01}^6\mu_{10}\mu_{11}\mu_{21} - 12\mu_{01}^5\mu_{10}^3\mu_{02}\mu_{11} \\ &+6\mu_{01}^6\mu_{10}^2\mu_{02}\mu_{20} + 12\mu_{01}^6\mu_{10}^2\mu_{11}^2 - 72\mu_{01}^2\mu_{10}^4\mu_{02}^3 + 120\mu_{01}^5\mu_{10}\mu_{11}^3 \\ &+3\mu_{01}^4\mu_{10}^4\mu_{02}^2 - \mu_{01}^4\mu_{10}^4\mu_{04} - 6\mu_{01}^6\mu_{10}^2\mu_{22} + 4\mu_{01}^5\mu_{10}^3\mu_{13} - 60\mu_{01}^6\mu_{11}^2\mu_{20} \\ &-84\mu_{01}^4\mu_{10}^2\mu_{02}^2\mu_{20} - \mu_{01}^8\mu_{40} + 12\mu_{01}^7\mu_{11}\mu_{30} + 4\mu_{01}^7\mu_{10}\mu_{31} \\ &-12\mu_{01}^6\mu_{02}\mu_{20}^2 + 12\mu_{01}^7\mu_{20}\mu_{21} + 3\mu_{01}^8\mu_{20}^2 - 12\mu_{01}^6\mu_{10}\mu_{02}\mu_{30} \\ &-24\mu_{01}^6\mu_{10}\mu_{20}\mu_{12} + 12\mu_{01}^5\mu_{10}^2\mu_{20}\mu_{30}).\end{aligned}$$

These values are computed by using Maple V. Once we obtain $\tilde{K}_n(t)$, then the general saddlepoint approximation to the density of W_n is:

$$\hat{f}_n(x) = [\frac{1}{2\pi\tilde{K}_n''(t_0)}]^{1/2}\exp\{\tilde{K}_n(t_0) - t_0 x\} \tag{4.38}$$

where t_0 is determined as a solution of the equation $\tilde{K}_n'(t) = x$. Thus the general saddlepoint approximation to the density of $G_n(\theta)/G_n'(\theta)$ can be derived straightforward. Though it can give a more approximate estimate, computation seems very tedious for our purpose.

4.3.4 Bootstrap Method

For a given sample $\boldsymbol{x}$, we may use the nonparametric or parametric bootstrap method to produce m independent bootstrap samples $\boldsymbol{x}^{*1}, \boldsymbol{x}^{*2}, \cdots, \boldsymbol{x}^{*m}$. Our goal is to estimate the value at zero of the density function of $R_n(\boldsymbol{\theta}, \boldsymbol{x}) = -G_n(\boldsymbol{\theta}, \boldsymbol{x})/G_n'(\boldsymbol{\theta}, \boldsymbol{x})$. Based on the m samples, we can calculate $r_i = R_n(\boldsymbol{\theta}, \boldsymbol{x}^{*i})$, $i = 1, 2, \cdots, m$. Thus we may use

$$\hat{\Delta}(\boldsymbol{\theta}) = \frac{1}{md}\sum_{i=1}^{m} K(\frac{r_i}{d}) \tag{4.39}$$

where d is bandwidth and $K(t)$ is a kernel function. Usually, the kernel $K(t)$ is a symmetric probability density function. The common kernels for one dimensional case are Silverman kernel (see Chapter 2), Epanechnikov, Biweight, Triangular, Gaussian (Silverman, 1986, p43). A special case is the Rectangular kernel:

$$K(t) = \begin{cases} \frac{1}{2} & \text{if } |t| < 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.40}$$

which gives the Obsenblatt estimator (or naive estimator). That is,

$$\hat{\Delta}(\boldsymbol{\theta}) = \frac{1}{2md}[\text{no. of } r_1, r_2, \cdots, r_m \text{ falling in } (-d, d)]. \tag{4.41}$$

4.3.5 Trimmed Method

In order to eliminate or diminish the effect of outliers, the trimmed method should be used. That is, for a given sample, we consider the corresponding order statistic: $x^{(1)}, x^{(2)}, \cdots, x^{(m)}$, then delete several largest values or smallest values. For symmetric case, we delete the several largest values and smallest values, so we only consider the following sample,

$$x^{(k+1)}, x^{(k+2)}, \cdots, x^{(n-k-1)}, x^{(n-k)}$$

then apply the above methods to this new sample.

4.3.6 Comments

In this section, we have discussed several methods to estimate the first order root intensity. The normal approximation is an easy method. Although it may give less accurate estimate, it can give the result much faster. Simulation results (see Section 5.5) also suggest that it is a good practical method. The generalized saddle-point method and bootstrap method need more complex algebraic and numerical calculations, thus they are less attractive. For an m dimensional parameter space, they may involve a very complex calculation, even though we use the multi-normal approximation, since we have to calculate all the sample means, variance and covariances of $\boldsymbol{g}(\boldsymbol{\theta})$ and $\partial \boldsymbol{g}(\boldsymbol{\theta})$ and perform the corresponding algebraic calculation . In this case, the bootstrap method may give a good alternative. The bootstrap method may need more computation time, which is not a problem in this computer stage, however we can save a lot of laborious work.

Basically, the estimate of the first order root intensity is based on the probability density estimation. All methods of the probability density estimation can be applied

to this case. Tapia and Thompson (1978) and Silverman (1986) discussed many different methods of the probability density estimation in their books. Besides the kernel method mentioned in Section 4.3.4, other methods include the nearest neighbour method, the variable kernel methods, the orthogonal series estimators, more generally, the general weight function estimator.

In order to avoid repetition, we will present simulation results about Cauchy location model in Chapter 5. In Section 5.5, we will compare four different approaches based on sample: likelihood, median, root intensity, shifted information methods and the fifth method which is based on the known true value. It will be shown that these methods can make the same choice with high probability. The likelihood method usually gives a consistent estimator, however it can only be applied to the score estimating functions. Thus we advocate the root intensity and the shifted information methods. They can be used in a far more larger class of estimating functions than the score estimating functions.

4.4 Examples

4.4.1 Regression with Measurement Error

Stefanski and Carroll (1987) considered the generalized linear models in which the covariates cannot be observed directly, but can only be measured with a certain amount of measurement error. Multiple root problem arises in this problem. A special case is the logistic regression with errors in covariates. Suppose that Y_i is a binary response with $p_i = P(Y_i = 1)$, we use a logistic model to fit it. That is,

$$\log \frac{p_i}{1 - p_i} = \alpha + \beta' x_i.$$

Suppose that the covariates x_i are not observed directly, but only indirectly through $z_i = x_i + \epsilon_i$, where the measurement error ϵ_i is assumed to be normally distributed with mean zero and covariance matrix Ω. Stefanski and Carroll (1987) obtained the conditional score when conditioning on the complete sufficient statistic for the nuisance parameters x_i, namely, $A_i = z_i + y_i\Omega\beta$. Hanfelt and Liang (1995) modified this conditional score by building the following estimating function:

$$g(\theta) = \sum_i \begin{pmatrix} 1 & d_i \end{pmatrix}^t (y_i - \mu_i^c) \tag{4.42}$$

where

$$\mu_i^c = \frac{\exp\{\alpha + (A_i - \frac{1}{2}\Omega\beta)'\beta\}}{1 + \exp\{\alpha + (A_i - \frac{1}{2}\Omega\beta)^t\beta\}}$$

and $d_i = A_i + (\mu_i^c - 1)\Omega\beta$. Wang and Small (1998) used a method based on local likelihood function to discuss its multiple root problem. Consider the case with only two parameters α and β. we generate a sample with size $n = 100$ with $x_i \sim N(0, 0.8^2)$ and $\epsilon_i \sim N(0, (0.8/3)^2)$, and the true values of parameters is $\alpha = -1.4$, $\beta = 1.4$. We will find two roots using Matlab: $(-1.05, 2.33)$ and $(-0.96, 8.83)$. In this case, there are only two parameters, the normal approximation is a good choice since its simplicity. Based on the normal approximation, the root intensity at $(-1.05, 2.33)$ is 0.7531, the root intensity at $(-0.96, 8.83)$ is 0.3645. So we should choose $(-1.05, 2.33)$, which is closer to the true value $(-1.4, 1.4)$. It seems that it is a right choice. At the true value, the root intensity is 1.8229, which is much larger than the value at other points. This supports the conclusion of Theorem 4.2.

