CHAPTER 1

INTRODUCTION

1.1 General Introduction

In the applications of statistical models, especially parametric ones, to real life phenomena, practitioners are confronted by the issues of parameter estimation, testing of relevant hypotheses and construction of confidence intervals of these models. For parameter estimation, maximum likelihood estimation (MLE) is a popular technique due to its efficiency. The statistical literature contains many papers about various aspects and applications of MLE. For instance, Kalbfleisch and Sprott (1970) introduced and exemplified likelihood methods to deal with some multi parameter problems, in particular the elimination of nuisance parameters from the likelihood function so that inferences could be made. In this paper, integrated likelihoods, maximum relative likelihoods, conditional likelihoods, marginal likelihoods and second-order likelihoods are introduced and illustrated in examples. Sprott (1980) discussed further on the various pivotal quantities associated with the application of MLE to small samples to estimates a parameter in the presence of other nuisance parameters. Next, Sprott (1983) examined the application of MLE to some convolution densities (the convolution of a Poisson and a binomial distribution, a Poisson and a normal distribution, and so on). The paper claimed that MLE of the parameters in the convolution densities is numerically intractable. However, it is shown that in a large class of such densities, the number of ML equations can be reduced by one. Thus for a two-parameter family of distributions only a single equation needs to be solved iteratively.

Although MLE is an efficient method of estimation there are many difficulties associated with the method: multiples maxima, starting values, convergence criteria and rate of convergence of the numerical algorithm. Douglas (1980) discussed the problem of inherently high parameter correlation encountered with the MLE for standard contagious distributions. The high correlation of ML estimators tends to lead to mathematical complexities, and causes difficulty or even errors in their interpretation. Besides, numerical difficulties may arise when using numerical procedures to locate the estimates. Orthogonal parameterization (reparameterization) is suggested to reduce or even eliminate such undesirable feature. Cox and Reid (1987) examined the consequences of parameter orthogonality in MLE. The complete literature reviews for orthogonal parameterization are given in Section 2.2. The orthogonality of a class of discrete distributions with two unknown parameters reparameterized into the mean, μ, with respect to the remaining parameter have been considered by Willmot (1990). Since parameter orthogonality has many consequences in statistical inference, we have extended the works of Willmot by deriving the orthogonality of μ for models with more than two parameters, where the remaining parameters are regarded as nuisance parameters.

ML estimate has to be solved numerically by maximizing the likelihood or log-likelihood function when its analytical solution is intractable to obtain. There are many optimization algorithms have been developed and applied in parameter estimation. Most of the optimization methods suffer from the problem of selecting the starting values for the optimization. One of the popular choices for MLE is the iteratively re-weighted least square method via EM algorithm. Although the EM algorithm is easy to implement, it suffers from slow convergence and dependence on the starting values. In our study, we consider the simulated annealing, an optimization algorithm which are specially designed to optimize functions that are not smooth and may have multiple local optimum values, in parameter estimation. Simulated annealing is guaranteed to converge regardless of the choice of starting values although the convergence could be slow. The algorithm has also introduced random elements into the iteration process in order for the algorithm to escape from a local optimum. Enhancement of optimization by simulated annealing has been examined by various authors (see, for example, Brooks and Morgan (1995) and references therein). Smyth (2002) discussed the various optimization methods used in parameter estimation.

White (1982) has considered the MLE in the presence of model misspecification, which is so-called quasi maximum likelihood estimation (QMLE). The major concern of QML method is to draw statistical inferences under potential model misspecification. QML method is robust to specification errors compared to the traditional ML method. Many extension and refinement works have been done for the recent 20 years related to the topics of White (see, for example, Fomby and Carter, 2003). In particular, we have derived the orthogonality for a class of Poisson-convolution models. For the application of the obtained orthogonality results, a uniformly most powerful test of the mean is developed based on the asymptotic result under model misspecification. The test of mean is implemented on the convolution of Poisson and negative binomial variables.

Suppose $X_1, X_2, ..., X_n$ are independently and identically observations from a distribution $F(x)$. One important issue that we considered here is the goodness-of-fit problem. Our interest is to determine whether a given random sample can be fitted well by a probability model. In order to do so, we have to test H_0 whether the sample $x_1, x_2, ..., x_n$ comes from a population with distribution function $F(x)$. A useful review on the goodness-of-fit problem is given by D'Agostino and Stephens (1986). In addition, Stuart and Ord (1991) and Lehmann (1999) have provided the details of some of the main goodness-of-fit techniques. The classical test for this problem is the chi-square test. It was introduced by Pearson (1900). There are some advantages of using the chi-square test; for example, it is more versatile and can be used for continuous as well as discrete data, and the test statistic is easily adjusted for the case when the parameters have to be estimated. However, the power of the test is comparatively low as reported in the literature.

There is another class of goodness-of-fit statistics which are widely used and they are known as Empirical Distribution Function (EDF) statistics. The practical guide to the use of EDF statistics for goodness of fit have been discussed by Stephens (1974). EDF based goodness-of-fit tests consider a comparison between the sample EDF $F_n\left(x\right)$ and the actual distribution function $F\left(x\right)$. If the assumed model is correct, we will have $F_n\left(x\right)$ close to $F\left(x\right)$. Generally, EDF statistics are easily calculated and they are shown to be more powerful in terms of hypothesis testing.

In this thesis, we introduce a goodness-of-fit test based on an information identity matrix known as Bartlett's First Identity which assumes that the model is correctly specified if the equality of the outer product and Hessian form of the information matrix is attained. White (1982) has used this identity to form the Information Matrix test for model misspecification. The proposed goodness-of-fit test is then illustrated using the negative binomial distribution and the simulation results are compared with EDF based goodness-of-fit tests.

Since there are still some statistical inference problems for the Delaporte distribution which do not seem to have been considered in the literature, we shall examine the parameter estimation, efficiency of estimation especially the method of moments and the maximum likelihood estimation, orthogonal parameters and other related inferences. We propose to estimate the parameters using a quadratic distance statistic which has been derived based on the Bartlett's First identity. The advantage of the quadratic distance method is that the global maximum is expected to be attained. Besides, the method can be simplified and eases computation if the model parameters are orthogonal. Furthermore, we have constructed confidence intervals under misspecification of model and the Delaporte distribution is used as an illustrative example. We found that the confidence intervals based on White's approach are more conservative. We have also studied the efficiencies of the estimations by the method of moments and maximum likelihood for the Delaporte distribution (Ruohenen, 1983).

1.2 Summary of Contributions of Thesis

Chapter 1 contains introductory summary on the materials underlying of this thesis. In Chapter 2, we give a brief literature review. We have discussed in detail maximum likelihood estimation, such as its application in parameter estimation and inference in statistics. The consequences of using orthogonal parameters in estimation have been studied and we present the procedure for constructing orthogonal parameters based on the paper by Willmot (1990). The effects of model misspecification on maximum likelihood estimation have also been considered here. We review White's information matrix test which has received much attention currently. An overview on the stochastic optimization method, simulated annealing, is given since we have used it to optimize the log-likelihood function in maximum likelihood estimation.

Chapter 3 proposes a goodness-of-fit test based on the Bartlett's First Identity. In fact, this identity is the basis of White's (1982) Information Matrix (IM) test for model misspecification. However, the proposed test is simplified since we consider it under orthogonality of the parameters with bootstrapped critical values. In addition, the direct application of Bartlett's First Identity as a goodness-of-fit test avoids the use of the complicated covariance matrix in White's IM test. This definitely reduces the difficulties in computation. Besides, the consistency and asymptotic normality of the proposed test have been derived. The proposed test is useful as an alternative method to determine the goodness–of-fit of a given random sample to a specified probability model in terms of power.

Chapter 4 examines the orthogonality of parameters for probability models with more than two parameters. This is an extension of Willmot's (1990) results. In particular, we derive the orthogonality for a class of Poisson-convolution models. This convolution may be regarded as a signal-plus-noise model and it is of practical importance. As an application of the orthogonality result, a uniformly most powerful test of the mean is developed based on an asymptotic result under model misspecification.

The Delaporte distribution, a Poisson-convolution model, is useful in fitting the number of claims in an insurance portfolio. The orthogonal parameters of the Delaporte distribution have been derived and discussed in Chapter 4. Since the Delaporte distribution has not been studied in detail, Chapter 5 first examines the efficiency of parameter estimation methods, such as the method of moments, maximum likelihood estimation and method of zero frequency. Furthermore, the confidence interval for Delaporte distribution has been constructed under two different conditions: (1) the model is correctly specified where the asymptotic variance is computed by inverse of the Fisher information matrix and (2) the model is misspecified where the asymptotic variance is computed by the outer product and Hessian form of information matrix as in White (1982). The robustness of the confidence interval based on the two approaches has been verified using generated random samples. In addition, we also propose to estimate the orthogonal parameters of the Delaporte distribution by using a quadratic distance statistic which is identical in form to White's

(1982) IM test. This is useful since the global maximum of the (log-) likelihood function is expected to be attained with the application of the proposed quadratic distance statistic.

1.3 Organization of Thesis

The organization of the thesis is as follows. Chapter 1 gives a general introduction including the overview on the topic of parameter estimation, goodness-of-fit test, orthogonal parameterization and model misspecification. It also provides a summary of contributions of the thesis.

Chapter 2 gives a brief literature survey about the background and motivation for the work of the thesis.

Chapter 3 proposes a new goodness-of-fit test and the method is illustrated using negative binomial (NB) distribution. Some empirical distribution function (EDF) statistics are considered for comparison purpose. Besides, we compare the proposed test with Jarque-Bera test which is well known as a goodness-of-fit test for normality.

Chapter 4 examines the orthogonality of the mean, μ , for models with more than two parameters. In particular, the orthogonal parameters for a class of Poisson-convolution models have been derived.

Chapter 5 studies on the Delaporte distribution, a model for claim number process. Statistical inference for this distribution has been examined exclusively.

Chapter 6 concludes the thesis with a summary of the research findings and proposes future work.

CHAPTER 2

LITERATURE REVIEW

2.1 Maximum Likelihood Estimation

Maximum likelihood estimation (MLE), one of the main tools in modern statistical inference, was proposed and developed by Fisher in the 1920s (see Fisher, 1925). Consider a random sample of size n, where x_i denotes the i-th observation from a population that has probability density function or probability mass function $P(x;\Theta)$. The function

$$L(\Theta)=\prod_{i=1}^{n}P(x_i;\Theta) \tag{2.1}$$

is known as the likelihood function, where Θ represents the vector of unknown parameters. The value of Θ that maximizes (2.1) is the ML estimate.

For discrete distributions which take non-negative integers, the likelihood function is expressed as

$$L(\Theta)=\prod_{k=0}^{t}P(k)^{f_k}$$

where $P(k)$ and f_k respectively denotes the probability mass function and the frequency of k counts for k = 0,1,2,…, t and t is the largest count in the data set. In practice, it is more convenient to deal with the log-likelihood function given by

$$\ln L(\Theta)=\sum_{k=0}^{t}f_k \ln P(k) \tag{2.2}$$

bearing in mind that

$$\sum_{k=0}^{t}f_k=n,\ \sum_{k=0}^{t}kf_k=\sum_{i=1}^{n}x_i .$$

Hence, from (2.2), the partial derivative of the log-likelihood function is given by

$$\frac{\partial \ln L(\Theta)}{\partial \Theta} = \sum_k f_k \frac{\partial \ln P(k)}{\partial \Theta} = \sum_k f_k \frac{1}{P(k)} \frac{\partial P(k)}{\partial \Theta}$$

The maximum likelihood estimator $\hat{\Theta}$ is given by

$$\hat{\Theta} = \underset{\Theta}{\arg\max} \log L(\Theta) .$$

MLE has wide application in parameter estimation and inference in statistics. MLE can be developed for a large variety of situations. One important aspect of MLE is that the method has desirable mathematical and optimality properties. For instance, the large sample theory of asymptotic normality, consistency and efficiency for MLE are well established if the model is correctly specified. Moreover, one of the most useful properties of MLE is the invariance property (see Zehna, 1966 and Casella and Berger, 2002), that is, if $\hat{\theta}$ is the ML estimator of θ and if *h* is a function, then $h(\hat{\theta})$ is the ML estimator of $h(\theta)$. The proof of invariance property is given by Berk (1967). Therefore, the same MLE solution is obtained independent of the parameterization used. Furthermore, there are a number of statistical inference methods which are developed using MLE. For example, MLE is the basis for the chi-square test, Bayesian methods, inference with missing data, modeling of random effects, and model selection criteria such as the Akaike information criterion (Akaike, 1973) and the Bayesian information criterion (Schwarz,1973).