4.4.2 Mixture Models

In this subsection, we apply our method to a practical example (see McLachlan and Basford, 1987). We obtain this data set from Professor McLachlan. This is a data

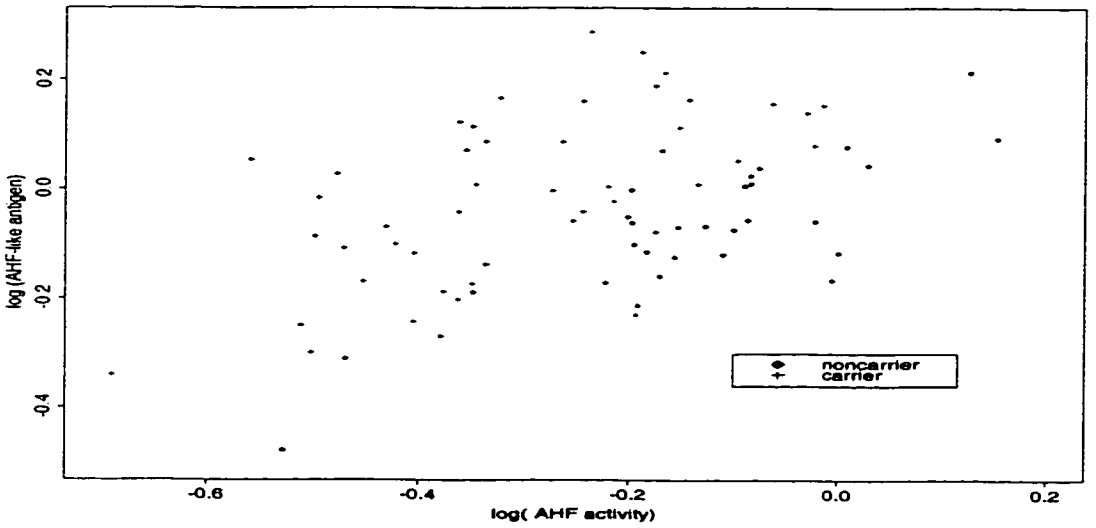

Figure 4.1: Data from Habbema, Hermans and van den Broek (1974)

set consisting only of two bivariate populations. This data set was the same as the one studied in Habbema, Hermans and van den Broek (1974). The question is how to discriminate between normal women and haemophilia A carriers based on two variables. $x_1 = \log_{10}(\text{AHF activity})$ and $x_2 = \log_{10}(\text{AHF-like antigen})$. The available data set contains 30 observations on known noncarriers and 45 observations on known obligatory carriers. We will denote these observations as x_{ij} $(i = 1, 2;$ $j = 1, \ldots, n_i)$. The data set is plotted in Figure (4.1). Let F_1 and F_2 be the populations of noncarriers and carriers respectively. It is assumed that the populations have a bivariate normal with means $\boldsymbol{\mu}_1 = (\mu_{11}, \mu_{12})'$ and $\boldsymbol{\mu}_2 = (\mu_{21}, \mu_{22})'$ respectively and common covariance matrix

$$\Omega = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2. \end{pmatrix}$$

Basford and McLachlan(1985) used a mixture model of a bivariate normal model to fit these 75 observations. That is, these observations are considered to be taken

Table 4.1: Estimates Under Homoscedasticity

p	μ_{11}	μ_{12}	μ_{21}	μ_{22}	σ_1^2	ρ	σ_2^2	L	RI
0.716	-0.206	-0.080	-0.321	0.079	0.0265	0.7422	0.0171	75.00	0.0251
0.528	-0.121	-0.019	-0.370	-0.052	0.0137	0.5760	0.0220	73.49	0.2486
0.681	-0.153	0.012	-0.420	-0.135	0.0138	0.2252	0.0175	73.29	0.3094

from a mixture of F_1 and F_2 with unknown proportions p and $1-p$. Thus the corresponding score estimating functions give three roots which are shown in Table 4.1. Since there are 8 parameters, the score estimating equation $G(\boldsymbol{\theta}) = 0$ consists of 8 estimating equations. Then $\partial G(\boldsymbol{\theta})$ is a 8 by 8 matrix, which has 36 different components since it is symmetric. Suppose that we use the normal approximation, we need to compute 44 different sample means, variances and their correlations. This is computationally burdensome. The same problem occurs for the Edgeworth approximation and the generalized saddlepoint approximation. Based on the bootstrap method, we got the estimated root intensity function shown in the last column in Table 4.1. We use 1000 bootstrap resamplings and the naive estimator to get the above estimations. According to these values, the first local maximum with the largest loglikelihood value is not a good choice. This agrees with Basford and McLachlan (1985). However, the root intensity estimation based on the bootstrap method suggests that the third root rather than the second one (Basford and McLachlan, 1985) is the best choice.

4.5 Discussion

This chapter is an important one in this thesis. We have developed the asymptotic result for the root intensity in this chapter. It says that under some regularity conditions, the first order root intensity at the true value tends to infinity for large samples, while its value at other point tends to zero quickly. This provides a basis for root selection of estimating functions. In Section 4.3, many different estimation methods for the root intensity were discussed. Two practical examples were also presented in Section 4.4. However, these theoretical results in this chapter are based on the estimating functions in the form of (4.2). It is of interest to investigate how to extend it to more general estimating functions such as martingale estimating functions. This is a future research topic.

Chapter 5

The Shifted Information Criterion

5.1 Preliminaries

In this chapter, an alternative criterion is proposed to choose among the multiple roots of estimating functions for transformation models. This information-based criterion is simple in practice. The typical example of multiple roots is the Cauchy location model. For this model, we will investigate and compare different approaches to choose the best root among multiple roots.

5.2 Single Parameter Location Models

5.2.1 Shifted Information Functions

Let $X_1, X_2, \cdots, X_n$ be independently and identically distributed with location model $f(x-\theta_0)$. The score estimating function is:

$$\sum_{i=1}^{n} g(X_i - \theta) = 0 \tag{5.1}$$

where $g(x) = -f'(x)/f(x)$.

For the true value θ_0, we define the *shifted information function* as

$$I(\theta_0, \theta) = \frac{[E_{\theta_0} g'(X-\theta)]^2}{E_{\theta_0} g^2(X-\theta)} \tag{5.2}$$

which is Godambe information at θ_0 when $\theta = \theta_0$ (see Godambe, 1960).
Assume that for all $t \in R$,

$$\lim_{|x|\to\infty} \frac{f(x+t)f'(x)}{f(x)} = 0 \tag{5.3}$$

which is satisfied by regular distributions including Cauchy distribution and normal distribution. In general, when the function $f(x)$ is in the form of:

$$f(x) \sim p(x)\exp(-kx^{\alpha}) \quad \text{as} \quad |x| \to \infty$$
$$f'(x) \sim q(x)\exp(-kx^{\alpha}) \quad \text{as} \quad |x| \to \infty$$

where $\alpha > 0$, $p(x) \sim x^{\beta}$, $q(x) \sim x^{\gamma}$, $-\infty < \beta, \gamma < \infty$, or $\alpha = 0$, $\gamma < 0$, then the equation (5.3) holds. The normal distribution is in the first case, but the Cauchy distribution is in the second case. There are no exceptions to (5.3) among distributions commonly used in statistical practice.

Proposition 5.1: Under condition (5.3), $I(\theta_0, \theta_0) \geq I(\theta_0, \theta)$ for all θ.

Proof: Since

$$\begin{aligned} E_{\theta_0}(g'(X-\theta)) &= \int_{-\infty}^{+\infty} g'(x-\theta) f(x-\theta_0) dx \\ &= \int_{-\infty}^{+\infty} g'(y) f(y+t) dy \quad t = \theta - \theta_0 \\ &= -\frac{f(y+t)f'(y)}{f(y)}|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} g(y) f'(y+t) dy \\ &= \int_{-\infty}^{\infty} \frac{f'(y)f'(y+t)}{f(y)} dy \end{aligned}$$

and

$$\begin{aligned} E_{\theta_0} g^2(X-\theta) &= \int_{-\infty}^{+\infty} g^2(x-\theta) f(x-\theta_0) dx \\ &= \int_{-\infty}^{+\infty} (\frac{f'(y)}{f(y)})^2 f(y+t) dy \quad t = \theta - \theta_0 \end{aligned}$$

In order to prove that $I(\theta_0, \theta_0) \geq I(\theta_0, \theta)$, it is sufficient to prove that

$$\begin{aligned} K(\theta_0, \theta) &= [E_{\theta_0} g'(X-\theta_0)]^2 E_{\theta_0} g^2(X-\theta) - [E_{\theta_0} g'(X-\theta)]^2 E_{\theta_0} g^2(X-\theta_0) \\ &= [\int_{-\infty}^{\infty} \frac{(f'(y))^2}{f(y)} dy]^2 \int_{-\infty}^{+\infty} (\frac{f'(y)}{f(y)})^2 f(y+t) dy \\ &\quad -[\int_{-\infty}^{\infty} \frac{f'(y)f'(y+t)}{f(y)} dy]^2 \int_{-\infty}^{+\infty} \frac{(f'(y))^2}{f(y)} dy \\ &\geq 0 \end{aligned}$$

Using the Cauchy-Schwartz inequality,

$$\begin{aligned} & [\int_{-\infty}^{\infty} \frac{f'(y)f'(y+t)}{f(y)} dy]^2 \\ &= [\int_{-\infty}^{\infty} \frac{f'(y)f'(y+t)}{f(y)f(y+t)} f(y+t) dy]^2 \\ &\leq \int_{-\infty}^{+\infty} (\frac{f'(y)}{f(y)})^2 f(y+t) dy \int_{-\infty}^{+\infty} (\frac{f'(y+t)}{f(y+t)})^2 f(y+t) dy \\ &= \int_{-\infty}^{+\infty} (\frac{f'(y)}{f(y)})^2 f(y+t) dy \int_{-\infty}^{\infty} \frac{(f'(y))^2}{f(y)} dy. \end{aligned}$$

Thus $K(\theta_0, \theta) \geq 0$, that is, $I(\theta_0, \theta) \geq I(\theta_0, \theta)$. This completes our proof.

To use this inequality as a method for a root selection, we need to find an estimate of the shifted information function $I(\theta_0, \theta)$, where θ_0 is the true value of the parameter. Naturally, we can use the sample mean as an estimate of the corresponding mean, that is,

$$\hat{I}_n(\theta) = \frac{(\sum_{i=1}^n g'(x_i - \theta))^2}{n \sum_{i=1}^n g^2(x_i - \theta)} \tag{5.4}$$

can be regarded as an estimate of the shifted information function. Based on the above result, we choose the root which maximizes the estimated shifted information function as the estimator of the location parameter θ. It will be shown that this method can be extended to a higher dimension and more general estimating functions.

5.2.2 General Estimating Functions

In the last section, we have shown that the shifted information is maximized when $\theta = \theta_0$. In the following, we will argue that this idea can be extended to more general estimating functions, though it lacks the rigorous proof. Let $g(x)$ be a function with a continuous derivative, which satisfies conditions (A), (B) and (C) in Chapter 6. Furthermore, we assume that $g(x)$ satisfies

$$\lim_{|x| \to \infty} \frac{g'(x)}{g(x)} = 0 \tag{5.5}$$

then

$$\begin{aligned} I(\theta_0, \theta) &= \frac{[\int_{-\infty}^{+\infty} g'(x - \theta) f(x - \theta_0) dx]^2}{\int_{-\infty}^{+\infty} g^2(x - \theta) f(x - \theta_0) dx} \\ &= \frac{[\int_{-\infty}^{+\infty} g'(x + t) f(x) dx]^2}{\int_{-\infty}^{+\infty} g^2(x + t) f(x) dx} \end{aligned}$$

$$\leq \frac{\int_{-\infty}^{+\infty}(\frac{g'(x+t)}{g(x+t)})^2 f(x)dx \int_{-\infty}^{+\infty}(g(x+t))^2 f(x)dx}{\int_{-\infty}^{+\infty} g^2(x+t)f(x)dx}$$

$$= \int_{-\infty}^{+\infty}(\frac{g'(x+t)}{g(x+t)})^2 f(x)dx$$

Under suitable regularity conditions, we can exchange the order of integration and limit, thus,

$$\lim_{|\theta|\to\infty} I(\theta_0,\theta) \leq \lim_{|t|\to\infty} \int_{-\infty}^{+\infty}(\frac{g'(x+t)}{g(x+t)})^2 f(x)dx = 0.$$

Assumption (5.5) implies the above result. Under conditions (A), (B) and (C) in Chapter 6, the probability that $G_n(\theta) = 0$ has a unique root on any finite interval approaches to 1 when the sample size is large. Furthermore, this root is close to the true value with high probability. Therefore, when $G_n(\theta) = 0$ has multiple roots, say $\hat{\theta}_0, \hat{\theta}_1, \cdots, \hat{\theta}_k$, all roots but one ($\hat{\theta}_0$) lie outside the given finite interval. When the finite interval is chosen to be wide enough, based on the above discussion, the corresponding shifted information $I(\theta_0,\hat{\theta}_i)(i = 1,\cdots,k)$ is small comparing with $I(\theta_0,\hat{\theta}_0)$, which is closer to $I(\theta_0,\theta_0)$. Thus it is reasonable to choose the root which has the largest estimated shifted information as the estimator of parameter.

5.3 Invariant Information Functions

Assume that the random variable X has the probability density function $f(x;\theta)$. Under the reparameterization $\theta = h(\alpha)$, where $h(\alpha)$ is a strictly monotone function with a continuous derivative, the probability density function becomes $f_h(x;\alpha) = f(x;h(\alpha))$, and the corresponding score function is

$$g_h(x;\alpha) = \frac{\partial \ln f_h(x;\alpha)}{\partial \alpha}$$

Define the information function as :

$$I_h(\alpha_0, \alpha) = \frac{\{E_{\alpha_0}[\frac{\partial}{\partial \alpha}(\frac{1}{h'(\alpha)} g_h(x;\alpha))]\}^2}{E_{\alpha_0}[(g_h(x;\alpha))^2]} \tag{5.6}$$

for any fixed α_0. Then we have the following result.

Proposition 5.2. The information function defined in (5.6) is independent of h in the sense that $I_h(h^{-1}(\theta_0), h^{-1}(\theta))$ is independent of h.

Proof: Since

$$\begin{aligned} g_h(x;\alpha) &= \frac{\partial \ln f_h(x;\alpha)}{\partial \alpha} = \frac{\partial \ln f(x;h(\alpha))}{\partial \alpha} \\ &= \frac{\partial \ln f(x;h(\alpha))}{\partial \theta} \frac{\partial \theta}{\partial \alpha} = h'(\alpha) g(x;h(\alpha)) \end{aligned}$$

where $g(x;\theta)$ is the score function corresponding to $f(x;\theta)$, that is,

$$g(x;\theta) = \frac{\partial \ln f(x;\theta)}{\partial \theta}$$

thus

$$\frac{\partial}{\partial \alpha}(\frac{1}{h'(\alpha)} g_h(x;\alpha)) = \frac{\partial g(x;h(\alpha))}{\partial \alpha} = h'(\alpha) \frac{\partial g(x;h(\alpha))}{\partial \theta}$$

$$\begin{aligned} \{E_{\alpha_0}[\frac{\partial}{\partial \alpha}(\frac{1}{h'(\alpha)} g_h(x;\alpha))]\}^2 &= (h'(\alpha))^2 [E_{\alpha_0}(\frac{\partial g(x;h(\alpha))}{\partial \theta})]^2 \\ &= (h'(\alpha))^2 [\int_R \frac{\partial g(x;h(\alpha))}{\partial \theta} f(x;h(\alpha_0)) dx]^2 \\ &= (h'(\alpha))^2 [\int_R \frac{\partial g(x;\theta)}{\partial \theta} f(x;\theta_0) dx]^2 \\ &= (h'(\alpha))^2 [E_{\theta_0}(\frac{\partial g(x;\theta)}{\partial \theta})]^2 \end{aligned}$$

where $\theta = h(\alpha)$, $\theta_0 = h(\alpha_0)$. Similarly,

$$E_{\alpha_0}[(g_h(x;\alpha))^2] = (h'(\alpha))^2 E_{\theta_0}[(g(x;\theta))^2]$$

therefore

$$I_h(\alpha_0,\alpha) = \frac{[E_{\theta_0}(\frac{\partial g(x;\theta)}{\partial \theta})]^2}{E_{\theta_0}[(g(x;\theta))^2]} = I(\theta_0,\theta)$$

which is independent of h.

Remarks:

(1) Based on Proposition 5.1, for any distribution which can be transformed to a location model, we can define its information function as in (5.6). Furthermore, we can conclude that

$$I_h(\alpha_0,\alpha_0) \geq I_h(\alpha_0,\alpha) \qquad \text{for all } \alpha$$

(2) In particular, for the scale model:

$$f_s(x;\alpha) = \frac{1}{\alpha} f_0(\frac{x}{\alpha}) \qquad x > 0$$

where $\alpha > 0$. Let $\alpha = e^{\theta}$, $x = e^{y}$, it can be transformed into

$$f_l(y;\theta) = f_0(e^{y-\theta})e^{y-\theta} \qquad y \in R$$

when we define

$$I_s(\alpha_0,\alpha) = \frac{E_{\alpha_0}[\frac{\partial}{\partial \alpha}(\alpha g_s(X;\alpha))]^2}{E_{\alpha_0}[g_s^2(X;\alpha)]}$$

where

$$g_s(x;\alpha) = \frac{\partial \ln f_s(x;\alpha)}{\partial \alpha}$$

then

$$I_s(\alpha_0, \alpha_0) \geq I_s(\alpha_0, \alpha) \qquad \alpha > 0$$

holds. The following example demonstrates this result.

Example 5.1. Consider the exponential distribution which has the probability density function:

$$f(x, \alpha) = \frac{1}{\alpha} e^{-\frac{x}{\alpha}} \qquad x > 0$$

where $\alpha > 0$. Assume the true value $\alpha_0 = 1$. In this case, the score estimating function is

$$g(x, \alpha) = \frac{\partial \ln f(x, \alpha)}{\partial \alpha} = \frac{x - \alpha}{\alpha^2}$$

thus

$$E_{\alpha_0}[g^2(X, \alpha)] = \frac{\alpha^2 - 2\alpha + 2}{\alpha^4}$$
$$E_{\alpha_0}[\frac{\partial}{\partial \alpha}(\alpha g(x; \alpha))] = -\frac{1}{\alpha^2}$$

Then

$$I(\alpha_0, \alpha) = \frac{1}{\alpha^2 - 2\alpha + 2} = \frac{1}{1 + (\alpha - 1)^2}$$

which is maximized at the true value $\alpha_0 = 1$.

5.4 Multiparameter Location Models

In the following, I will discuss the multiparameter location models. Assume that $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ are identically and independently m-dimensional random vectors from

$f(\boldsymbol{x}-\boldsymbol{\theta}_0)$. The estimating functions are

$$G(\boldsymbol{x},\boldsymbol{\theta}) = \begin{pmatrix} g_1(\boldsymbol{x}-\boldsymbol{\theta}_0) \\ g_2(\boldsymbol{x}-\boldsymbol{\theta}_0) \\ \vdots \\ g_m(\boldsymbol{x}-\boldsymbol{\theta}_0) \end{pmatrix} = \mathbf{0} \tag{5.7}$$

In the following discussion, assume that the following condition is satisfied:

$$\text{(A)} \qquad \lim_{|\boldsymbol{x}|\to\infty} f(\boldsymbol{x}+\boldsymbol{t})g(\boldsymbol{x}) = 0 \qquad \text{for any } \boldsymbol{t}$$