However, one still needs to be aware of the limitations of MLE. One of the problems is that ML estimates can be heavily biased for finite (small) sample. Besides, the complexity of MLE depends on the form of the likelihood equation which involves the density function of the model. It is often not trivial to find analytical expressions for many problems and we must resort to more elaborate techniques. In such situations, it is necessary to use numerical methods to evaluate the MLE by successive iteration. There is a variety of numerical methods which are available for locating the root of an equation. For example, Kale (1961) has discussed several of these methods (the fixed-derivative Newton, Newton-Raphson and 'scoring for parameters' methods) for obtaining the MLE of a single parameter under the usual regularity conditions, from the point of view whether or not they satisfy certain desirable probabilistic properties as $n \to \infty$. In a subsequent paper, Kale (1962) makes a similar study for the multi-parameter case. Furthermore, in any practical problem, the associated existence of a unique consistent root and regularity conditions are no guarantee that a single root of the likelihood equation will exist for a simple sample as above. In fact, multiple roots often exist, corresponding to multiple relative maxima of the likelihood function, even if the regularity conditions are satisfied. Barnett (1966) has also pointed out that there are cases where the likelihood equation may have an unbounded number of roots. When this problem occurs, we find that MLE can be sensitive to the choice of starting values in order to obtain the global optimum solutions. Kirkpatrick, Gelatt and Vecchi (1983) have introduced a stochastic optimization algorithm known as simulated annealing (SA). In theory, SA is able to locate the global maximum (see section 2.5 for details). However, SA in practice takes a comparative longer time to find the solution and SA does encounter problem when the log-likelihood function is ill-conditioned; for example, the log-likelihood is "flat" for some parts of the parameter space. Gan and Jiang (1999) have proposed a test for global maximum given that the model is correctly specified and claimed that the global maximum will be obtained if and only if the following condition (Bartlett's First Identity) is satified,

$$\left(\frac{\partial \log L}{\partial \theta}\right)^2 + \frac{\partial^2 \log L}{\partial \theta^2} \approx 0,$$

where L is the likelihood function and θ is an unknown parameter.

Various aspects of MLE have been discussed in the statistical literature. Sprott (1983) has examined a large class of convolution and generalized models where the number of ML equations may be reduced. Thus, for two-parameter distribution with the sample mean as a solution of one of the score function, there is only a single equation to be solved iteratively. The problem of high correlation between ML estimators can be solved or reduced by orthogonal parameterization (see section 2.2). Habibullah and Katti (1991) have proposed a modified procedure to improve the convergence rate of steepest descent method in the maximization of the likelihood functions of some widely used generalized distributions. Futhermore, Yanagimoto (1991) has studied an estimation problem of a model using conditional maximum likelihood estimator.

In general, ML estimators are not robust and inconsistent when the model is misspecified. There are some studies about the behaviour of maximum likelihood estimators under model misspecification. Quasi-maximum likelihood estimation (QMLE) in the presence of misspecified models has been considered, for instance, by White (1982), and Gourieroux, Monfort and Trognon (1984).

2.2 Orthogonal Parameterization

Cox and Reid (1987) have studied and summarized some of the desirable properties and consequences of parameter orthogonality in maximum likelihood estimation. Parameter orthogonality is defined as follows. Consider a vector of unknown parameters, $\boldsymbol{\theta}$ partitioned into two vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ of length p_1 and p_2 respectively, where $p_1 + p_2 = p$. $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are said to be orthogonal if the elements of the information matrix satisfy

$$i_{\theta_s\theta_t} = n^{-1}E\left(\frac{\partial \ln L}{\partial \theta_s}\frac{\partial \ln L}{\partial \theta_t};\theta\right) = -n^{-1}E\left(\frac{\partial^2 \ln L}{\partial \theta_s \partial \theta_t};\theta\right) = 0,$$

for $s = 1, \ldots, p_1$, $t = p_1+1, \ldots, p_2$ and L is the likelihood function.

The construction of orthogonal parameters has been generalized based on the arguments of Huzubazar (1950) and Jeffreys (1961, p.208) (refer to Cox and Reid, 1987, p.3). Orthogonal parameters for some well known distributions have been derived. The two papers are then examined for the application of orthogonal parameters to conditional inference.

Orthogonal parameterizations of distributions with two parameters are well reported. For examples, Huzurbazar (1956) has considered the orthogonal parameterization for an exponential family type of distributions. Huzurbazar (1950) discussed the orthogonal parameterization of the negative binomial distribution while for the Poisson-inverse Gaussian it has been considered by Stein, Zucchini and Juritz (1987). Willmot (1988) has examined orthogonal parameterization for a large family of discrete distributions which includes many well-known distributions. The orthogonal parameterization of the mean and a shape parameter is derived. Willmot also showed that the problem of inherently high correlation of the maximum likelihood estimators for the standard contagious distribution given by Douglas (1980) can be solved or reduced by using orthogonal parameterization.

Willmot (1990) considered orthogonal parameterization involving the mean for distributions with two unknown parameters. He has demonstrated a simple construction of a parameter orthogonal to the mean and the method is given as follows. Let $f(x)$ and $C(t) = \log\left[E\left(e^{tX}\right)\right]$ be the probability mass function or probability density function and the cumulant generating function (cgf) for a two-parameter distribution respectively, where X is a random variable. Suppose that $f(x)$ has two parameters, μ and θ, where μ is the population mean and $\sigma = \sigma(\mu, \theta)$ is the population standard deviation. The maximum likelihood estimate, $\hat{\mu}$ for the mean μ, is the sample mean under the usual regularity conditions. Therefore, as noted by Sprott (1983), one must have the functions $g = g(\mu, \theta)$ and $h = h(\mu, \theta)$ such that

$$\frac{\partial}{\partial \mu} \ln f(x) + g \frac{\partial}{\partial \theta} \ln f(x) + h(x - \mu) = 0, \tag{2.3}$$

or equivalently,

$$\frac{\partial}{\partial \mu} C(t) + g \frac{\partial}{\partial \theta} C(t) + h\left(\frac{\partial}{\partial t} C(t) - \mu\right) = 0. \tag{2.4}$$

We then define the asymptotic variance-covariance matrix of $(\hat{\mu}, \hat{\theta})$ as the inverse of the 2 by 2 Fisher information matrix, $A = (a_{ij})$; for example,

$$a_{22} = nE[\{\partial \ln f(X) / \partial \theta\}^2]. \tag{2.5}$$

By differentiating the identity $\int (x - \mu) f(x) dx = 0$ with respect to μ, we get

$$E_\theta[(X - \theta_1)\{\partial \ln f(X) / \partial \theta_j\}] = 0. \tag{2.6}$$

Thus, from (2.3), (2.5) and (2.6),

$$a_{11} = g^2 a_{22} + nh^2 \sigma^2, \qquad a_{12} = a_{21} = -g a_{22}.$$

where $\sigma^2 = E\left[(X - \mu)^2\right]$. So we have

$$A = \begin{bmatrix} g^2 a_{22} + nh^2 \sigma^2 & -g a_{22} \\ -g a_{22} & a_{22} \end{bmatrix} \tag{2.7}$$

Note that $\det(A) = a_{11} a_{22} - a_{12}^2 = nh^2 \sigma^2 a_{22}$. By writing $A^{-1} = (a^{ij})$ and letting $\hat{\mu}$ denote the sample mean, (2.7) leads to

$$\text{asvar}(\hat{\mu}) = a^{11} = \sigma^2 / n, \qquad \text{ascov}(\hat{\mu}, \hat{\theta}) = a^{12} = g\sigma^2 / n.$$

Suppose that $\phi = \phi(\mu, \theta)$ is the new proposed parameter, we find that

$$\text{ascov } (\hat{\mu}, \hat{\phi}) = \text{ascov } (\hat{\mu}, \hat{\theta}) \frac{\partial \phi}{\partial \theta} + \text{asvar } (\hat{\mu}) \frac{\partial \phi}{\partial \mu}.$$

The parameter ϕ will be orthogonal to μ if ascov $(\hat{\mu}, \hat{\phi}) = 0$ (Willmot, 1990) and it has to satisfy

$$\frac{\partial \phi}{\partial \mu} + g \frac{\partial \phi}{\partial \theta} = 0.$$

The above results are then applied to the construction of orthogonal parameterizations for a fairly general class of compound distributions and convolutions. In particular, Willmot (1990) first considered the models which may be expressed in compound Poisson form with the power series. The cgf is defined by

$$C(t) = \lambda \left\{ \frac{A(\theta e^t)}{A(\theta)} - 1 \right\} \tag{2.8}$$

where $A(\theta)$ is the series function. From (2.8), one has

$$\mu = \lambda \theta A'(\theta) / A(\theta), \qquad \sigma^2 = \mu \{1 + \theta A''(\theta) / A'(\theta)\}.$$

It follows that, for the family of distributions with cgf (2.8), the mean $\mu = \lambda \theta A'(\theta) / A(\theta)$ is orthogonal to

$$\phi = \frac{\mu}{\theta A'(\theta)} = \frac{\lambda}{A(\theta)}.$$

As an illustration, consider the negative binomial distribution with cgf

$$C(t) = r \ln \{(1-\theta) / (1-\theta e^t)\},$$

which is of the same form as (2.6) with

$$\lambda = -r \ln(1-\theta), \qquad A(\theta) = -\ln(1-\theta).$$

Intuitively, the mean $\mu = r\theta/(1-\theta)$ is orthogonal to $\phi = r$. The result is used in Chapter 3 in order to simplify the proposed test statistic for the goodness-of-fit problem.

Willmot (1990) has also discussed orthogonal parameterization for convolution models which involve the convolution of a Poisson and a power series distribution with cgf

$$C(t) = \lambda(e^t - 1) + A(\theta e^t) - A(\theta), \tag{2.9}$$

where $\exp\{A(\theta)\}$ is the series function. One finds from (2.9) that

$$\mu = \lambda + \theta A'(\theta), \qquad \sigma^2 = \mu + \theta^2 A''(\theta).$$

Hence, the mean $\mu = \lambda + \theta A'(\theta)$ is orthogonal to

$$\phi = \frac{\mu}{\theta} - A'(\theta) = \frac{\lambda}{\theta}$$

for the family of distributions with cgf defined by (2.9).

Willmot (1988, 1990) has considered the orthogonality for a wide class of discrete models with two parameters. Results for distributions with more than two parameters do not seem to be well-publicized in the statistical literature. In Chapter 4, we have extended the work of Willmot (1990) to models with more than two parameters. The remaining parameters after reparameterization are treated as nuisance parameters. In addition, we have derived the orthogonal parameters for a class of Poisson-convolution models.

2.3 Maximum Likelihood Estimation under Model Misspecification

In practice, model misspecification occurs rather frequently. We shall focus on the effects of model misspecification on maximum likelihood estimators (MLEs). The misspecification of the model would normally lead to the violation of the properties of the maximum likelihood estimator. The consistency and asymptotic normality of the estimator will need further justification if one does not assume that the probability model is correctly specified. Berk (1966, 1970) has considered the consistency question by using the Bayesian approach. Huber (1967) has given a general condition where the MLEs converge to a limit although the probability model is not correctly specified. The approach follows Wald (1943). Huber has also studied the asymptotic normality of the MLEs. Akaike (1973) pointed out that the maximum likelihood estimator is just an estimator for the parameters which minimize the Kullback-Leibler Information Criterion (KLIC) when the true model is unknown.

White (1982) has reviewed and studied properties of MLEs in the presence of misspecification. Some simple conditions and treatments for consistency and asymptotic normality were given. In addition, White has proposed a useful diagnostic test on misspecification based on the quasi-maximum likelihood estimator (QMLE) and the information matrix. The word "quasi" means that the ML estimators have been obtained from the log-likelihood of a misspecified model. When the model is correctly specified, one finds that the QMLE is the same as the usual MLE. Based on Assumptions A1 and A2 (see Appendix A), the QMLE of the sample is defined as (see White, 1982, p. 2-3),

$$\ln L_n(\theta; X) \equiv n^{-1}\sum_{i=1}^{n} \ln f(X_i; \theta),$$

where $X_i, i = 1, ..., n,$ is the independent random $1 \times M$ vectors and a QMLE is denoted as a parameter vector $\hat{\theta}_n$ which solves the following condition,

$$\max_{\theta \in \Theta} \ln L_n(\theta; X).$$

In other words, the QMLE is generally a strongly consistent estimator for the parameter vector which minimizes the KLIC. Besides, QMLE is considered as a special case of LeCam's (1953) asymptotic normality result where the asymptotic covariance matrix of the

QMLE no longer equal to the inverse Fisher's information matrix. The more general form of the asymptotic covariance matrix which can be estimated consistently is given by

$$C(\theta)=A(\theta)^{-1}B(\theta)A(\theta)^{-1}.$$

When the model is correctly specified, $A(\theta)=-B(\theta)$ and $C(\theta)=-A(\theta)^{-1}=B(\theta)^{-1}$, where $-A(\theta)$ is the Fisher's information matrix. White (1982) introduced the IM test, a new test for misspecification by using the latter properties of QMLE. The details of the IM test are provided in section 2.4.

White's results have received much attention from researchers and extensive work has been done on the theoretical aspect of misspecification. In addition, many statistical techniques that are robust to misspecification have been proposed (see Fomby and Carter Hill, 2003 for details) and these techniques are found to be useful in empirical research.

2.4 Information Matrix Test

White (1982) introduced the Information Matrix (IM) test as a general test to detect model misspecification based on the information matrix equality. According to the information matrix equivalence theorem, when the model is correctly specified, the Hessian form of the information matrix $-A(\theta)$ and its outer product form $B(\theta)$ satisfies $A(\theta)+B(\theta)=0$. The failure of this equality implies misspecification of the model. However, $A(\theta)+B(\theta)=0$ does not imply that the model is correctly specified.

In general, if expectations exist, the Hessian matrix $A(\theta)$ and outer product matrix $B(\theta)$ for the information matrix are defined as (see White, 1982)

$$A(\theta)=E\left[\frac{\partial^2 \ln P(x;\theta)}{\partial\theta_i\partial\theta_j}\right],$$

$$B(\theta) = E\left[\frac{\partial \ln P(x;\theta)}{\partial \theta_i} . \frac{\partial \ln P(x;\theta)}{\partial \theta_j}\right].$$

The formal definition of IM test is given by Theorem 1 (see Appendix A). Notice that the original form of White's IM test is not easy to compute since it involves analytical third derivatives of the covariance matrix. In addition, available evidence showed that the asymptotic χ^2 distribution is always a poor approximation to the finite sample distribution of the test statistic. Therefore, the true size of IM test in finite samples often differs greatly from the nominal size derived from asymptotic theory. This has been shown in Monte Carlo experiments as reported by Taylor (1987), Orme (1990), Chesher and Spady (1991), Davidson and MacKinnon (1992), and Horowitz (1994). Several methods have been advocated to overcome the problem. For instance, Chester and Spady (1991), by using a higher-order Edgeworth expansion, improved upon the critical values for the IM test statistic of some specific models; but this is algebraically tedious and difficult to implement. Davidson and MacKinnon (1992) have developed a variant of the IM test based on the double-length artificial regressions while Horowitz (1994) suggested bootstrap-based critical values. Recently, Dhaene and Hoorelbeke (2004) have examined a new form of the IM test where the sample covariance matrix is estimated by parametric bootstrap. Dhaene and Hoorelbeke (2004) stated that their version of the test is easier to compute and requires no analytical derivations. However, it can be time consuming if one decides to use bootstrap-based critical values because a nested bootstrap is required. Furthermore, Croux, Dhaene, Hoorelbeke (2006) have studied the behavior of IM test when the ML estimators are replaced by robust estimators in constructing the test. The purpose of this change is to reduce the masking effect when the outliers are present and it can also improve the power of the test. In short, IM test do play an important role for the detection of misspecification, however, precautions have to be taken in handling the estimation of the covariance matrix. We have applied the IM test proposed by Dhaene and Hoorelbeke (2004) in the goodness-of-fit problem mentioned in Chapter 3 for comparison purpose.

2.5 Simulated Annealing

In recent years, many researchers have considered stochastic methods in the numerical optimization of an objective function. Simulated annealing (SA) proposed by Kirkpatrick, Gelett and Vecchi (1983) is one of these stochastic methods and it makes less assumptions compared to the "classical" optimization methods. SA is first developed based on the ideas of Metropolis et al. (1953). Its genesis involves the simulation of a system of particles with a change in temperature and such system will try to find an equilibrium point that minimizes the total energy under perturbation. Metropolis et al. have applied statistical thermodynamics in order to estimate the equilibrium points. Hence, Kirkpatrick et al. (1983) have implemented these ideas in a more general optimization problem, where the value of the objective function represents the energy and Metropolis' temperature is treated as a control parameter in the optimization process. Therefore, SA's roots are in thermodynamics, where one studies a system's thermal energy. The concept of the algorithm is motivated by the cooling of molten metal. The metal reaches a low energy stage after slow cooling (annealing). Inherent random fluctuations in energy allow the annealing system to escape local energy minima and to achieve the global minimum.

In fact, many researchers would usually experience the common difficulties (slow rate of convergence, run-time execution, etc) when implementing some "classical" optimization methods such as Newton-Raphson, fixed derivative Newton, Davidon-Fletcher-Powell, and the simplex method for a sophisticated distribution which have multiple roots. However, SA is specially designed for functions with multiple optima and it makes fewer assumptions on the “shape” and nature of the function. SA explores the entire surface of the function very roughly, and tries to optimize the function during its uphill and downhill moves. Thus, it is much more robust than the classical algorithms. This robustness comes at a cost-longer run time. However, in an era of increasingly cheaper computing, this substitution of computer time for trial, error, and frustration should be encouraged.

We give a brief technical explanation of the SA algorithm. The algorithm at the beginning randomly chooses a trial point within the step length v (a vector of length n) of the user-selected starting point. The function is evaluated at this trial point, and is compared to its value at the initial point. In a maximization problem, all uphill moves are accepted, and the algorithm continues from an accepted trial point (note the step length is always centered on the trial point and not 0). However, downhill moves may be accepted; the decision is made by the *Metropolis* criteria, which uses temperature, T, and the magnitude of the downhill move in a probabilistic manner. That is, the higher the temperature and the smaller the downhill move, the more likely that the move will be accepted. If the trial point is rejected, then another point is chosen for a trial evaluation.

Each element of v, the step length is adjusted periodically so that half of all function evaluations in that direction are accepted. A fall in temperature is imposed upon the system with the r_T variable by $T_{i+1} = r_T \cdot T_i$ where i denotes the i^{th} iteration. Downhill moves are less likely to be accepted, and the percentage of rejections rises as the temperature declines. Given the scheme for the selection of the step length, it falls as a result. Thus, as temperature declines, the step length falls, and simulated annealing focus upon the most promising area for optimization. More details about the algorithm, advantages and disadvantages of SA have been reviewed by Fouskakis and Drapper (2002). We have used the SA algorithm in Matlab (see Goffe, 1996) to handle most of the parameter estimation problems appeared in the next few Chapters.

Bohachevsky, Johnson and Stein (1986) have presented a generalized simulated annealing for function optimization and the method is applied to some complicated examples. The better and improved optimum is determined. Bertsimas and Tsitsiklis (1993) have studied the convergence and behavior of simulated annealing in its applications. Furthermore, Brooks and Morgan (1995) have given an overview on the theory of simulated annealing. A hybrid approach which combines the simulated annealing and a traditional optimization algorithm has been developed. The hybrid approach has performed better than the two algorithms if they are used separately.

CHAPTER 3

A GOODNESS-OF-FIT TEST BASED ON AN INFORMATION MATRIX IDENTITY

3.1 Introduction

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution $F(x)$. An important issue in a statistical analysis is to determine if a given random sample fits a probability model well. This is a goodness-of-fit problem and it leads to the consideration of the following null and alternative hypotheses for a given distribution $G(x)$:

$$\begin{aligned} &H_0\text{: } F(x) = G(x), \\ &H_a\text{: } F(x) \neq G(x). \end{aligned} \tag{3.1}$$

Many goodness-of-fit tests have been proposed and studied. Tests like the Kolmogorov-Smirnov, Anderson-Darling and Cramér-von Mises based on the empirical distribution function (EDF) $F_n(x)$ have been considered (see Stephens, 1974). Here, we consider a goodness-of-fit test based on an information matrix identity known as Bartlett's First Identity which states that the outer product and Hessian form of the information matrix are equal under correct model specification. This identity is the basis of White's (1982) Information Matrix (IM) test for model misspecification. However, the proposed test differs from the IM test in two ways. Firstly, the goodness-of-fit test statistic will be considered under orthogonality of the parameters with the bootstrapped critical values adjusted for bias by using the bias-corrected accelerated or BC_a method. When parameters are orthogonal, the proposed test statistic is simplified and this reduces computation. Willmot (1988, 1990) has derived orthogonal parameters for a wide range of discrete distributions. Secondly, the direct application of Bartlett's First Identity as a goodness-of-fit test avoids the need to use the complicated covariance matrix. This also leads to simplification and reduces computation. In contrast White's IM test involves derivatives of the covariance matrix. The application of Bartlett's First Identity for goodness-of-fit does not seem to have been widely reported in the statistical literature. The proposed goodness-of-fit test will be illustrated with the negative binomial (NB) distribution.

The NB distribution is a well-known model which may be formulated as a mixed or compound Poisson distribution. There are various parameterizations of the NB (see Johnson, Kemp and Kotz, 2005, p.209) and the following parameterization for the probability mass function (pmf) will be adopted in this study:

$$P(x;\mu,\phi)=\begin{cases}\binom{\phi+x-1}{\phi-1}\left(\frac{\mu}{\mu+\phi}\right)^{x}\left(1-\frac{\mu}{\mu+\phi}\right)^{\phi}, & x=0,1,2,...,\\ 0 & ,\ \text{otherwise.}\end{cases} \tag{3.2}$$

where $\mu > 0$ and $\phi > 0$ are the mean and index parameters respectively. For this parameterization, the two parameters are orthogonal, that is, the maximum likelihood estimators (MLEs) of μ and ϕ are uncorrelated.

Parameter orthogonality is formally defined as follows. Consider a vector of unknown parameters, $\boldsymbol{\theta}$ partitioned into two vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ of length p_1 and p_2 respectively, where $p_1 + p_2 = p$ (Cox and Reid, 1987). $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are said to be orthogonal if the elements of the information matrix satisfy

$$i_{\theta_s\theta_t} = n^{-1}E\left(\frac{\partial \ln L}{\partial \theta_s}\frac{\partial \ln L}{\partial \theta_t};\theta\right) = -n^{-1}E\left(\frac{\partial^2 \ln L}{\partial \theta_s \partial \theta_t};\theta\right) = 0,$$

for $s = 1, \ldots, p_1$, $t = p_1+1, \ldots, p_2$ and L is the likelihood function.

This chapter is arranged as follows. Section 3.2 gives preliminaries on the goodness-of-fit tests based on the EDF for discrete distributions (see Famoye, 2000). A new test statistic based on the Bartlett's First Identity (BFI) is then proposed in section 3.3. Some asymptotic properties of the proposed test are discussed in section 3.3.2. In section 3.3.3 and 3.3.4, the procedure for bootstrapping the critical values of the proposed test statistic with corrections by the BC_a method is given. Using Monte Carlo simulations, the proposed test is then compared with the discrete EDF tests and the IM test (see Section 2.4) as implemented recently by Dhaene and Hoorelbeke (2004). The Pearson chi-square goodness-of-fit test has not been considered due to its poor power. The results of the simulations are discussed in section 3.3.6. Section 3.4 gives the results of the comparison between BFI Goodness-of-fit test and Jarque-Bera test. The conclusion is given in section 3.5.

3.2 Empirical Distribution Function based Goodness-of-fit Tests

Empirical Distribution Function (EDF) based goodness-of-fit tests consider the discrepancy between the sample EDF $F_n(x)$ and the actual distribution function $F(x;\theta)$ where θ is the vector of parameters. If the assumed model is correct, $F_n(x)$ is expected to be close to $F(x;\theta)$. If not, one suspects that the hypothetical distribution function $F(x;\theta)$ is not the correct model.

For a random sample $X_1, X_2, \ldots, X_n$ of size *n* from a discrete distribution, let f_x be the observed frequencies, where x = 0, 1, 2, …, *k*, with *k* being the largest observation and

$$n = \sum_{x=0}^{k} f_x .$$

The EDF for the sample is given by

$$F_n(x) = \frac{1}{n}\sum_{i=0}^{[x]} f_{i,} \quad [x] = 0,1,2,3,...,k.$$

where $[x]$ denotes the greatest integer that is less than or equal to *x* (or integer part of *x*).

The theoretical distribution function is

$$F(x;\theta) = \sum_{i=0}^{x} P(i;\theta), \quad x \geq 0,$$

where $P(i;\theta)$ is the probability mass function.

In order to test the hypotheses in (3.1), we consider the four modified EDF test statistics for observed count data,

(a) Kolmogorov-Smirnov statistic, K_d

$$K_d = \sup_x \left| F_n(x) - F(x;\hat{\theta}) \right|.$$

(b) Cramér-von Mises statistic, W_d

$$W_d = n\sum_{x=0}^{k} \left[F_n(x) - F(x;\hat{\theta}) \right]^2 P(x;\hat{\theta}).$$

(c) Anderson-Darling statistic, A_d

$$A_d = n\sum_{x=0}^{k} \frac{\left[F_n(x) - F(x;\hat{\theta}) \right]^2 P(x;\hat{\theta})}{F(x;\hat{\theta})\left[1 - F(x;\hat{\theta})\right]}.$$

(d) Watson statistic, U_d

$$U_d = n\left\{ \sum_{x=0}^{k} \left[F_n(x) - F(x;\hat{\theta}) \right]^2 P(x;\hat{\theta}) - Q^2 \right\},$$

where $Q = \sum_{x=0}^{k} \left[F_n(x) - F(x;\hat{\theta}) \right] P(x;\hat{\theta}).$

3.3 A Goodness-of-fit Test based on Bartlett's First Identity

3.3.1 Introduction

White's IM test has been defined based on Bartlett's First Identity (BFI). According to BFI, when the model is correctly specified, the Hessian form of the information matrix $-A(\theta)$ and its outer product form $B(\theta)$ satisfy $A(\theta)+B(\theta)=0$. The failure of this equality implies misspecification of the model.