Let us denote

$$A(\boldsymbol{\theta}_0,\boldsymbol{\theta}) = E_{\boldsymbol{\theta}_0}(\dot{G}) = (a_{ij}(\boldsymbol{\theta}_0,\boldsymbol{\theta}))_{m\times m}$$

where

$$\begin{aligned} a_{ij}(\boldsymbol{\theta}_0,\boldsymbol{\theta}) &= E_{\boldsymbol{\theta}_0}[\frac{\partial g_i(\boldsymbol{x}-\boldsymbol{\theta})}{\partial\theta_j}] = -\int_{R^m} \frac{\partial g_i(\boldsymbol{x}-\boldsymbol{\theta})}{\partial x_j} f(\boldsymbol{x}-\boldsymbol{\theta}_0)d\boldsymbol{x} \\ &= -\int_{R^m} \frac{\partial g_i(\boldsymbol{x})}{\partial x_j} f(\boldsymbol{x}+\boldsymbol{t})d\boldsymbol{x} \qquad \boldsymbol{t}=\boldsymbol{\theta}-\boldsymbol{\theta}_0 \\ &= \int_{R^m} g_i(\boldsymbol{x})\frac{\partial f(\boldsymbol{x}+\boldsymbol{t})}{\partial x_j}d\boldsymbol{x} \end{aligned}$$

vand

$$B(\boldsymbol{\theta}_0,\boldsymbol{\theta}) = E_{\boldsymbol{\theta}_0}(GG') = (b_{ij}(\boldsymbol{\theta}_0,\boldsymbol{\theta}))_{m\times m}$$

where

$$\begin{aligned} b_{ij}(\boldsymbol{\theta}_0,\boldsymbol{\theta}) &= E_{\boldsymbol{\theta}_0}[g_i(\boldsymbol{x}-\boldsymbol{\theta})g_j(\boldsymbol{x}-\boldsymbol{\theta})] \\ &= \int_{R^m} g_i(\boldsymbol{x})g_j(\boldsymbol{x})f(\boldsymbol{x}+\boldsymbol{t})d\boldsymbol{x} \qquad \boldsymbol{t}=\boldsymbol{\theta}-\boldsymbol{\theta}_0 \end{aligned}$$

Also denote

$$c_j = \int_{R^m} [\frac{\partial f(\boldsymbol{x})}{\partial x_j}]^2 \frac{1}{f(\boldsymbol{x})} d\boldsymbol{x}$$

By using Cauchy-Schwartz inequality, under condition (A), we have

$$\begin{aligned} (a_{ij}(\boldsymbol{\theta}_0, \boldsymbol{\theta}))^2 &= [\int_{R^m} g_i(\boldsymbol{x}) \frac{\partial f(\boldsymbol{x}+\boldsymbol{t})}{\partial x_j} d\boldsymbol{x}]^2 \qquad (5.8) \\ \leq & \int_{R^m} g_i^2(\boldsymbol{x}) f(\boldsymbol{x}+\boldsymbol{t}) d\boldsymbol{x} \int_{R^m} [\frac{\partial f(\boldsymbol{x}+\boldsymbol{t})}{\partial x_j}]^2 \frac{1}{f(\boldsymbol{x}+\boldsymbol{t})} d\boldsymbol{x} \\ = & \ b_{ii}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) c_j. \end{aligned}$$

As an extension of the shifted information function for the single parameter case, we define the *shifted information function* for multiparameter as

$$I(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = \frac{\sum_{i=1}^m \{E_{\boldsymbol{\theta}_0} [\frac{\partial g_i(\boldsymbol{x}-\boldsymbol{\theta})}{\partial \theta_i}]\}^2}{\{\sum_{i=1}^m [E_{\boldsymbol{\theta}_0} g_i^2(\boldsymbol{x}-\boldsymbol{\theta})]^2\}^{\frac{1}{2}}} = \frac{\sum_{i=1}^m a_{ii}^2(\boldsymbol{\theta}_0, \boldsymbol{\theta})}{(\sum_{i=1}^m b_{ii}^2(\boldsymbol{\theta}_0, \boldsymbol{\theta}))^{\frac{1}{2}}} \qquad (5.9)$$

When the estimating functions are the score estimating functions, that is,

$$g_i(\boldsymbol{x}-\boldsymbol{\theta}) = \frac{\partial \ln f(\boldsymbol{x}-\boldsymbol{\theta})}{\partial \theta_i} = -\frac{\partial f(\boldsymbol{x}-\boldsymbol{\theta})}{\partial \theta_i} \frac{1}{f(\boldsymbol{x}-\boldsymbol{\theta})}$$

thus $a_{ii}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) = -b_{ii}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) = -c_i$, for $i = 1, 2, \cdots, m$.

Proposition 5.3. Under conditions (A), for the score estimating functions, $I(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) \geq I(\boldsymbol{\theta}_0, \boldsymbol{\theta})$.

By (5.8), we have

$$\sum_{i=1}^m a_{ii}^2(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \leq \sum_{i=1}^m b_{ii}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) c_i \leq [\sum_{i=1}^m b_{ii}^2(\boldsymbol{\theta}_0, \boldsymbol{\theta})]^{\frac{1}{2}} [\sum_{j=1}^m c_i^2]^{\frac{1}{2}}.$$

Thus.

$$\begin{aligned}
& [\sum_{i=1}^{m} a_{ii}^2(\boldsymbol{\theta}_0, \boldsymbol{\theta})][\sum_{i=1}^{m} b_{ii}^2(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)]^{\frac{1}{2}} \\
\leq\ & [\sum_{i=1}^{m} b_{ii}^2(\boldsymbol{\theta}_0, \boldsymbol{\theta})]^{\frac{1}{2}}[\sum_{i=1}^{m} c_i^2]^{\frac{1}{2}}[\sum_{i=1}^{m} b_{ii}^2(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)]^{\frac{1}{2}} \\
=\ & [\sum_{i=1}^{m} b_{ii}^2(\boldsymbol{\theta}_0, \boldsymbol{\theta})]^{\frac{1}{2}}[\sum_{i=1}^{m} a_{ii}^2(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)]^{\frac{1}{2}}[\sum_{i=1}^{m} a_{ii}^2(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)]^{\frac{1}{2}} \\
=\ & [\sum_{i=1}^{m} b_{ii}^2(\boldsymbol{\theta}_0, \boldsymbol{\theta})]^{\frac{1}{2}}[\sum_{i=1}^{m} a_{ii}^2(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)].
\end{aligned}$$

Therefore,

$$\frac{\sum_{i=1}^{m} a_{ii}^2(\boldsymbol{\theta}_0, \boldsymbol{\theta})}{[\sum_{i=1}^{m} b_{ii}^2(\boldsymbol{\theta}_0, \boldsymbol{\theta})]^{\frac{1}{2}}} \leq \frac{\sum_{i=1}^{m} a_{ii}^2(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)}{[\sum_{i=1}^{m} b_{ii}^2(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)]^{\frac{1}{2}}}.$$

That is, $I(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \leq I(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)$ for all $\boldsymbol{\theta}$.

Similar to the single parameter case, we can extend the *shift information function* (5.9) to multiparameter transformation models.

5.5 Simulation Results

Let us consider the Cauchy location model. In this case, the score estimating equation is

$$G(\theta, \boldsymbol{x}) = \sum_{i=1}^{n} \frac{2(x_i - \theta)}{1 + (x_i - \theta)^2}. \tag{5.10}$$

We will use five different methods to choose the best root from multiple roots:

- **Likelihood method**: Choose the root which has the maximum value of likelihood function as the estimator of parameter.

- **Median method**: Choose the root which is the closest to sample median.
- **Root intensity method**: Choose the root which has the maximum value of estimated root intensity function. In order to get a more robust estimation, we use the trimmed method. Here we discard the observation which has the largest absolute value, then use the normal approximation method discussed in Section 4.3.1 to estimate the root intensity function.
- **Shifted information method**: Choose the root which has the maximum value of estimated shifted information function. For the same reason, we discard the observation which has the largest absolute value, then use the method discussed in Section 5.2 to estimate the shifted information.
- **True value method:** The above four methods are based only on the sample. In this simulation, we have known that the true value of parameter is zero. Thus we choose the root which is closest to the true value as the estimate of the parameter. Then we will compare this estimate with the choices based on the other four methods.

We use Splus to implement the simulation for the standard Cauchy distribution. In each trial, Splus generates a sample with size 10, the C program calculates all roots of the corresponding score estimating equation (5.10) using the bisection method, then we get the corresponding values of the likelihood function, the distances to the sample median, the distance to the true value, the values of the estimated root intensity function, the values of the estimated shifted information function. In order to get a more robust estimation, we use the trimmed method for estimations of the root intensity function and the shifted information function. Here we discard the observation which has the largest absolute value. Then use the normal approximation method discussed in subsection 4.3.1 to estimate the root intensity function.

The estimated shifted information function can also be obtained by using sample mean to estimate the corresponding mean. Finally, we use the above methods to choose the best root.

Based on 2000 trials, there are at least two roots in 591 trials. We find that in 1959 trials, the root which maximizes the estimated root intensity function is the one closest to the true value; in 1951 trials, the root corresponding to the largest estimated shifted information function is the same as the one closest to the true value; in 1981 trials, the global maximum of likelihood is just the root closest to the true value; while in 1987 trials, the root which is the closest to the sample median and the one closest to the true value are the same. We may consider the choice of *true value method* as the 'right' choice. This simulation also indicates that that in 1962 trials, the root which maximizes the estimated root intensity function is the global maximum of likelihood; in 1954 trials, the root corresponding to the largest estimated shifted information function is the global maximum of likelihood; while in 1989 trials, the root which is the closest to sample median is the global maximum of likelihood.

The multiple root case is of special interest. The following table compares the root intensity method and the shifted information method with the likelihood method and the 'right' method.

Table 5.1: Comparison between Different Methods

	Likelihood		*'Right'*	
Root Intensity	553	93.6%	550	93.1%
Shifted Information	545	92.2%	542	91.7%

The above simulation results show that among the multiple root cases, at 93.6% of times, the likelihood and root intensity methods gives the same choices; at 92.2% of times, the likelihood and the shifted information methods is consistent; at 93.1% of times, the root intensity method gives the 'right' choice; while at 91.7% of times, the selection of roots based on the shifted information method is the same as the 'right' choice.