In general, the sample quantities of Hessian matrix $A(\theta)=\left(A_{ij}(\theta,n)\right)$ and outer product matrix $B(\theta)=\left(B_{ij}(\theta,n)\right)$ for the information matrix are defined as (see White, 1982)

$$A_{ij}(\theta,n)=n^{-1}\sum_{x=0}^{k} f_x \frac{\partial^2 \ln P(x;\theta)}{\partial\theta_i\partial\theta_j}, \tag{3.3}$$

$$B_{ij}(\theta,n)=n^{-1}\sum_{x=0}^{k} f_x \left[\frac{\partial \ln P(x;\theta)}{\partial\theta_i}.\frac{\partial \ln P(x;\theta)}{\partial\theta_j}\right]. \tag{3.4}$$

The proposed BFI goodness-of-fit test statistic, under the assumption of orthogonality of parameters, is defined as

$$D=tr\left[A(\boldsymbol{\theta})+B(\boldsymbol{\theta})\right], \tag{3.5}$$

where $\boldsymbol{\theta}$ is the parameter vector of the model. The trace (*tr*) of the matrices $A(\theta)$ and $B(\theta)$ are considered because off-diagonal elements of the information matrices are zero when parameters are orthogonal.

Due to the orthogonality of parameters and direct application of BFI, which dispenses the need to use the covariance matrix, the proposed test statistic is of a much simpler form than the IM test. We focus on the goodness-of-fit test for NB distribution as given in (3.1). Since the probability mass function (3.2) is in terms of orthogonal parameters μ and ϕ, where the ML estimator $\hat{\mu}$ of the mean μ is the sample mean $\overline{X}$, the test statistic may be based on ϕ only, given $\hat{\mu} = \bar{X}$. In this case, the Hessian and outer products matrices are reduced to a scalar. Let $P(x) = P(x;\hat{\mu},\phi)$. Therefore, we can rewrite (3.3) and (3.4) as

$$A(\phi,n) = n^{-1}\sum_{x=0}^{k} f_x \frac{\partial^2 \ln P(x)}{\partial \phi^2}, \tag{3.6}$$

$$B(\phi,n) = n^{-1}\sum_{x=0}^{k} f_x \left[\frac{\partial \ln P(x)}{\partial \phi} . \frac{\partial \ln P(x)}{\partial \phi}\right]. \tag{3.7}$$

The first and second derivatives of the log-likelihood function, with μ known, are

$$\frac{\partial \ln P(x)}{\partial \phi} = \begin{cases} \left[\Phi + \ln(1-\Phi)\right], & x = 0, \\ 1 - \dfrac{1}{\mu}(x+1)\dfrac{P(x+1)}{P(x)} + \ln(1-\Phi) + \dfrac{1}{P(x)}\displaystyle\sum_{j=0}^{x-1}\frac{\Phi^{j+1}}{j+1}P(x-j-1) & \\ & x = 1,2,3,\dots \end{cases}$$

$$\frac{\partial^2 \ln P(x)}{\partial \phi^2} = \begin{cases} \dfrac{1}{\phi}\Phi^2, & x = 0, \\ -\dfrac{1}{\mu}(x+1)\dfrac{P(x+1)}{P(x)}\left[\dfrac{\partial \ln P(x+1)}{\partial \phi} - \dfrac{\partial \ln P(x)}{\partial \phi}\right] + \Phi\left(\dfrac{1}{\phi} + \dfrac{1}{P(x)}\right) & \\ \cdot\displaystyle\sum_{j=0}^{x-1}\Phi^{j}\left[-\frac{P(x-j-1)}{(\mu+\phi)} + \frac{P(x-j-1)}{j+1}\frac{\partial \ln P(x-j-1)}{\partial \phi}\right] & \\ -\dfrac{1}{P(x)}\dfrac{\partial \ln P(x)}{\partial \phi}\displaystyle\sum_{j=0}^{x-1}\frac{\Phi^{j+1}}{j+1}P(x-j-1), & x = 1,2,3,\dots \end{cases}$$

where $\Phi = \mu/(\mu+\phi)$. For the NB goodness-of-fit problem, the goodness-of-fit test statistic is given by

$$D = A(\phi,n) + B(\phi,n).$$

The rejection region, R of the test statistic of size α is given by

$$R = \{|D| \geq c_{n,\alpha}\},$$

where $c_{n,\alpha}$ is a bootstrapped critical value determined from $P\Big[|D| \geq c_{n,\alpha} \,\Big|\, H_0 \text{ is true}\Big] = \alpha.$

3.3.2 Asymptotic Properties

In this section we summarize some properties of the proposed goodness-of-fit test given in section 3.4.1. As in White's IM test, for the proposed goodness-of-fit test, it is assumed that

(i) global ML estimates (MLE) are used (consistent estimates).

(ii) the null hypothesis is that the model is correctly specified.

If local ML estimates (which are inconsistent estimates) are used in the proposed goodness-of-fit test then Bartlett's First Identity is not satisfied leading to the conclusion that model is not correctly specified even though the model is correct. That is the use of *local* MLE gives rise to model misspecification when the proposed goodness-of-fit is applied. Therefore, the BFI goodness-of-fit test is not applicable if global ML estimates are not used.

(a) *Consistency of Goodness-of-Fit Test*

We wish to show that the proposed goodness-of-fit test is consistent, that is, the null hypothesis (of correct specification of model) is rejected if the test statistic $D_n(\theta)$ which is defined in (3.8) has a value that differs sufficiently from 0.

As in White (1982, p.9), we define

$$d_\ell(x_k, \theta) = \frac{\partial^2 \ln P(x_k;\theta)}{\partial\theta_i \partial\theta_j} + \left[\frac{\partial \ln P(x_k;\theta)}{\partial\theta_i} . \frac{\partial \ln P(x_k;\theta)}{\partial\theta_j}\right],$$

$\ell = 1, 2, \ldots, p(p+1)/2, \quad i, j = 1, 2, \ldots, p, \quad k = 1, 2, 3, \ldots, n$

and $D_\ell\left(\hat{\theta}_n\right)=\frac{1}{n}\left\{\sum_{k=1}^{n} d_\ell\left(x_k,\hat{\theta}_n\right)\right\}$. Define a $q\times 1$ vector

$$D_n\left(\hat{\theta}_n\right)=\left(D_1,D_2,\ldots,D_q\right)',\ q\le p(p+1)/2$$

so that

$$D_n\left(\hat{\theta}_n\right)=\frac{1}{n}\left\{\sum_{k=1}^{n} d\left(x_k,\hat{\theta}_n\right)\right\}, \tag{3.8}$$

where $d\left(x_k,\hat{\theta}_n\right)=\left(d_1,d_2,\ldots,d_q\right)'$.

Let $D(\theta)=A(\theta)+B(\theta)$ (see White, 1982, p.5), where

$$A(\theta)=E\left(\frac{\partial^2 \ln P\left(x_k;\theta\right)}{\partial\theta_i\partial\theta_j}\right) \text{ and } B(\theta)=E\left(\frac{\partial \ln P\left(x_k;\theta\right)}{\partial\theta_i}.\frac{\partial \ln P\left(x_k;\theta\right)}{\partial\theta_j}\right).$$

The definitions above are used in the following Proposition.

Proposition 3.4.2.1. *Consider a probability space* $(\Omega,\Im,P)$. *Let* $\Theta\subset R^p$ *be compact and* $\{D_n\}$ *be a sequence of measurable functions which are continuous with respect to* Θ. *Suppose* $\{\hat{\theta}_n\}$ *is a sequence of measurable functions such that* $\hat{\theta}_n\to\theta_*$ *in probability where* $\theta_*\in\Theta$ *and* $D(\theta)$ *is a continuous function of* $\theta\in\Theta$. *Then* $D_n\left(\hat{\theta}_n\right)\to D(\theta_*)$ *in probability.*

Proof: By the assumption of the compactness of the parameter space $\Theta\subset R^p$, there exists a unique maximizer θ_* of the log-likelihood function. Hence, by the Weak Law of Large Numbers, $D_n(\theta)\to D(\theta)$ in probability or $D_n(\theta)-D(\theta)\underset{P}{\to}0$ for $\theta\in\Theta$. Let $T_n(\theta)=D_n(\theta)-D(\theta)$ and $T_n(\theta)\underset{P}{\to}0$.

As in White (1982, p.20), proof of Theorem 4.1, we assume that Assumptions A8 and A9 (see Appendix A) hold, Taylor's expansion (Mean Value Theorem, see Appendix B) gives

$$T_n\left(\hat{\theta}_n\right) = T_n\left(\theta_*\right) + \left(\hat{\theta}_n - \theta_*\right)\nabla T_n\left(\bar{\theta}_n\right)$$

where $\bar{\theta}_n$ is between $\hat{\theta}_n$ and θ_*. Thus

$$\left|T_n\left(\hat{\theta}_n\right) - T_n\left(\theta_*\right)\right| = \left|\hat{\theta}_n - \theta_*\right|\left|\nabla T_n\left(\bar{\theta}_n\right)\right|$$

Since $\left|\nabla T_n\left(\bar{\theta}_n\right)\right|$ is bounded, by Assumption A9 and $\hat{\theta}_n \to \theta_*$, we have

$$T_n\left(\hat{\theta}_n\right) - T_n\left(\theta_*\right) \underset{P}{\to} 0 .$$

In fact $\left|\nabla T_n\left(\bar{\theta}_n\right)\right|$ is stochastically bounded, that is, $\left|\nabla T_n\left(\bar{\theta}_n\right)\right| = O_P(1)$.

Therefore, by the continuity of $D(\theta)$, $D_n\left(\hat{\theta}_n\right) \to D\left(\theta_*\right)$ in probability, leading to the conclusion. Note that in the above Proposition, convergence in probability may be replaced by convergence with probability 1 (or almost surely) throughout since $\hat{\theta}_n \to \theta_*$ with probability 1 (see Theorem 2, Appendix A).

Remarks: Note that $D(\theta) = 0$ if the model is correctly specified. In White's Information Matrix test (Theorem 1, Appendix A), the test statistic is derived under the assumption that the model is correctly specified. In the above proposition, the model is not assumed to be correctly specified. If the model is not correctly specified, $D(\theta) \neq 0$, and as $n \to \infty$, $D_n\left(\hat{\theta}_n\right)$ tends to a non-zero $D\left(\theta_*\right)$ provided $\hat{\theta}_n \to \theta_*$. We do not want $D_n\left(\hat{\theta}_n\right)$ to tend to a zero value when the model is not correct. This is intuitive but it has to be proved mathematically.

(b) *Asymptotic normality of* $\sqrt{n}D_n(\theta)$

We define $\sqrt{n}D_n(\theta)=\sqrt{n}\left(\sum_{i=1}^{p}D_{ii}(\theta,n)\right)$, $i=1,2,\ .\ .\ .,p$. The asymptotic normality of $\sqrt{n}D_n(\theta)$ is given by Theorem 1 (see Appendix A). The details have been derived by White (1982).

3.3.3 The Bootstrap Procedure for Bartlett's First Identity Goodness-of-fit Test

If the null hypothesis in (3.1) is true, that the model is correctly specified, the finite sample distribution of the test statistic is dependent on the parameters of the model tested. Parametric bootstrap can be used to test (3.1). Therefore, Monte Carlo simulation is used to estimate the exact finite sample critical value of the test statistic from the sample data. For the IM test, Horowitz (1994) has shown that the finite sample critical values obtained from the bootstrap method are more accurate than the asymptotic χ^2 critical values.

Let $T(f_x)$ denote the BFI test statistic where f_x is the observed frequency. Let $F(t;\mu,\phi)$ denote the distribution function under the null hypothesis when μ and ϕ are the true but unknown NB parameter values. For a test at a significance level of α, let $c_{n,\alpha}$ denote the $1-\alpha$ quantile of $F(t;\mu,\phi)$ which is the critical value of $T(f_x)$. The Monte Carlo procedure to estimate the bootstrapped critical value is as follows:

1. Generate a random sample of size *n* from NB (μ,ϕ) with observed frequencies f_x, $x = 0,1,2,\ldots,k$.

2. Based on the sample from Step 1, estimate the sample mean $\hat{\mu}$. Conditional on $\hat{\mu}$, $\hat{\phi}$ is obtained by maximum likelihood estimation.

3. Calculate test statistic $T\left(f_x\right)$.

4. Replicate bootstrap samples, B_n as follows:

(a) Generate a random sample of size n from NB ($\hat{\mu},\hat{\phi}$) with observed frequencies f_{jx} , x = 0, 1,2,…,k, and $1 \le j \le B_n$.

(b) Obtain the bootstrap estimators $\hat{\mu}_j$ and $\hat{\phi}_j$, $1 \le j \le B_n$.

(c) Calculate the test statistic $T_j^*\left(f_{jx}^*\right)$, $1 \le j \le B_n$.

(d) Arrange $T_1^*, T_2^*, ..., T_{B_n}^*$ in ascending order to obtain

$$T_{1:B_n}^* \le T_{2:B_n}^* \le ... \le T_{B_n:B_n}^* .$$

5. $c_{n,\alpha}$ (see Baringhaus and Henze, 1992) of F^* is given by

$$c_{n,\alpha} = T_{\alpha_n:B_n}^* + \left(1-\gamma_n\right)\left[T_{\alpha_n+1:B_n}^* - T_{\alpha_n:B_n}^*\right],$$

where $\alpha_n = B_n - \left[\alpha\left(B_n+1\right)\right]$ and $\gamma_n = \alpha\left(B_n+1\right) - \left[\alpha\left(B_n+1\right)\right]$. Note that $F^*\left(t\right)$ is the empirical distribution function of $T_1^*, T_2^*, ..., T_{B_n}^*$ and $\left[u\right]$ represent the integer part of u. Baringhaus and Henze (1992) have advocated the idea that the bootstrap sample size $B_n = \max\left\{n, \left[1/\alpha\right]\right\}$ is sufficient to ensure that the actual level of the test is close to the chosen nominal level.

6. Reject the null hypothesis at level α if $T\left(f_x\right)$ exceeds $c_{n,\alpha}$.

N random samples, each of size *n*, are generated, and Steps 1-6 are implemented for each of them in order to estimate the significance level α. The estimated value of α is the proportion of number of times that the null hypothesis is rejected in the *N* random samples.

In Step 2, the NB parameter μ is estimated by the sample mean while the parameter ϕ is estimated by the ML method. It is known for the NB distribution that the ML estimate has small standard errors even for small or moderate sample sizes (*n* = 100 or 200). The optimum solution of ML estimator $\hat{\phi}$ may be obtained by using a stochastic optimization algorithm such as simulated annealing (see Kirkpatrick, 1983; Goffe, 1996).

Even though the bootstrapped critical values can bring the empirical level of the BFI test close to its nominal level, errors due to the bootstrap are still encountered. Beran (1988) suggested a technique called "prepivoting" to overcome this problem. In general, a nested double-bootstrap Monte Carlo simulation is used to refine the approximation of the actual significance level of the test. Nevertheless, Horowitz (1994) has pointed out that the Monte Carlo experiments are very time-consuming because for each replication of the Monte Carlo sample there are many inner replications in which the parameters of the null hypothesis are re-estimated from the corresponding bootstrap samples. Due to the connection of hypothesis testing with confidence interval, we propose to reduce the error of bootstrapped critical values by the BC_a method (Efron, 1993). The BC_a method is described further in the next section.

3.3.4 The Bias Corrected and Accelerated method

The bias corrected and accelerated $\left(BC_a\right)$ method is commonly used to correct deficiencies of the standard and percentile methods. Efron (1993) has defined the $\alpha-$level endpoint of the BC_a interval as

$$\hat{\phi}_{BC_a}[\alpha] = \hat{G}^{-1}\left(\Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{a}\left(\hat{z}_0 + z^{(\alpha)}\right)}\right)\right),$$

where $\hat{G}$ is the cumulative distribution function of the bootstrap replications $\hat{\phi}^*$, $z^{(\alpha)}$ is the 100αth percentile point of a standard normal distribution. The bias and acceleration adjustment in the BC_a method are $\hat{z}_0$ and $\hat{a}$, and Φ is the standard normal cumulative distribution function. The details for the construction of $\hat{z}_0$ and $\hat{a}$ are given by Efron (1993). Efron (1993) has shown that the BC_a interval is second order accurate, that is,

$$P\left(\phi \le \hat{\phi}_{BC_a}[\alpha]\right) = \alpha + O\left(n^{-1}\right).$$

Furthermore, the bootstrap-t endpoint $\hat{\phi}_{\text{STUD}}$ and BC_a endpoint agree to second order accurate (see Efron, 1993, p. 160-162):

$$\hat{\phi}_{BC_a}[\alpha] = \hat{\phi}_{\text{STUD}}[\alpha] + O_p\left(n^{-3/2}\right).$$

These facts can be proven by using Edgeworth expansions as given in Hall (1988).

To implement the BC_a method, we have modified step 5 of the Monte Carlo procedure:

$$c_{n,\alpha} = T^*_{\alpha_n:B_n},$$

where $\alpha_n = \left[\alpha B_n\right]$ and α is re-defined as:

$$\alpha = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{a}\left(\hat{z}_0 + z^{(\alpha)}\right)}\right).$$

The comparative simulation results for the BFI test statistic with and without the BC_a correction are reported in section 3.3.5.

3.3.5 Monte Carlo Experiments and Results

In this section, we report the results of a Monte Carlo investigation of the finite sample empirical level for the BFI test. Next, we examine the power of the BFI test applied in goodness-of-fit problem when the critical values are obtained using the bootstrap procedure in Section 3.3.3. The EDF tests and the recent implementation of the IM test (ω_B test) by Dhaene and Hoorelbeke (2004) are used as comparison.

Table 3.1: Maximum Likelihood estimates for NB, $\mu = 6.75$ and $\phi = 4.50$.

N	$\hat{\mu} = \bar{X}$	$\hat{\phi}$
100	7.0300	4.8719
200	6.8200	4.1269
500	6.5700	4.5666
1000	6.6980	4.3909
5000	6.7830	4.5790
10000	6.7815	4.5884

In the Monte Carlo investigation the effect of small sample size is considered. It is well-known that ML estimates have large biases when the sample size is small. Following Famoye (2000), a simulation study indicates that the ML estimates for the NB parameters have large biases for a sample size of n = 100. The bias reduces when n increases. As an illustration, Table 3.1 displays the simulation results based on 100 replications for the NB parameter $\mu = 6.75$ and $\phi = 4.5$. It is observed that the ML estimates approaches the true parameters value when n exceeds 1000. Therefore, sample sizes of n = 100 and n = 200 are considered small for the NB model.

Table 3.2: Empirical level for the BFI test with and without BC_a correction, $\alpha = 0.05$.

μ	ϕ	n	B_n	*BFI* (with BC_a)	*BFI* (without BC_a)
1.0	4.0	100	100	0.005	0.000
6.0	9.0	100	100	0.019	0.008
3.15	2.1	100	100	0.051	0.028
2.0	0.5	100	100	0.050	0.016

B_n – number of bootstrap samples

Table 3.3: Empirical level for the EDF, BFI and ω_B tests, $\alpha = 0.05$.

μ	ϕ	n	B_n	K_d	W_d	A_d	U_d	BFI	ω_B
1.11	10.0	100	100	0.053	0.045	0.051	0.053	0.005	0.002
5.11	46.0			0.062	0.053	0.054	0.063	0.006	0.001
1.11	10.0	200	200	0.040	0.040	0.036	0.050	0.004	0.001
5.11	46.0			0.054	0.053	0.054	0.052	0.001	0.001
1.0	4.0	100	100	0.048	0.049	0.047	0.049	0.011	0.005
4.0	16.0			0.071	0.057	0.054	0.077	0.012	0.002
1.0	4.0	200	200	0.045	0.044	0.041	0.044	0.016	0.004
4.0	16.0			0.054	0.055	0.050	0.081	0.009	0.004
2.0	3.0	100	100	0.063	0.064	0.070	0.058	0.028	0.025
6.0	9.0			0.063	0.072	0.072	0.071	0.030	0.019
2.0	3.0	200	200	0.072	0.071	0.070	0.071	0.029	0.028
6.0	9.0			0.063	0.066	0.055	0.056	0.030	0.020
3.15	2.1	100	100	0.045	0.047	0.053	0.065	0.021	0.051
6.75	4.5			0.065	0.067	0.066	0.063	0.035	0.061
3.15	2.1	200	200	0.045	0.098	0.093	0.058	0.032	0.067
6.75	4.5			0.065	0.068	0.069	0.063	0.035	0.080
2.0	0.5	100	100	0.064	0.068	0.065	0.069	0.050	0.084
8.0	2.0			0.067	0.063	0.066	0.057	0.032	0.125
2.0	0.5	200	200	0.050	0.045	0.043	0.045	0.055	0.110
8.0	2.0			0.073	0.064	0.066	0.071	0.065	0.163

The simulation study for the EDF test statistics, BFI test statistics and ω_B test statistics have been developed for different values of parameters μ and ϕ. To study the effect of tail lengths, the combinations of NB parameters are chosen to represent short ($\mu < \phi$) and long tails ($\mu > \phi$) of the distribution. The empirical levels for the EDF and BFI test are based on 1000 Monte Carlo samples. Each entry in Table 3.3 and 3.4 represents the proportion of 1000 Monte Carlo samples declared significant by each test using bootstrapped critical values. It is found that the EDF test statistics have estimated significance levels of α very close to the chosen nominal level for n = 100. For BFI test and ω_B test, the empirical levels of the tests tend to be smaller than the chosen nominal level for certain parameter values where the NB distribution has a shorter tail. However, empirical $\alpha-$levels of the BFI test are closer to the selected nominal level when the NB distribution has a longer tail while the ω_B test has empirical levels exceeding the nominal level. This is due to the use of the

Hotelling's T^2 distribution. For example, by bootstrapping critical values, as suggested by Horowitz (1994), for the parameters $\mu = 8.0$ and $\phi = 2.0$ gives the empirical level for ω_B test as 0.063 instead of 0.125. Noted that we have only reported the results of ω_B test for $\alpha = 0.05$ which is in Table 3.3 since the test has same performance for other significance levels. Table 3.2 gives the empirical level for the BFI test with and without BC_a correction for some parameter combinations. The empirical level for the BFI test is closer to the nominal level with BC_a correction. The empirical levels from the simulation study with 1000 Monte Carlo samples are reported in Tables 3.3 and 3.4. Each entry in the tables represents the proportion of Monte Carlo samples rejected by each test using the various critical values discussed in section 3.3.3. For $\alpha = 0.05$, the results of empirical level of the EDF and BFI tests are very similar to those for $\alpha = 0.10$.

Table 3.4: Empirical level for the EDF and BFI tests, $\alpha = 0.10$.

μ	ϕ	n	B_n	K_d	W_d	A_d	U_d	BFI
1.11	10.0	100	100	0.111	0.109	0.117	0.106	0.010
5.11	46.0			0.108	0.100	0.098	0.095	0.015
1.11	10.0	200	200	0.108	0.101	0.097	0.102	0.005
5.11	46.0			0.108	0.106	0.108	0.104	0.006
1.0	4.0	100	100	0.085	0.084	0.090	0.087	0.026
4.0	16.0			0.129	0.137	0.140	0.139	0.013
1.0	4.0	200	200	0.085	0.087	0.088	0.087	0.030
4.0	16.0			0.134	0.135	0.143	0.133	0.014
2.0	3.0	100	100	0.108	0.118	0.121	0.115	0.069
6.0	9.0			0.107	0.121	0.122	0.127	0.087
2.0	3.0	200	200	0.116	0.130	0.136	0.124	0.060
6.0	9.0			0.106	0.109	0.108	0.107	0.060
3.15	2.1	100	100	0.107	0.112	0.122	0.113	0.101
6.75	4.5			0.110	0.120	0.125	0.113	0.127
3.15	2.1	200	200	0.107	0.119	0.115	0.119	0.098
6.75	4.5			0.115	0.114	0.117	0.114	0.103
8.0	2.0	100	100	0.114	0.103	0.113	0.104	0.102
2.0	0.5			0.117	0.107	0.119	0.106	0.103
8.0	2.0	200	200	0.118	0.116	0.119	0.115	0.119
2.0	0.5			0.107	0.094	0.094	0.095	0.103

In the next part of the simulation study, the powers of the test statistics are compared for two alternative hypotheses H_a:

(a) Neyman type-A distribution with parameters λ and θ, NTA$\left(\lambda,\theta\right)$.

(b) Poisson-inverse Gaussian with parameters γ and β, P-IG$\left(\gamma,\beta\right)$.

The NTA has pmf (Johnson, Kemp and Kotz, 2005, p. 404) given by

$$P\left(x;\lambda,\theta\right)=\frac{e^{-\lambda}\theta^{x}}{x!}\sum_{j=0}^{\infty}\frac{\left(\lambda e^{-\theta}\right)^{j}j^{x}}{j!}, \qquad x=0,1,2,...$$

$\theta>0$, and it satisfies the recurrence relation

$$P\left(x\right)=\frac{\lambda\theta e^{-\theta}}{x}\sum_{j=0}^{x-1}\frac{\theta^{j}}{j!}P\left(x-1-j\right), \qquad x\geq1$$

where $P\left(0\right)=e^{-\lambda+\lambda e^{-\theta}}$.