Since the likelihood method is widely accepted method, and the 'right' method is a natural choice, the simulation results show that both methods are reliable methods. However, the likelihood method can be only used in specific parametric models, where we know the form of distributions except for the parameters. In contrast to the likelihood method, the root intensity method can be applied in a very wide class of estimating functions. The shifted information method can also used to a more general class of estimating function for location models or related transformation models.

5.6 Discussion

In this Chapter, we have considered an information-based criterion for root selection in the location models. This method is based on the fact that the shifted information function for the score estimating function gets its maximum at the true value. This method can be extended to any transformation models which can be transformed into location models. In addition to the score estimating functions, this method can also be applied to some more general estimating functions. The higher dimension case is also discussed in Section 5.4.

Chapter 6

Number of Roots for Large Samples

6.1 Preliminaries

In this chapter, we will discuss the number of the roots of estimating functions for large samples. Assume that $X_1, X_2, \cdots, X_n$ are identically and independently distributed random variables with density function $f(x, \theta_0)$, we will show that under some regularity conditions, an estimating function in the form of

$$G(\theta) = \sum_{i=1}^{n} g(X_i, \theta)$$

has a unique solution in any fixed interval which includes θ_0 for large samples.

6.2 Review

In this section, we will introduce some related concepts and results (Billingsley, 1968). Let X be a mapping from (Ω, β, P) into $C[0,1]$, which is the set of all continuous functions on $[0,1]$. That is, for every $\omega \in \Omega$, $X(\omega)$ is a continuous function on $[0,1]$.

Definition 6.1 Let $C[0,1]$ be endowed with the supremum norm and its associated topology. We say $\{X_n\}$ is *tight* when $\{P_n\}$ is tight, where $\{P_n\}$ is the distribution of $\{X_n\}$. That is, for any positive ϵ, there exists a compact set K such as $P(X_n \in K) > 1 - \epsilon$ for all n.

Definition 6.2 The *modulus of continuity* of an element x of $C[0,1]$ is defined by

$$w_x(\delta) = w(x,\delta) = \sup_{|s-t|<\delta} |x(s) - x(t)| \quad 0 < \delta \leq 1. \tag{6.1}$$

Then we have the following result (Billingsley, 1968, p55-58):

Theorem 6.1 The sequence $\{X_n\}$ is tight if and only if the two conditions hold:

(i) For each positive η, there exists an α such that

$$P\{|X_n(0)| > \alpha\} \leq \eta \qquad n \geq 1.$$

(ii) For each positive ϵ and η, there exists a δ, $0 < \delta < 1$, and an integer n_0 such that

$$P\{w(X_n,\delta) \geq \epsilon\} \leq \eta \qquad n \geq n_0.$$

In section 6.3, we will use the following important **Tightness Criterion** (Billingsley, 1968, p95):

Theorem 6.2 The sequence $\{X_n\}$ is tight if it satisfies the two conditions:

(i) The sequence $\{X_n(0)\}$ is tight on the line, that is, for each positive η, there exists an α such that

$$P\{|X_n(0)| > \alpha\} \leq \eta \qquad n \geq 1$$

(ii) There exists constants $\gamma \geq 0$ and $\alpha > 1$ and a nondecreasing, continuous function F on $[0,1]$ such that

$$P\{|X_n(t_2) - X_n(t_1)| \geq \lambda\} \leq \frac{1}{\lambda^\gamma}|F(t_2) - F(t_1)|^\alpha \tag{6.2}$$

holds for all t_1, t_2, n and all positive λ.

Remarks:

1. From Chebyshev's inequality, the moment condition

$$E\{|X_n(t_2) - X_n(t_1)|^\gamma\} \leq |F(t_2) - F(t_1)|^\alpha \tag{6.3}$$

implies (6.2). Especially, we may take $\gamma = 2$ and $\alpha = 2$ and $F(t) = kt$ $(k > 0)$.

2. Obviously, all the results hold when $[0,1]$ is replaced by any closed interval $[a,b]$.

6.3 Convergence Results

Proposition 6.1. Assume that for all $\theta \in [a,b]$, $X_1(\theta), \cdots, X_n(\theta)$ are identically and independently distributed random variables with the properties:

(i) $E[X_1(\theta)] = 0$ and $\text{var}(X_1(\theta)) = \sigma^2(\theta)$ for all $\theta \in [a, b]$.

(ii) $E[(X_1(\theta_1) - X_1(\theta_2))^2] \leq C(\theta_1 - \theta_2)^2$ for all $\theta_1, \theta_2 \in [a, b]$.

Let

$$Y_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i(\theta)$$

then $\{Y_n\}$ is tight in $C[a, b]$.

Proof: For any $\theta_0 \in [a, b]$, by Chebyshev's inequality, for any $\epsilon > 0$, when M is large enough,

$$P\{|Y_n(\theta_0)| > M\} \leq \frac{\text{var}(Y_n(\theta_0))}{M^2} = \frac{\sigma^2(\theta_0)}{M^2} < \epsilon \quad \text{for all } n$$

That is, $\{Y_n(\theta_0)\}$ is tight on the line. Furthermore,

$$\begin{aligned} & E\{[Y_n(\theta_1) - Y_n(\theta_2)]^2\} = \text{var}(Y_n(\theta_1) - Y_n(\theta_2)) \\ = \; & \frac{1}{n} \sum_{i=1}^{n} \text{var}(X_i(\theta_1) - X_i(\theta_2)) = E[(X_1(\theta_1) - X_1(\theta_2))^2] \\ \leq \; & C(\theta_1 - \theta_2)^2. \end{aligned}$$

It follows from Theorem 6.2 that $\{Y_n\}$ is tight on $C[a, b]$.

Proposition 6.2. Under the conditions of Proposition 6.1, for any positive ϵ and η, there exists n_0 such that

$$P\left\{ \sup_{\theta_1, \theta_2 \in [a,b]} |\bar{X}_n(\theta_1) - \bar{X}_n(\theta_2)| > \epsilon \right\} < \eta \quad n \geq n_0 \tag{6.4}$$

where

$$\bar{X}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} X_i(\theta) = \frac{1}{\sqrt{n}} Y_n(\theta)$$

that is $\sup_{\theta_1,\theta_2\in[a,b]} |\bar{X}_n(\theta_1) - \bar{X}_n(\theta_2)|$ converges to 0 in probability.

Proof: Let $T = b-a$, from Proposition 6.1, for any positive ϵ and η, there exists m independent of n, such that

$$P\left\{ \sup_{|\theta_1-\theta_2|<\frac{T}{m}} |Y_n(\theta_1) - Y_n(\theta_2)| > \epsilon \right\} < \eta \qquad n \geq n_1. \tag{6.5}$$

Since

$$\sup_{\theta_1,\theta_2\in[a,b]} |Y_n(\theta_1) - Y_n(\theta_2)| \leq m \sup_{|\theta_1-\theta_2|<\frac{T}{m}} |Y_n(\theta_1) - Y_n(\theta_2)| \tag{6.6}$$

when taking $n_0 = \max(m^2, n_1)$, from (6.5) and (6.6), we have: for $n \geq n_0$,

$$\begin{aligned} & P\left\{ \sup_{\theta_1,\theta_2\in[a,b]} |\bar{X}_n(\theta_1) - \bar{X}_n(\theta_2)| > \epsilon \right\} \\ = \; & P\left\{ \sup_{\theta_1,\theta_2\in[a,b]} |Y_n(\theta_1) - Y_n(\theta_2)| > \sqrt{n}\epsilon \right\} \\ \leq \; & P\left\{ m \sup_{|\theta_1-\theta_2|<\frac{T}{m}} |Y_n(\theta_1) - Y_n(\theta_2)| > m\epsilon \right\} \\ < \; & \eta. \end{aligned}$$

This completes the proof.

6.4 Main Result

Let $X_1, X_2, \cdots, X_n$ be identically and independently distributed as X with density function $f(x,\theta_0)$, and $g(x,\theta)$ have a continuous derivative with respect to θ. We consider the following estimating function:

$$G_n(\theta) = \sum_{i=1}^{n} g(X_i,\theta) = 0 \tag{6.7}$$

and assume that the following conditions hold:

(A): For any $\theta_1, \theta_2 \in [-K, K]$,

$$E_{\theta_0}\{[g(X,\theta_1) - g(X,\theta_2)]^2\} \leq C_1(\theta_1 - \theta_2)^2 \tag{6.8}$$

$$E_{\theta_0}\left\{\left[\frac{\partial g(X,\theta_1)}{\partial\theta} - \frac{\partial g(X,\theta_2)}{\partial\theta}\right]^2\right\} \leq C_2(\theta_1 - \theta_2)^2. \tag{6.9}$$

(B): $\theta_0 \in (-K, K)$ and

$$E_{\theta_0} g(X,\theta_0) = 0 \tag{6.10}$$

$$E_{\theta_0} g(X,\theta) = \mu(\theta) \neq 0 \quad \text{for all} \quad \theta \neq \theta_0 \tag{6.11}$$

(C):

$$E_{\theta_0}\left[\frac{\partial g(X,\theta_0)}{\partial\theta}\right] \neq 0 \tag{6.12}$$

Note that the above conditions are satisfied easily. For instance, when $E_{\theta_0}[\partial g(X,\theta)/\theta]$ and $E_{\theta_0}[\partial^2 g(X,\theta)/\theta^2]$ are bounded, the condition (A) holds. Under the above assumptions, we can derive the following main result.

Theorem 6.3 Under the conditions (A),(B) and (C), the probability that $G_n(\theta) = 0$ on $[-K, K]$ has a unique root tends to 1 as n approaches to ∞.

In order to prove the above theorem, we first state two Lemmas:

Lemma 6.1. If

(i) $Y_n(\theta_0) \to 0$ in probability for some $\theta_0 \in [a, b]$;

(ii) $\sup_{\theta\in[a,b]} |Y_n(\theta) - Y_n(\theta_0)| \to 0$ in probability.

Then for any positive ϵ, η, there exists n_0 such that $n \geq n_0$

$$P\left\{\sup_{\theta\in[a,b]} |Y_n(\theta)| > \epsilon\right\} < \eta \tag{6.13}$$

That is, $\sup_{\theta\in[a,b]} |Y_n(\theta)| \to 0$ in probability.

Proof: For any $\epsilon > 0$ and $\eta > 0$, since

$$\begin{aligned}
& P\left\{\sup_{\theta\in[a,b]} |Y_n(\theta)| > \epsilon\right\} \\
\leq\ & P\left\{\sup_{\theta\in[a,b]} |Y_n(\theta) - Y_n(\theta_0)| + |Y_n(\theta_0)| > \epsilon\right\} \\
=\ & P\left\{\sup_{\theta\in[a,b]} |Y_n(\theta) - Y_n(\theta_0)| + |Y_n(\theta_0)| > \epsilon, |Y_n(\theta_0| > \frac{\epsilon}{2}\right\} \\
& + P\left\{\sup_{\theta\in[a,b]} |Y_n(\theta) - Y_n(\theta_0)| + |Y_n(\theta_0)| > \epsilon, |Y_n(\theta_0| \leq \frac{\epsilon}{2}\right\} \\
\leq\ & P\left\{|Y_n(\theta_0| > \frac{\epsilon}{2}\right\} + P\left\{\sup_{\theta\in[a,b]} |Y_n(\theta) - Y_n(\theta_0)| > \frac{\epsilon}{2}\right\} \\
\to\ & 0
\end{aligned}$$

there exists a n_0 such that $n \geq n_0$,

$$P\left\{\sup_{\theta\in[a,b]} |Y_n(\theta)| > \epsilon\right\} < \eta.$$

This completes the proof.

Lemma 6.2. If $\lim P(A_n) = 1$ and $\lim P(B_n) = 1$, then $\lim P(A_n B_n) = 1$.

This proof is standard.

We also introduce some notations:

$$G_n(\theta) = \sum_{i=1}^{n} g(X_i; \theta)$$

$$\overline{G}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(X_i;\theta)$$

$$\overline{G}_n^*(\theta) = \frac{1}{n}\sum_{i=1}^{n}[g(X_i;\theta) - \mu(\theta)] = \overline{G}_n(\theta) - \mu(\theta)$$

$$\overline{G}_{1n}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial g(X_i;\theta)}{\partial\theta}$$

$$\overline{G}_{1n}^*(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial g(X_i;\theta)}{\partial\theta} - \nu(\theta)\right] = \overline{G}_{1n}(\theta) - \nu(\theta)$$

where $\nu(\theta) = E_{\theta_0}[\partial g(X,\theta)/\partial\theta]$.

Proof of Theorem 6.3: We will prove the above theorem in three steps.

Step 1: In this step, we will prove that for some positive δ,

$$\lim_{n\to\infty} P\{G_n(\theta) = 0 \text{ has at most one root in } [\theta_0 - \delta, \theta_0 + \delta]\} = 1 \qquad (6.14)$$

Without loss of generality, assume that

$$c_0 = E_{\theta_0}\left[\frac{\partial g(X,\theta_0)}{\partial\theta}\right] > 0$$

then there exists $\delta > 0$ such that

$$\nu(\theta) = E_{\theta_0}\left[\frac{\partial g(X,\theta)}{\partial\theta}\right] > \frac{c_0}{2} \qquad \theta \in [\theta_0 - \delta, \theta_0 + \delta]$$

Consider $g_1(X,\theta) = \partial g(X,\theta)/\partial\theta - \nu(\theta)$, then $E_{\theta_0}(g_1(X,\theta)) = 0$, and

$$\begin{aligned}
& E_{\theta_0}\{[g_1(X,\theta_1) - g_1(X,\theta_2)]^2\} \\
= \; & E_{\theta_0}\left\{\left[\left(\frac{\partial g(X,\theta_1)}{\partial\theta_1} - \frac{\partial g(X,\theta_1)}{\partial\theta_1}\right) - (\nu(\theta_1) - \nu(\theta_2))\right]^2\right\} \\
= \; & E_{\theta_0}\left\{\left[\frac{\partial g(X,\theta_1)}{\partial\theta_1} - \frac{\partial g(X,\theta_1)}{\partial\theta_1}\right]^2\right\} - [\nu(\theta_1) - \nu(\theta_2)]^2 \\
\leq \; & C_2(\theta_1 - \theta_2)^2
\end{aligned}$$

By applying Proposition 6.2, for any ϵ and η, there exists n_0 such that

$$P\left\{\sup_{\theta\in[\theta_0-\delta,\theta_0+\delta]}|\bar{G}^*_{1n}(\theta)-\bar{G}^*_{1n}(\theta_0)|>\epsilon\right\}<\eta \quad n\geq n_0$$

By the law of large numbers, $\bar{G}^*_{1n}(\theta_0)$ converges to 0 in probability. From Lemma 6.1, for any ϵ,

$$\lim_{n\to\infty} P\left\{\sup_{\theta\in[\theta_0-\delta,\theta_0+\delta]}|\bar{G}^*_{1n}(\theta)|>\epsilon\right\}=0$$

By taking $\epsilon = c_0/4$, then

$$\lim_{n\to\infty} P\left\{\sup_{\theta\in[\theta_0-\delta,\theta_0+\delta]}\left|\overline{G}^*_{1n}(\theta)\right|\leq\frac{c_0}{4}\right\}=1. \tag{6.15}$$

Since

$$|\overline{G}_{1n}(\theta)| = |\overline{G}^*_{1n}(\theta)+\nu(\theta)| \geq \nu(\theta)-|\overline{G}^*_{1n}(\theta)|$$

thus

$$\begin{aligned}\inf_{\theta\in[\theta_0-\delta,\theta_0+\delta]}|\overline{G}_{1n}(\theta)| &\geq \inf_{\theta\in[\theta_0-\delta,\theta_0+\delta]}(\nu(\theta)-|\overline{G}^*_{1n}(\theta)|)\\ &\geq \inf_{\theta\in[\theta_0-\delta,\theta_0+\delta]}\nu(\theta)-\sup_{\theta\in[\theta_0-\delta,\theta_0+\delta]}|\overline{G}^*_{1n}(\theta)|\\ &\geq \frac{c_0}{2}-\sup_{\theta\in[\theta_0-\delta,\theta_0+\delta]}|\overline{G}^*_{1n}(\theta)|\end{aligned}$$

therefore

$$\left\{\sup_{\theta\in[\theta_0-\delta,\theta_0+\delta]}|\overline{G}^*_{1n}(\theta)|\leq\frac{c_0}{4}\right\}\subset\left\{\inf_{\theta\in[\theta_0-\delta,\theta_0+\delta]}|\overline{G}_{1n}(\theta)|\geq\frac{c_0}{4}\right\}. \tag{6.16}$$

If $G_n(\theta)=\sum_{i=1}^n g(X_i;\theta)=0$ has at least two roots on $[\theta_0-\delta,\theta_0+\delta]$, then there exists $\theta^*\in(\theta_0-\delta,\theta_0+\delta)$ such that

$$\frac{\partial G_n(\theta^*)}{\partial\theta}=\sum_{i=1}^n\frac{\partial g(X_i;\theta^*)}{\partial\theta}=0$$

that is, $\overline{G}_{1n}(\theta^*) = 0$. Let

$$A_n = \{G_n(\theta) = 0 \quad \text{has at most one root in} \quad [\theta_0 - \delta, \theta_0 + \delta]\}$$

then

$$\left\{ \inf_{\theta \in [\theta_0 - \delta, \theta_0 + \delta]} |\overline{G}_{1n}(\theta)| \geq \frac{c_0}{4} \right\} \subset A_n. \tag{6.17}$$

From (6.15), (6.16) and (6.17), we have,

$$\lim_{n \to \infty} P(A_n) = 1. \tag{6.18}$$

Step 2: We will prove

$$\lim_{n \to \infty} P\{G_n(\theta) = 0 \quad \text{has no root in} \quad [-K, \theta_0 - \delta] \cup [\theta_0 + \delta, K]\} = 1. \tag{6.19}$$

Since

$$E_{\theta_0} g(X; \theta) = \mu(\theta) \neq 0 \qquad \text{for all} \qquad \theta \neq \theta_0$$

we may assume

$$|\mu(\theta)| \geq c_1 > 0 \qquad \text{for} \qquad \theta \in I = [-K, \theta_0 - \delta] \cup [\theta_0 + \delta, K].$$

Since

$$E_{\theta_0}\{[g(X; \theta_1) - g(X; \theta_2)]^2\} \leq C_1(\theta_1 - \theta_2)^2$$

similar to (6.15), we have

$$\lim_{n \to \infty} P\left\{ \sup_{\theta \in [-K, K]} |\overline{G}_n^*(\theta)| \leq \frac{c_1}{2} \right\} = 1. \tag{6.20}$$