The pmf of the P-IG distribution (Willmot, 1987) is

$$P(x;\gamma,\beta)=P(0)\frac{\gamma^{x}}{x!}\sum_{j=0}^{x-1}\frac{\left(x-1+j\right)\,!}{\left(x-1-j\right)\,!\,j\,!}\left(\frac{\beta}{2\lambda}\right)^{j}\left(1+2\beta\right)^{-(x+j)/2}, \quad x=1,2,3,\,...,$$

$\gamma>0$, $\beta>0$ and it has the following recursion formula

$$P(x)=\frac{2\beta}{1+2\beta}\left(1-\frac{3}{2x}\right)P\left(x-1\right)+\frac{\gamma^{2}}{x\left(x-1\right)\left(1+2\beta\right)}P\left(x-2\right), \quad x\geq2,$$

where $P\left(0\right)=e^{\gamma\beta^{-1}\left\{1-\left(1+2\beta\right)^{1/2}\right\}}$ and $P\left(1\right)=\gamma\left(1+2\beta\right)^{-1/2}P\left(0\right)$.

The NTA and P-IG distributions have been chosen because they are popular distributions in various applications. The P-IG distribution is known to have a behavior similar to the NB distribution and has been proposed as its alternative (Willmot, 1987) while the NTA distribution can be multimodal. Tables 3.5 and 3.6 show the results of 1000 Monte Carlo samples generated from distributions under the alternative hypotheses, for sample sizes of

n = 100 and n = 200, and tested for the NB model. The modified Anderson-Darling EDF test performs well. When the model is correctly specified (NB), the BFI test performs the best. When the samples come from the NTA or P-IG distributions, it is comparable to the other tests. Among the EDF tests, the modified Anderson-Darling test performs well. Overall, the BFI test is comparable or more powerful than the other tests.

Table 3.5: Power comparison for the EDF tests and BFI test when $\alpha = 0.05$ and n = 100.

Model	K_d	W_d	A_d	U_d	*BFI*
NB(1.11,10.0)	0.053	0.045	0.051	0.053	0.005
NB(1.0,4.0)	0.048	0.049	0.047	0.049	0.011
NB(6.75,4.5)	0.065	0.067	0.066	0.063	0.035
NB(8.0,2.0)	0.067	0.063	0.066	0.057	0.032
P-IG(2.50,5.0)	0.202	0.271	0.305	0.239	0.217
P-IG(4.75,2.0)	0.104	0.104	0.117	0.099	0.152
P-IG(8.0,7.5)	0.389	0.472	0.481	0.473	0.319
NTA(2.0,3.5)	0.635	0.540	0.585	0.747	0.549
NTA(4.75,1.5)	0.114	0.130	0.135	0.103	0.150
NTA(8.93,0.58)	0.070	0.073	0.072	0.068	0.099

Table 3.6: Power comparison for the EDF tests and BFI test when $\alpha = 0.05$ and n = 200.

Model	K_d	W_d	A_d	U_d	*BFI*
NB(1.11,10.0)	0.040	0.040	0.036	0.050	0.004
NB(1.0,4.0)	0.045	0.044	0.041	0.044	0.016
NB(6.75,4.5)	0.065	0.068	0.069	0.063	0.035
NB(8.0,2.0)	0.073	0.064	0.066	0.071	0.065
P-IG(2.50,5.0)	0.423	0.490	0.606	0.633	0.605
P-IG(4.75,2.0)	0.172	0.159	0.207	0.240	0.190
P-IG(8.0,7.5)	0.560	0.737	0.750	0.686	0.655
NTA(2.0,3.5)	0.929	0.972	0.990	0.956	0.872
NTA(4.75,1.5)	0.144	0.170	0.193	0.141	0.164
NTA(8.93,0.58)	0.070	0.061	0.064	0.057	0.064

3.4 Comparison between Bartlett's First Identity Goodness-of-fit Test and Jarque-Bera test

The Jarque-Bera test (JB) is well known as a goodness-of-fit test for normality. It is a particular case of White's IM test if the distribution is normal and the parameters are estimated by maximum likelihood estimation. The objective of this section is to derive the BFI test for the normal distribution and compare it with the JB test.

The JB test (Jarque and Bera, 1987) is defined as

$$JB = \frac{n}{6}.\left(S^2 + \frac{(k-3)^2}{4} \right)$$

where the sample skewness and the sample kurtosis are given by $S = \frac{\hat{\mu}_3}{\hat{\mu}_2^{3/2}}$ and $K = \frac{\hat{\mu}_4}{\hat{\mu}_2^{\,2}}$.

μ_2, μ_3 and μ_4 are the theoretical second, third and fourth central moments respectively and they are estimated by

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^j, \quad j = 2,3,4.$$

Bowman and Shenton (1975) have stated that the JB statistic is the sum of squares of two asymptotically independent standardized normal distributions. Therefore, *JB* is asymptotically chi-squared distributed with two degrees of freedom. We reject H_0 at $\alpha -$significant level if $JB \geq \chi^2_{1-\alpha,2}$.

If we consider the normal distribution and estimate parameters by maximum likelihood estimation, then White's IM test is exactly the JB test. However, it is cumbersome to implement White's IM test since we need to compute the covariance matrix. Therefore, the BFI goodness-of-fit test as defined in eq. (3.5) is used for a comparison with JB test. Let $f(x) = f(x;\hat{\mu},\hat{\sigma})$. Since normal random variable is continuous, we have rewritten equations (3.3) and (3.4) as

$$A_{ij}\left(\theta,n\right) = n^{-1}\sum_{t=1}^{n}\frac{\partial^2 \ln f\left(x_t;\theta\right)}{\partial\theta_i\partial\theta_j},$$

$$B_{ij}\left(\theta,n\right) = n^{-1}\sum_{t=1}^{n}\left[\frac{\partial \ln f\left(x_t;\theta\right)}{\partial\theta_i}.\frac{\partial \ln f\left(x_t;\theta\right)}{\partial\theta_j}\right].$$

The first and second derivatives of the natural log-likelihood function for the two parameters are as given below:

$$\frac{\partial \ln f(x)}{\partial \mu} = \frac{x-\mu}{\sigma^2}, \frac{\partial^2 \ln f(x)}{\partial \mu^2} = -\frac{1}{\sigma^2};$$

$$\frac{\partial \ln f(x)}{\partial \sigma^2} = \frac{1}{2} . \frac{(x-\mu)^2}{\sigma^4} - \frac{1}{2\sigma^2}, \frac{\partial^2 \ln f(x)}{\partial (\sigma^2)^2} = \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6}.$$

For the normal goodness-of-fit problem, the goodness-of-fit test statistic is given by

$$D = tr\left[A(\hat{\mu},\hat{\sigma};n) + B(\hat{\mu},\hat{\sigma};n)\right].$$

The rejection region, *R* of the test statistic is given by $R = P\left[D \geq c_{n,\alpha} \middle| H_0 \text{ is true}\right]$ where $c_{n,\alpha}$ is a bootstrapped critical value.

We have conducted a simulation study on the empirical level of the test. The results based on 1000 Monte Carlo samples are reported in Tables 3.7 and 3.8. To compare the JB test and BFI goodness-of-fit test, the empirical values have been computed according to different $\alpha-$significance levels and sample sizes *n*.

Table 3.7: Empirical level for the BFI test.

	n					
α	10	20	50	100	200	500
0.01	0.006	0.007	0.012	0.015	0.013	0.015
0.02	0.016	0.020	0.023	0.024	0.022	0.021
0.05	0.046	0.052	0.049	0.052	0.051	0.048
0.10	0.099	0.108	0.104	0.106	0.099	0.096
0.20	0.198	0.207	0.213	0.199	0.203	0.201

Table 3.8: Empirical level for the JB test.

	n					
α	10	20	50	100	200	500
0.01	0.003	0.008	0.015	0.019	0.026	0.018
0.02	0.006	0.013	0.019	0.023	0.032	0.022
0.05	0.009	0.023	0.032	0.048	0.047	0.041
0.10	0.015	0.040	0.050	0.074	0.076	0.096
0.20	0.026	0.064	0.084	0.129	0.160	0.172

Overall, we observe that the computed empirical values are quite close to the nominal level. In some cases ($\alpha = 0.05, 0.10, 0.20$) when the sample sizes are small, the empirical values of JB test are even smaller than the nominal level. However, the results for the BFI test seem to agree with the JB test as the sample size increases.

3.5 Conclusion

Based on the simulation study conducted, it appears that the proposed Bartlett First Identity test under orthogonality of model parameters, as exemplified by the negative binomial distribution, is viable as a goodness-of-fit test. Although the proposed test is closely related to White's Information Matrix test, parameter orthogonality and obviating the need to use the covariance matrix make the proposed test statistic much simpler than the information test. The trace of the Hessian and outer product forms of the information matrix has been considered. If a simpler goodness-of-fit test statistic is required, the modified Anderson-Darling test is to be recommended and this agrees with Famoye's (2000) recommendation in his study of the generalized logarithmic distribution. However, if better power is needed, the BFI test can be considered. For the comparison with JB test, the BFI test has almost similar performance with JB test.

CHAPTER 4

PARAMETERS ORTHOGONAL TO THE MEAN

4.1 Introduction

Two parameters of a distribution are said to be orthogonal if their maximum likelihood estimators are asymptotically uncorrelated. Parameter orthogonality has many advantages in statistical inference and a good review has been given by Cox and Reid (1987). For instance, in maximum likelihood estimation, reparameterization in terms of orthogonal parameters speeds up convergence of the numerical method employed (Sprott, 1983). Willmot (1990) has considered orthogonality of a class of discrete models with two unknown parameters reparameterized in terms of the mean μ with respect to the remaining parameter. Here we examine the orthogonality of μ for models with *more than two* parameters, where the remaining parameters are regarded as nuisance parameters. In particular we derive orthogonality for a class of Poisson-convolution models. The Poisson-convolution is defined as $Z = X + Y$ where X is a Poisson random variable (rv) with mean λ and Y is another nonnegative integer valued rv with X and Y being independent. This convolution may be regarded as a signal-plus-noise model and it occurs in many settings (see Samaniego, 1976, and references therein) to make it a model of practical importance. Examples of applications are aerial prospecting for uranium, synaptic transmission of neural impulses, misclassification model and stochastic occupancy model in hospitals (Shonick, 1970).

As an application of the orthogonality result, a uniformly most powerful test of the mean is developed based on the asymptotic result under model misspecification. The test of mean is exemplified for the convolution of Poisson and negative binomial variables. A small Monte Carlo power study of the proposed test has been conducted.

This chapter is organized as follows: In section 4.2, the condition for parameters orthogonal to the mean is derived. Section 4.3 proposes the test of mean based upon asymptotic normality of the maximum likelihood estimators under model misspecification. Section 4.4 discusses a small Monte Carlo power study of the proposed test.

4.2 Parameters Orthogonal to the Mean

4.2.1 Condition for Orthogonal Parameters

Let $f(x)$ be a probability mass or density function which involves r (≥ 2) unknown parameters $\theta_1, \ldots, \theta_r$, among which θ_1 equals the population mean. Willmot (1990) considered the case $r = 2$ and our primary interest lies in the situation $r \geq 3$. We assume standard regularity conditions. Let $\theta = (\theta_1, ..., \theta_r)^T$, the parameter space for θ being an open set in the r-dimensional Euclidean space. Denote the per observation Fisher information matrix by $A = (a_{ij})$, where

$$a_{ij} \equiv a_{ij}(\theta) = E_\theta[\{\partial \log f(x) / \partial \theta_i\}\{\partial \log f(x) / \partial \theta_j\}], \qquad 1 \leq i, j \leq r \tag{4.1}$$

It is supposed that A is positive definite at every θ. Let $M(t)$ be the moment generating function corresponding to $f(x)$ and $C \equiv C(t) = \log M(t)$ be the cumulant generating function. We assume that for any θ, $M(t)$ exists finitely for t in an interval in the positive part of the real line.

Suppose there exist functions $g_j(\theta)$ $(2 \le j \le r)$ and $h(\theta)$ such that

$$\frac{\partial \log f(x)}{\partial \theta_1} + \sum_{j=2}^{r} g_j(\theta) \frac{\partial \log f(x)}{\partial \theta_j} + h(\theta)(x-\theta_1) = 0, \tag{4.2}$$

or equivalently,

$$\frac{\partial C}{\partial \theta_1} + \sum_{j=2}^{r} g_j(\theta) \frac{\partial C}{\partial \theta_j} + h(\theta)(\frac{\partial C}{\partial t} - \theta_1) = 0. \tag{4.3}$$

The equivalence of (4.2) and (4.3) holds because

$$M(t) \times \{\text{LHS of (4.3)}\} = \int e^{tx} \{\text{LHS of (4.2)}\} \; f(x)dx\,.$$

For $2 \le j \le r$ differentiating the identity $\int (x-\theta_1) f(x)dx = 0$ with respect to θ_j, we get

$$E_\theta[(x-\theta_1)\{\partial \log f(x)/\partial \theta_j\}] = 0.$$

Hence by (4.1) and (4.2),

$$a_{11} = \{h(\theta)\sigma\}^2 + g^T A_{22} g\,, \qquad (a_{12},\ldots,a_{1r}) = -g^T A_{22}\,,$$

where $\sigma \equiv \sigma(\theta)$ is the population standard deviation, $g = (g_2(\theta),\ldots,g_r(\theta))^T$ and A_{22} is the principal submatrix of A given by the last $r-1$ rows and columns of the latter. Thus

$$A = \begin{bmatrix} \{h(\theta)\sigma\}^2 + g^T A_{22} g & -g^T A_{22} \\ -A_{22} g & A_{22} \end{bmatrix}, \tag{4.4}$$

Note that det(A) = $\{h(\theta)\sigma\}^2$ $\det(A_{22})$ so that $h(\theta)\sigma \ne 0$, because A is positive definite. Hence writing $A^{-1} = (a^{ij})$, from (4.4), we get

$$a^{11} = \{h(\theta)\sigma\}^{-2}\,, \qquad (a^{12},\ldots,a^{1r}) = \{h(\theta)\sigma\}^{-2} g^T\,.$$

Therefore, as in Willmot (1990), a parameterization $\phi = (\phi_1,\ldots,\phi_r)^T$, where $\phi_1 = \theta_1$ and ϕ is a one to one function of θ, ensures parametric orthogonality with respect to $\phi_1(=\theta_1)$ if and only if

$$\frac{\partial \phi_k}{\partial \theta_1} + \sum_{j=2}^{r} g_j(\theta) \frac{\partial \phi_k}{\partial \theta_j} = 0, \qquad 2 \le k \le r. \tag{4.5}$$

Equation (4.5) extends the main result of Willmot (1990) to the multiparameter case.