When

$$\sup_{\theta \in [-K, K]} \left|\overline{G}_n^*(\theta)\right| \leq \frac{c_1}{2}$$

we have

$$\begin{aligned}
&\inf_{\theta\in I}|\overline{G}_n(\theta)| = \inf_{\theta\in I}|\mu(\theta)+\overline{G}_n^*(\theta)| \\
&\geq \inf_{\theta\in I}|\mu(\theta)| - \sup_{\theta\in I}|\overline{G}_n^*(\theta)| \\
&\geq \inf_{\theta\in I}|\mu(\theta)| - \sup_{\theta\in[-K,K]}|\overline{G}_n^*(\theta)| \\
&\geq c_1 - \frac{c_1}{2} = \frac{c_1}{2}.
\end{aligned}$$

Let

$$B_n = \{G_n(\theta) = 0 \text{ has no root in } I\}$$

then

$$\left\{\sup_{\theta\in[-K,K]}|\overline{G}_n^*(\theta)| \leq \frac{c_1}{2}\right\} \subset \left\{\inf_{\theta\in I}|\overline{G}_n(\theta)| \geq \frac{c_1}{2}\right\} \subset B_n. \tag{6.21}$$

By (6.20), we get

$$\lim_{n\to\infty} P(B_n) = 1. \tag{6.22}$$

Step 3: We will prove

$$\lim_{n\to\infty} P\{G_n(\theta) = 0 \text{ has at least one root in } [\theta_0-\delta, \theta_0+\delta]\} = 1 \tag{6.23}$$

In fact,

$$\begin{aligned}
\inf_{\theta\in[\theta_0-\delta,\theta_0+\delta]}\overline{G}_{1n}(\theta) &\geq \inf_{\theta\in[\theta_0-\delta,\theta_0+\delta]}\nu(\theta) - \sup_{\theta\in[\theta_0-\delta,\theta_0+\delta]}|\overline{G}_{1n}^*(\theta)| \\
&\geq \frac{c_0}{2} - \sup_{\theta\in[\theta_0-\delta,\theta_0+\delta]}|\overline{G}_{1n}^*(\theta)|
\end{aligned}$$

we can obtain from (6.15) and (6.16)

$$\lim_{n\to\infty} P\left\{\inf_{\theta\in[\theta_0-\delta,\theta_0+\delta]}\overline{G}_{1n}(\theta) \geq \frac{c_0}{4}\right\} = 1. \tag{6.24}$$

On the other hand, since $E_{\theta_0} g(X;\theta_0) = 0$, $\overline{G}_n(\theta_0) \to 0$ in probability, that is

$$\lim_{n\to\infty} P\left\{|\overline{G}_n(\theta_0)| < \frac{c_0\delta}{4}\right\} = 1. \tag{6.25}$$

Let

$$C_n = \{G_n(\theta) = 0 \text{ has at least one root in } [\theta_0 - \delta, \theta_0 + \delta]\}$$

we claim that

$$\left\{\inf_{\theta\in[\theta_0-\delta,\theta_0+\delta]} \overline{G}_{1n}(\theta) \geq \frac{c_0}{4}\right\} \cap \left\{|\overline{G}_n(\theta_0)| < \frac{c_0\delta}{4}\right\} \subset C_n. \tag{6.26}$$

In fact, since $\overline{G'}_n(\theta) = \overline{G}_{1n}(\theta) \geq c_0/4 > 0$, $\overline{G}_n(\theta)$ is a strictly increasing function.

(i) If $\overline{G}_n(\theta_0) = 0$, then $G(\theta_0) = 0$. In other words, $G_n(\theta) = 0$ has at least one root in $[\theta_0 - \delta, \theta_0 + \delta]$.

(ii) If $\overline{G}_n(\theta_0) > 0$, by mean value theorem for differentiation, there exists a $\theta_1^* \in (\theta_0 - \delta, \theta_0)$ such that

$$\overline{G}_n(\theta_0) - \overline{G}_n(\theta_0 - \delta) = \overline{G'}_n(\theta_1^*)\delta = \overline{G}_{1n}(\theta_1^*)\delta \geq \frac{c_0\delta}{4}$$

then $\overline{G}_n(\theta_0) < c_0\delta/4$ implies

$$\overline{G}_n(\theta_0 - \delta) < 0$$

thus there exists $\theta_1 \in (\theta_0 - \delta, \theta_0)$ such that $\overline{G}_n(\theta_1) = 0$.

(iii) Similar to (ii), if $\overline{G}_n(\theta_0) < 0$, then

$$\overline{G}_n(\theta_0 + \delta) > 0$$

therefore there exists $\theta_2 \in (\theta_0 - \delta, \theta_0)$ such that $\overline{G}_n(\theta_2) = 0$.

In all cases, $G_n(\theta) = 0$ has at least one root in $[\theta_0 - \delta, \theta_0 + \delta]$. Thus (6.23) holds. From (6.24), (6.25) and (6.26), we obtain

$$\lim_{n\to\infty} P(C_n) = 1 \tag{6.27}$$

(6.18), (6.22) and (6.27) implies $\lim_{n\to\infty} P(A_n B_n C_n) = 1$, that is, the probability that $G_n(\theta) = 0$ has unique root in [-K,K] approaches to 1 as $n \to \infty$.

Chapter 7

Summary and Future Work

In this thesis, we have discussed the multiple root problems of estimating functions and proposed two new approaches to solve these problems. One is based on the root intensity, which can be applied to more general estimating functions. For transformation models, in particular, the location models, the shifted information methods can be used. In Chapter 1, the basic theory of estimating functions was reviewed and the problem of multiple roots was put forward. In Chapter 2, some examples in statistical theory, biostatistics and economics and many different methods to solve this problem were presented. These methods include *projected likelihood ratio* methods (McLeish and Small, 1992; B. Li, 1993; Hanfelt and Liang, 1995; B. Li, 1997), *approximate one-root estimating function* methods (Kolkiewicz,1995; McLeish and Small, 1988) and *statistical information* methods (Heyde and Morton, 1998; Singh and Morton, 1999; Wang and Small, 1998). In Chapter 3, the concept of root intensity was introduced and its properties and approximation methods were discussed. The root intensity for Cauchy location models was studied in more detail. Based on root intensity, a new approach to choose the best root was devel-

oped in Chapter 4. This involves both theoretical foundation and practical methods such as the normal approximation and the bootstrap methods. These methods were also applied to two practical examples, namely, the logistic regression models with measurement error and the normal mixture models for clustering data. In Chapter 5, the shifted information was proposed for transformation models, which can be used in the root selection for transformation models, for example, Cauchy location model. Different approaches were also compared for Cauchy location models. Although the multiple root problems may appear in many cases, it can be proved that for regular estimating function, with high probability, there is only one root for any given compact set including the true value in parameter space. The mathematical proof of this result for one dimension case was given in Chapter 6.

Although we have studied the multiple root problem of estimating functions in detail, there are many issues which is still worth studying. Future work includes two aspects: one is to extend the existing methods to a more general case, another is to find some new approaches. The future work can be the following problems:

- We have proposed the root selection method based on root intensity in Chapter 4, this method can be applied to estimating function in the form of

$$G(\boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta})$$

  where $\boldsymbol{x}_i$ $(i = 1, 2, \cdots, n)$ are identically and independently distributed random vectors. Can the approach be extended to dependent random variables? For stochastic processes, more complex estimating functions can be derived in principle. For instance, in mathematical finance, diffusion processes are widely used. The typical examples are geometric Brownian models for stocks, Vasicek and CIR interest rate models (see Hull, 1997). The one-dimensional

diffusion processes defined by the following class of stochastic differential equations:

$$dX_t = b(X_t;\boldsymbol{\theta})dt + \sigma(X_t;\boldsymbol{\theta})dW_t \tag{7.1}$$
$$X_0 = x_0$$

The function σ is assumed to be positive, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset R^k$. There are many different methods to build estimating functions for this model (see Sørensen, 1997). Kessler (1997) built explicit estimating functions using differential operators; Kessler and Sørensen (1995) constructed martingale estimating functions based on eigenfunctions; McLeish and Kolkiewicz (1997) proposed estimating functions based on higher order Itô-Taylor expansions. Kloeden and Platen (1992) considered an estimating function using the normal density approximation of the transition density. On the basis of this estimating function, Bibby and Sørensen's (1995,1996) derived the martingale estimating function. In particular, when the quadratic term is neglect, the martingale estimating function is

$$G^*(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\dot{F}(X_{(i-1)\Delta};\boldsymbol{\theta})}{\phi(X_{(i-1)\Delta};\boldsymbol{\theta})}(X_{i\Delta} - F(X_{(i-1)\Delta};\boldsymbol{\theta})) \tag{7.2}$$

where

$$F(x;\boldsymbol{\theta}) = E_\theta(X_\Delta | X_0 = x) \tag{7.3}$$

and

$$\phi(x;\boldsymbol{\theta}) = Var_\theta(X_\Delta | X_0 = x) \tag{7.4}$$

It follows from Theorem 1.1 that the optimal estimating function in the class of the martingale estimating functions is of the form

$$G(\boldsymbol{\theta}) = \sum_{i=1}^{n} g_{i-1}(\boldsymbol{\theta})(X_{i\Delta} - F(X_{(i-1)\Delta};\boldsymbol{\theta})) \tag{7.5}$$

where g_{i-1} is $\mathcal{F}_{i-1}$ measurable and a continuously differentiable function of $\boldsymbol{\theta}$. Since these estimating functions usually are nonlinear functions in the parameters, they may have several roots.

- The shifted information works well in transformation models. Is it possible to define a more general information which can be applied to more general models? That is, we need to find an information function which has the property that the information function at the true parameter is distinguishable. This seems a very challenging problem.