There are many models, such as those based on convolutions or compound distributions, where the form of $f(x)$ is complicated and hence an explicit determination of the information matrix A is difficult. However, such models often entail a relatively simple form of $C(t)$, so that one can check if (4.3) holds for some $g_j(\theta)$ $(2 \le j \le r)$ and $h(\theta)$. If indeed (4.3) holds, then (4.2) also holds and hence an orthogonal parameterization with respect to θ_1 can be obtained via (4.5) even without explicit determination of A. In the next section we consider parameters orthogonal to the mean for Poisson-convolution models.

4.2.2 Orthogonal Parameters for Poisson-convolution Models

Consider the convolution of the Poisson distribution with mean λ and a power series distribution with probability mass function of the form $\psi(x) = b(x\,;\xi)\rho^x / B(\xi,\rho)$ $(x = 0,1,2,\ldots)$, where $B(\xi,\rho) = \sum_{x=0}^{\infty} b(x\,;\xi)\rho^x$. The unknown parameters are λ, ρ and ξ; among these, $\lambda\,(> 0)$ and $\rho\,(> 0)$ are scalar-valued while ξ is possibly vector-valued. Thus altogether there are $r = s + 2$ parameters, where $s\,(\ge 1)$ is the dimension of ξ. Here

$$M(t) = [\exp\{\lambda(e^t - 1)\}]\ \{B(\xi,\rho e^t)/B(\xi,\rho)\}. \tag{4.6}$$

Let $Q(\xi,\rho) = \log B(\xi,\rho)$, $Q_v(\xi,\rho)$ = $\partial^v Q(\xi,\rho)/\partial\rho^v$ (v = 1,2). Then by (4.6), $C \equiv C(t) = \lambda(e^t - 1) + Q(\xi,\rho e^t) - Q(\xi,\rho)$, and the population mean is given by $\mu =$ $\{\partial C/\partial t\}_{t=0} = \lambda + \rho Q_1(\xi,\rho)$, i.e., $\lambda = \mu - \rho Q_1(\xi,\rho)$. One can express C in terms of μ, ρ and ξ as

$$C = \{\mu - \rho Q_1(\xi,\rho)\}\,(e^t - 1) \; + Q(\xi,\rho e^t) - Q(\xi,\rho)\,. \tag{4.7}$$

Therefore,

$$\frac{\partial C}{\partial \mu} = e^t - 1, \quad \frac{\partial C}{\partial \rho} = -\{Q_1(\xi,\rho) + \rho Q_2(\xi,\rho)\}\,(e^t - 1) + e^t Q_1(\xi,\rho e^t) - Q_1(\xi,\rho),$$

$$\frac{\partial C}{\partial t} - \mu = \mu(e^t - 1) + \rho e^t\{Q_1(\xi,\rho e^t) - Q_1(\xi,\rho)\},$$

so that

$$\frac{\partial C}{\partial \mu} + \frac{\rho}{\mu + \rho^2 Q_2(\xi,\rho)}\frac{\partial C}{\partial \rho} - \frac{1}{\mu + \rho^2 Q_2(\xi,\rho)}\,(\frac{\partial C}{\partial t} - \mu) = 0. \tag{4.8}$$

Note that by (4.7), $\{\partial^2 C/\partial t^2\}_{t=0} = \mu + \rho^2 Q_2(\xi,\rho)$, i.e., $\mu + \rho^2 Q_2(\xi,\rho)$ equals the population variance and hence is positive; thus the left-hand side of (4.8) is well-defined. Comparing (4.8) with (4.3), it follows from (4.5) that a parameterization $\phi = (\phi_1,\ldots,\phi_r)^T$, where $\phi_1 = \mu$, ensures parametric orthogonality with respect to $\phi_1 (= \mu)$ if and only if

$$\frac{\partial \phi_k}{\partial \mu} + \frac{\rho}{\mu + \rho^2 Q_2(\xi,\rho)}\frac{\partial \phi_k}{\partial \rho} = 0, \qquad 2 \le k \le r. \tag{4.9}$$

It is easily seen that the conditions in (4.9) are met by

$$\phi_1 = \mu\,, \qquad \phi_2 = (\mu/\rho) - Q_1(\xi,\rho), \qquad (\phi_3,\ldots,\phi_r)^T = \xi\,. \tag{4.10}$$

As a specific application, consider the convolution of the Poisson and negative binomial distributions. Then $b(x\,;\xi) = \begin{pmatrix} \xi + x - 1 \\ x \end{pmatrix}$, $B(\xi,\rho) = (1-\rho)^{-\xi}$, $0 < \rho < 1$ and $\xi\ (>0)$ is scalar-valued. Consequently, $Q_1(\xi,\rho) = \xi/(1-\rho)$, and by (4.10), an orthogonal parameterization with respect to μ is given by

$$\phi_1 = \mu\,, \qquad \phi_2 = \frac{\mu}{\rho} - \frac{\xi}{1-\rho}, \qquad \phi_3 = \xi\,. \tag{4.11}$$

Note that the transformation in (4.11) is one to one because $\partial\phi_2 / \partial\rho < 0$, and for fixed $\mu, \xi,$ the parameter ϕ_2 tends to $+\infty$ or $-\infty$ as ρ tends to 0 or 1 respectively.

4.3 Test of the Mean under Model Misspecification

4.3.1 Uniformly Most Powerful Test of Hypotheses

Let $Z_1, Z_2, \ldots, Z_n$ be *n* rv's with pmf $p(z;\boldsymbol{\theta})$, parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_q)$. Let $\theta_1 = \mu$ be the parameter of interest.

The hypothesis to be tested is $H_0 : \mu \leq \mu_0$ versus the alternative $H_1 : \mu > \mu_0$. A uniformly most powerful (UMP) test may be developed based upon the asymptotic distribution of the QML estimators derived by White (1982) under the assumption of model misspecification. White (1982) called the ML estimators of $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_q)$ under the misspecified model as QML estimators. Under a correctly specified model it is well-known that $\sqrt{n}(\hat{\theta} - \theta)$ converges in distribution to $N(0, I^{-1})$ where I is the information matrix. With model misspecification, the QML estimators are also asymptotically normal but the covariance matrix I^{-1} is replaced by $I^{-1} R I^{-1}$ (Theorem 3.2, White, 1982), where $I = -E[\partial^2 \ln L / \partial\theta^2]$ (Hessian form) and $R = E[(\partial \ln L / \partial\theta)^2]$ (outer product form). Note that if the model is correctly specified, $I = R$ (Bartlett First Identity).

The UMP test for the mean involving the asymptotic normal distribution is given in the ensuing result.

Result 4.3.1: Let $X_1, X_2, \ . \ . \ , X_n$ be a random sample from a pdf $f(x;\theta)$ with parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_q)$ and let θ_1 be the parameter of interest. Suppose that $\sqrt{n}(\hat{\theta}_1 - \theta_1)$ converges in distribution to $N(0, \hat{\sigma}^2)$.

A uniformly most powerful test for the hypotheses $H_0 : \theta_1 \leq \theta_0$ versus $H_1 : \theta_1 > \theta_0$ is given by the critical region of size α

$$C = \left\{ (x_1, x_2, \ . \ . \ ., x_n) : \frac{\sqrt{n}(\hat{\theta}_1 - \theta_1)}{\hat{\sigma}} > c \right\}, \tag{4.12}$$

where *c* is the critical value determined from $\Pr((X_1, X_2, . \ ., X_n) \in C; H_0) = \alpha.$

The proof of Result 4.3.1 is straightforward and is given in Appendix C.

4.3.2 Uniformly Most Powerful Test for Mean of Convolution of Poisson and Power Series Distributions

Suppose the parameters of the convolution of Poisson and power series distributions are (ϕ_1, ϕ_2, ϕ_3), where $\phi_1 = \mu$ is the mean, and the remaining nuisance parameters $(\phi_2, \phi_3 = \xi)$ are orthogonal to μ. Under model misspecification, $\sqrt{n}(\hat{\mu} - \mu)$ converges in distribution to the normal distribution $N(0, I^{-1} R I^{-1})$. The critical region is given by (4.12). A UMP test is constructed based on model misspecification because the true model for a data generating process is seldom known. As a consequence, this test is expected to be robust. In general, we give the formula of $\hat{\sigma}^2$ for these convolution distributions. Let

$$I = -E\left[\frac{\partial^2 \ell n L}{\partial \theta^2}\right], R = E\left[\left(\frac{\partial \ell n L}{\partial \theta}\right)^2\right], \text{ and } P_k = P(X = k) = P(x; \phi_1, \phi_2, \phi_3).$$

where θ represents the parameters. The log-likelihood function for a given group data with sample frequency, f_k where *c* is the largest count data is defined as

$$\ln L = \sum_{k=1}^{c} f_k \ln P_k \ ,$$

and the partial derivative with respect to θ are

$$\frac{\partial \ln L}{\partial \theta} = \sum_k f_k \frac{\partial \ln P_k}{\partial \theta}, \quad \frac{\partial^2 \ln L}{\partial \theta^2} = \sum_k f_k \frac{\partial^2 \ln P_k}{\partial \theta^2}.$$

Under model misspecification, $\hat{\sigma}^2$ is obtained from the Hessian, *I*, and outer product matrices, where *I* is expressed as

$$I = -E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right] = -\sum_k E\left[f_k\right] \frac{\partial^2 \ln P_k}{\partial \theta^2} = -\sum_k nP_k \frac{\partial^2 \ln P_k}{\partial \theta^2} = n\sum_k \frac{1}{P_k}\left(\frac{\partial P_k}{\partial \theta}\right)^2,$$

because $n\sum_k \frac{\partial^2 \ln P_k}{\partial \theta^2} = 0.$ The expression for *R* is

$$\begin{aligned} R &= E\left[\left(\frac{\partial \ell nL}{\partial \theta}\right)^2\right] \\ &= E\left[\sum_k f_k \frac{\partial \ln P_k}{\partial \theta} \sum_j f_j \frac{\partial \ln P_j}{\partial \theta}\right] = E\left[\sum_k \sum_j f_k f_j \frac{1}{P_k P_j} \frac{\partial P_k}{\partial \theta} \frac{\partial P_j}{\partial \theta}\right] \\ &= \sum_k \sum_j E\left[f_k f_j\right] \frac{1}{P_k P_j} \frac{\partial P_k}{\partial \theta} \frac{\partial P_j}{\partial \theta}. \end{aligned} \tag{4.13}$$

Since $\left(f_0, f_1, ..., f_c\right)$ follows a multinomial distribution with parameters $\left(n, P_0, P_1, ..., P_c\right)$, we have

$$E\left[f_i\right] = nP_i, \operatorname{var}\left(f_i\right) = nP_i\left(1-P_i\right), \operatorname{cov}\left(f_i f_j\right) = -nP_i P_j,$$

which lead to

$$E\left[f_i f_j\right] = -nP_i P_j, \ E\left[f_i^2\right] = nP_i\left(1-P_i\right) + n^2 P_i^2. \tag{4.14}$$

Eq. (4.14) is substituted into (4.13) to compute *R*.

For comparison, the Monte Carlo simulation study on power of the UMP test given in the next section is constructed under two different approaches: (a) UMP test when the model under misspecification where $\hat{\sigma}^2$ is computed from covariance matrix, $I^{-1}R\,I^{-1}$, (b) UMP test when the model is correctly specified where $\hat{\sigma}^2$ is computed from covariance matrix, I^{-1}.

Since the Poisson-convolution distributions have many useful applications as signal plus noise models, the UMP test may prove useful. For instance, the mean μ may represent the average daily arrival of patients at a hospital (Shonick, 1970) and it is of interest to determine if this mean exceeds μ_0.