- Once we have provided a theoretical foundation to a method, we need a good approximation to the related information function based on the given sample. Althrough the normal approximation and the bootstrap method have been shown to be useful methods, a further exploration into alternative method is still needed.

Bibliography

[1] Barendregt, L.G. and Van Pul, M.C. (1995) On the estimation of the parameters for the Littlewood model in software reliability *Statistica Neerlandica* **49** 165-184.

[2] Barndorff-Nielsen, O.E. and Cox, D.R. (1979). Edgeworth and Saddlepoint approximations with statistical applications (with discussion). *J. Roy. Statist. Soc. Ser. B* **41** 279-312.

[3] Barndorff-Nielsen, O.E. and Cox, D.R. (1989) *Asymptotic Techniques for Use in Statistics* London, Chapman & Hall.

[4] Barndorff-Nielsen, O.E. and Cox, D.R. (1994) *Inference and Asymptotics*. London, Chapman & Hall.

[5] Barnett, V.D. (1966) Evaluation of the maximum-likelihood estimator where the likelihood equation has multiple roots. *Biometrika* **52**, 151-165.

[6] Basford, K.E. and McLachlan, G.J. (1985) Likelihood estimation with normal mixture models. *Appl. Statist.* **34**, 282-289.

[7] Bibby, B.M. and Sørensen, M. (1995) Martingale estimation functions for discretely observed diffusion processes. *Bernoulli* **1**, 17-39.

[8] Bibby, B.M. and Sørensen, M. (1996) On estimation for discretely observed diffusions: A review. *Theory of Stochastic Processes* **2 (18)**, 49-56.

[9] Billingsley, P. (1968) *Convergence of Probability Measures.* New York, Wiley.

[10] Bleistein, N. and Handelsman, R.A. (1975) *Asymptotic Expansions of Integrals.* Holt, Rinehart and Winston, New York.

[11] Bruijn, N.G. (1981) *Asymptotic Methods in Analysis* (third edition). Dover Publications, Inc. New York.

[12] Crowder, M. (1986) On consistency and inconsistency of estimating equations. *Econometric Theory* **2**, 305-330.

[13] Daniels, H.E. (1954) Saddlepoint approximations in statistics. *Ann. Math. Statist.* **25** 631-650.

[14] Daniels, H.E. (1960) The asymptotic efficiency of a maximum likelihood estimator. *Proc. 4th Berkeley Symp. Math. Statist. and Prob.* **1** 151-163.

[15] Diggle, P., Liang, K.Y. and Zeger, S.L. (1994) *Analysis of Longitudinal Data.* Chapman & Hall.

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

[17] Easton, G.S. and Ronchetti, E. (1986) General saddlepoint approximations with applications to L statistics. *J. Amer. Statist. Assoc.* **81** 420-430.

[18] Geary, R.C. (1944) Extension of a theorem by Harald Cramèr on the frequency distribution of a quotient of two variables. *J. Roy. Stat. Soc.* **17**, 56-57.

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

[20] Godambe, V.P. (1985) The foundations of finite sample estimation in stochastic processes. *Biometrika* **63**, 419-428.

[21] Godambe, V.P. and Thompson, M.E. (1989) An extension of quasi-likelihood estimation (with discussion). *J. Statist. Planning Inf.* **22**, 137-172.

[22] Greene, W. (1990) Multiple roots of the Tobit log-likelihood. *Journal of Econometrics* **46** 365-380.

[23] Habbema, J.D.F., Hermnas, J. and van den Broek, K. (1974). A stepwise discriminant analysis program using density estimation. *Compstat 1974, Proc. Computational Statistics.* Vienna: Physica-Verlag, pp. 101-110.

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

[25] Heyde, C.C. (1988) Fixed sample and asymptotic optimality for classes of estimating functions. *Comtemp. Math.* **80**, 241-247

[26] Heyde, C.C. (1997) *Quasi-likelihood and Its Application: A General Approach to Optimal Parameter Estimation.* Springer-Verlag, New York.

[27] Heyde, C.C. and Morton, R. (1998) Multiple roots in general estimating equations. *Biometrika* **85** 954-959.

[28] Hull, J. (1997) *Options, Futures, and Other Derivatives* (third edition). Prentice Hall International, Inc.

[29] Hull, J. and White, A. (1988) An analysis of the bias in option pricing caused by a stochastic volatility. *Advances in Future and Options Research* **3**, 29-61.

[30] Jensen, J.L. and Wood, A.T.A. (1999) Large deviation results for minimum contrast estimators. *Ann. Inst. Statist. Math.*. To appear.

[31] Kalbfleisch, J. D. and Sprott, D.A. (1970) Applications of likelihood methods to models involving large numbers of parameters (with discussion). *J. Roy. Statist. Soc. Ser. B* **32**, 175-208.

[32] Kessler, M. (1997) Estimation of an ergodic diffusion from discrete observations. *Scand. J. Statist.* **24**, 211-219.

[33] Kessler, M and Sørensen, M. (1995) Estimating equations based on eigenfunctions for discretely observed diffusion process. Research Report No. 332, Department of Theoretical Statistics, University of Aarhus.

[34] Kloeden P.E. and Platen, E. (1992) *Numerical Solution of Stochastic Differential Equations*. Springer-Verlag, New York.

[35] Kolkiewicz, A.W. (1995) *M-estimation for Autoregression Processes with Stable Innovations*. Ph.D thesis in the University of Waterloo.

[36] LeCam, L. (1979) *Maximum Likelihood: an Introduction*. Lecture Notes in Statistics, No. 18. University of Maryland, College Park, Md.

[37] Lehmann, E.L. (1983) *Theory of Point Estimation*. Wiley, New York.

[38] Li, B (1993) A deviance function for the quasilikelihood method. *Biometrika* **80**, 741-753.

[39] Li, B. (1996) A minimax approach to consistency and efficiency for estimating equations. *Ann. Statist.* **24**, 1283-1297.

[40] Li, B. (1997) On the consistency of generalized estimating equations. In *Selected Proceedings of the Symposium on Estimating Equations* **32**, p115-136.

[41] Li, X. (1999) Value at Risk based on the Volatility, Skewness and Kurtosis. Preprint.

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

[43] Littlewood, B. (1980) Theories of software reliability: How good are they and how can they be improved? *IEEE Transactions on Software Engineering* **6**, 489-500.

[44] McCullagh, P. (1987) *Tensor Methods in Statistics*. London: Chapman and Hall.

[45] McCullagh, P and Nelder, J.A (1989) *Generalized Linear Models*, London: Chapman and Hall.

[46] McCullagh, P. (1990) Quasi-likelihood and estimating functions. *In Statistical Theory and Modeling: In Honour of Sir David Cox* (D.V. Hinkley, N. Reid and E.J. Snell, eds.). Chapman and Hall, London.

[47] McLachlan, G.J. and Basford, K.E. (1987) *Mixture models: inference and applications to clustering*. Marcel Dekker, Inc.

[48] McLeish, D.L. and Kolkiewicz, A.W. (1997) Fitting diffusion models in finance. In *Selected Proceedings of the Symposium on Estimating Equations* **32**, p327-350.

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

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

[51] Neyman, J. and Scott, E.L. (1948) Consistent estimates based on partially consistent observations. *Econometrica* **16** 1-32.

[52] Olsen, R. (1978) Note on the uniqueness of the maximum likelihood estimator of the Tobit model. *Econometrica* **46** 1211-1215.

[53] Perlman, M.D. (1983) The limiting behavior of the multiple roots of the likelihood equation. In *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on his Sixtieth Birthday*, 339-370. Academic, New York.

[54] Raper, L.R., Balkau, B., Taylor, R., Milne, B., Collins, V. and Zimmet, P. (1983) Plasma glucose distributions in two Pacific populations: the bimodality phenomenon. *Tohoku J. exp. Med.* **141**, suppl., 199-206.

[55] Reeds, J.A. (1985) Asymptotic number of roots of Cauchy location likelihood equations. *Ann. Statist.* **13** 775-784.

[56] Reid, N. (1988) Saddlepoint Methods and Statistical Inference. *Statistical Science* **3**, 213-238.

[57] Serfling, R.J. (1980) *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons.

[58] Silverman, B.W. (1978) Choosing the window width when estimating a density. *Biometrika* **65** 1-11.

[59] Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis.* Chapman & Hall.

[60] Singh, A.C. and Mantel, H.J. (1999) Minimum chi-square estimating function and the problem of choosing among multiple roots. *Proc. Amer. Statist. Assoc.* (To appear).

[61] Skovgaard, I.M. (1990) On the density of minimum contrast estimators. *Ann. Statist.* **18**, 775-784.

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

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

[64] Small, C.G. and McLeish, D.L. (1994) *Hilbert Space Methods in Probability and Statistical Inference.* John Wiley & Sons, Inc.

[65] Small, C.G. and Yang, Z. (1999) Multiple roots of estimating functions. *Canadian J. Statist.* **27** 585-598.

[66] Small, C.G., Wang, J. and Yang, Z. (1999) Eliminating multiple root problems in Estimation. (Accepted for publication in Statistical Science).

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

[68] Sørensen, M (1997) Estimating functions for discretely observed diffusions: A review. In *Selected Proceedings of the Symposium on Estimating Equations* **32**, p305-325.

[69] Tapia, R.A. and Thompson, J.R. (1978) *Nonparametric Probability Density Estimation* Baltimore: Johns Hopkins University Press.

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

[71] Wang, J. and Small, C.G. (1998) Semiparametric local likelihood functions with applications to root selection. *Working Paper* 98-10 Department of Statistics and Actuarial Science, University of Waterloo.

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

[73] Wedderburn, R.W.M. (1976) On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* **63**, 27-32.