4.4 Monte Carlo Simulation Study on Power of Uniformly Most Powerful Test

A Monte Carlo simulation study is conducted for the convolution of Poisson and negative binomial distributions to study the effect on the power of the test due to the (a) finite (small) sample size, and (b) substitution of arbitrary values for the nuisance parameters $(\phi_2, \phi_3 = \xi)$. Note that due to parameter orthogonality, the variance of the ML estimator $\hat{\mu}$ is not affected by the estimators of $(\phi_2, \phi_3 = \xi)$. The robustness of the test due to model misspecification is also examined. Note that the UMP test is constructed based on two approaches (see Section 4.3.2) and are defined as UMP test under model misspecification, *UMP*$_{\mathrm{W}}$, and UMP test under correctly specified model, *UMP*$_{\mathrm{IM}}$.

The results shown here are based on 5000 Monte Carlo samples. The parameters, $\hat{\phi}_2$ and $\hat{\xi}$ are estimated using the maximum likelihood method. Table 4.1 gives the results of a Monte Carlo study of the effect of finite sample sizes on the empirical level of the UMP test. Random samples are generated from the convolution of Poisson and binomial distribution with the value of the parameter, μ_0, stated in $H_0 : \mu \leq \mu_0$, where μ_0 is set to be 6.15, 7.45, 9.85 and 13.00 respectively as shown in Table 4.1.

Table 4.1: Empirical level for the UMP test for mean, $\alpha = 0.05$.

$\mu = \mu_0$	ϕ_2	ξ	n	UMP_{W}	UMP_{IM}
6.15	1.32	3.65	50	0.062	0.072
			100	0.060	0.068
			200	0.048	0.024
			500	0.034	0.012
7.45	3.15	2.80	50	0.058	0.194
			100	0.051	0.100
			200	0.050	0.061
			500	0.030	0.011
9.85	5.17	4.50	50	0.046	0.141
			100	0.038	0.068
			200	0.024	0.057
			500	0.012	0.013
13.00	5.56	3.50	50	0.054	0.165
			100	0.032	0.128
			200	0.036	0.086
			500	0.026	0.053

It is observed that the empirical levels of the UMP_{W} test tend to be closer to the nominal level as the sample size increases. When the model is correctly specified, UMP_{W} is more conservative in terms of the power of the test compared to UMP_{IM} especially for the case when the sample sizes are small, for example, n = 50 and n = 100. The choice of the sample size has a significant influence on the power of UMP_{IM}. Table 4.2 reports the simulation results on the power study for alternative hypotheses H_a with $\mu > 6.15$ for the convolution of Poisson and negative binomial model. Note that ϕ_2 and ξ are fixed at 1.32 and 3.65 respectively for all the tested alternative hypotheses. Monte Carlo samples that are generated from distributions with different values of mean μ under the alternative hypotheses for some different sample sizes are tested. When the mean μ is set to a value far bigger than the mean value under null hypothesis, we find that the UMP test has great power even for small sample sizes. The UMP_{W} and UMP_{IM} have similar performances when the values of mean, μ is set greater than the value specified in the null hypothesis.

Table 4.2: Power comparison for the UMP test for mean, $\alpha = 0.05$.

μ	ϕ_2	ξ	n	UMP_{W}	UMP_{IM}
6.50	1.32	3.65	50	0.080	0.072
			100	0.132	0.210
			200	0.224	0.306
			500	0.404	0.416
7.50	1.32	3.65	50	0.446	0.418
			100	0.816	0.936
			200	0.982	0.994
			500	1.000	1.000
9.00	1.32	3.65	50	0.939	1.000
			100	1.000	1.000
			200	1.000	1.000
			500	1.000	1.000
12.00	1.32	3.65	50	1.000	1.000
			100	1.000	1.000
			200	1.000	1.000
			500	1.000	1.000

Table 4.3: Empirical level for the UMP test for mean, $\alpha = 0.05$ and $n = 50$.
(ϕ_2^* and ξ^* are the arbitrary values for the nuisance parameters ϕ_2 and ξ)

μ	ϕ_2	ξ	ϕ_2^*	ξ^*			
				1.0	2.5	4.0	6.5
6.15	1.32	3.65	0.8	0.066	0.028	0.014	0.008
			1.0	0.076	0.032	0.014	0.008
			3.0	0.064	0.058	0.026	0.004
			5.0	0.079	0.062	0.036	0.014
7.45	3.15	2.80	0.8	0.052	0.024	0.008	0.008
			1.0	0.060	0.024	0.010	0.008
			3.0	0.055	0.050	0.022	0.008
			5.0	0.078	0.062	0.032	0.010
9.85	5.17	4.50	1.0	0.012	0.006	0.006	0.006
			3.0	0.044	0.010	0.006	0.006
			5.0	0.030	0.018	0.008	0.006
			7.0	0.060	0.024	0.008	0.006
13.00	5.56	3.50	1.0	0.014	0.008	0.008	0.006
			3.0	0.038	0.014	0.014	0.006
			5.0	0.064	0.026	0.014	0.008
			7.0	0.062	0.040	0.016	0.008

Table 4.3 reports the simulation results of the UMP test for UMP_{W} when the two nuisance parameters are substituted by some arbitrary values. Each entry in the table represents the empirical level for the UMP test on mean. Since the results are quite similar for various sample sizes, we only present the results for $n = 50$. We observed that UMP test has the same performance as reported in Table 4.1. This indicates that the values of the two nuisance parameters do not have significant influence on the UMP test. To examine the robustness of the test due to model misspecification, random samples from the Neyman type-A distribution with orthogonal parameters μ and ϕ, $\mathrm{NTA}(\mu,\phi)$ are generated and the mean value is stated in H_0. The simulation results are shown in Table 4.4. For illustrative purpose, we have fixed the null hypothesis, $H_0 : \mu = 7.45$, in our study. It is seen that the UMP test for mean is robust to model misspecification if we perform the test, UMP_{W}. UMP_{IM} does not seem to perform well under model misspecification for small sample sizes.

Table 4.4: Empirical level for the small sample UMP test for mean when $\alpha = 0.05$.

Model	μ	ϕ	n	UMP_{W}	UMP_{IM}
NTA	7.45	1.30	50	0.019	0.188
			100	0.012	0.118
			200	0.008	0.032
			500	0.002	0.005
NTA	7.45	3.15	50	0.020	0.163
			100	0.028	0.073
			200	0.016	0.026
			500	0.008	0.010
NTA	7.45	5.56	50	0.023	0.151
			100	0.012	0.060
			200	0.008	0.010
			500	0.002	0.003

4.5 Conclusion

Based on the simulation study, we observed that the UMP test for the mean under model misspecification, UMP_{w}, performs well even when the sample sizes are small. Nevertheless, the power of the test increases with an increase in sample size. The test tends to be more conservative when the sample size becomes larger. Moreover, the simulation results have shown that the power of the test increases once the random sample is generated with the parameter μ larger than the value stated in H_0. The implementation of orthogonal parameterization in UMP test for mean has reduced or eliminated the effects of the nuisance parameters in constructing the UMP test. The robustness of the test due to the model misspecification has been justified in the Monte Carlo study as well.

CHAPTER 5

THE DELAPORTE DISTRIBUTION

5.1 Introduction

In this chapter, we consider applications of the results obtained in the previous chapters to a discrete distribution known as the Delaporte distribution arising in actuarial studies. Delaporte (1959, 1960, 1972a, b) and Ruohenen (1988) have considered a mixed Poisson distribution with a Gamma mixing distribution. The density function of the mixing distribution is given as

$$f(\gamma) = \frac{\beta^{\upsilon}}{\Gamma(\upsilon)} (\gamma - \lambda)^{\upsilon - 1} e^{-\beta\gamma},$$

where $\upsilon, \beta > 0; \gamma > \lambda \geq 0$. This leads to a three-parameter mixed Poisson distribution with probability generating function (pgf), see Johnson, Kemp and Kotz (2005), p.242,

$$\begin{aligned} G(z) &= \int_{\lambda}^{\infty} e^{\gamma(z-1)} f(\gamma) d\gamma \\ &= e^{\lambda(z-1)} \left(\tfrac{\beta+1}{\beta} - \tfrac{z}{\beta} \right)^{-\upsilon}. \end{aligned}$$

A further parameterization that has received much attention is $\rho = 1/(1+\beta)$, giving

$$G(z) = e^{\lambda(z-1)} \left(\frac{1-\rho}{1-\rho z} \right)^{\upsilon}. \tag{5.1}$$

The distribution with this pgf is known as the Delaporte distribution. It has been proposed as an alternative model in the insurance claim problem to the usual negative binomial distribution, another two-parameter gamma mixture. Willmot (1989) has studied on the tail behavior of this distribution and some of the asymptotic results are given in Willmot and Sundt (1989). Willmot and Sundt (1989) have developed the recursive algorithm for the purpose of evaluating Delaporte distribution. The Delaporte distribution has probability mass function

$$P(x;\lambda,\rho,\upsilon)=\sum_{r=0}^{x}\frac{\Gamma(r+\upsilon)}{\Gamma(\upsilon)r!}(1-\rho)^{\upsilon}\rho^{r}\frac{\lambda^{x-r}e^{-\lambda}}{(x-r)!}, \qquad x=0,1,2,... \tag{5.2}$$

The Delaporte distribution can be viewed as the convolution of the Poisson and negative binomial distributions. Ruohonen (1988) has studied the data fitting of the Delaporte distribution by the method of moments, moments and zero frequency, and maximum likelihood.

Parameters of the Delaporte distribution orthogonal to the mean have been derived in section 4.2.2. Orthogonal parameters have desirable properties as mentioned in Chapter 4; one of the important consequences is that they are not correlated and this would speed up convergence in an iterative method of estimation (Willmot, 1988, 1990). The efficiency of estimation is considered in section 5.2 since it does not seem to have been reported in the statistical literature. The comparative study of interval estimation for correctly specified and misspecified models is also discussed and illustrations using the Delaporte distribution are given in section 5.3. Section 5.4 gives some computational results of parameters estimation for the Delaporte distribution by using a proposed quadratic distance statistic.

5.2 Efficiency of Estimation

5.2.1 Introduction

Ruohonen (1988) has considered the method of moments and the maximum likelihood estimation in fitting the Delaporte distribution to real data. However, the maximum likelihood estimators for the parameters are calculated by optimizing the likelihood function numerically since they do not exist in a closed form. The objective of this section is to study the efficiencies of the methods discussed in Ruohonen's paper.

5.2.2 Evaluation of the Information Determinant

The efficiency of estimators can be computed by the formula (see, for instance, Katti and Gurland, 1962)

$$E = \frac{\text{var(ML estimate)}}{\text{var(other estimate)}}$$

For multi-parameter estimation, the formula above is modified and given by,

$$E = 1\Big/(\text{Generalized Variance} \times \text{Information Determinant}) \tag{5.3}$$

The information determinant is given by

$$I = \begin{vmatrix} nE\left[\left(\frac{\partial \ln P(x)}{\partial \theta_1}\right)^2\right] & nE\left[\left(\frac{\partial \ln P(x)}{\partial \theta_1}\right)\left(\frac{\partial \ln P(x)}{\partial \theta_2}\right)\right] & nE\left[\left(\frac{\partial \ln P(x)}{\partial \theta_1}\right)\left(\frac{\partial \ln P(x)}{\partial \theta_3}\right)\right] \\ nE\left[\left(\frac{\partial \ln P(x)}{\partial \theta_2}\right)\left(\frac{\partial \ln P(x)}{\partial \theta_1}\right)\right] & nE\left[\left(\frac{\partial \ln P(x)}{\partial \theta_2}\right)^2\right] & nE\left[\left(\frac{\partial \ln P(x)}{\partial \theta_2}\right)\left(\frac{\partial \ln P(x)}{\partial \theta_3}\right)\right] \\ nE\left[\left(\frac{\partial \ln P(x)}{\partial \theta_3}\right)\left(\frac{\partial \ln P(x)}{\partial \theta_1}\right)\right] & nE\left[\left(\frac{\partial \ln P(x)}{\partial \theta_3}\right)\left(\frac{\partial \ln P(x)}{\partial \theta_2}\right)\right] & nE\left[\left(\frac{\partial \ln P(x)}{\partial \theta_3}\right)^2\right] \end{vmatrix} \tag{5.4}$$

where $\theta_i, i = 1,2,3$, are the parameters of the distribution (see Shenton, 1949).

Let $P_k = P\left(X = k\right) = P\left(x; \lambda, \rho, \upsilon\right)$. The derivatives of $\ln P_k$ for the Delaporte Distribution are needed in order to evaluate I and they are given by

$$\frac{\partial \ln P_k}{\partial \lambda} = \frac{P_{k-1}}{P_k} - 1, \quad P_{k-i} = 0 \text{ for } k < i, \qquad \frac{\partial \ln P_0}{\partial \lambda} = -1$$

$$\frac{\partial \ln P_k}{\partial \rho} = \frac{1}{\rho(1-\rho)}\left\{k - t_k - \lambda\left(\frac{\partial \ln P_k}{\partial \lambda}\right)\right\}, \quad t_k = (k+1)\frac{P_{k+1}}{P_k}$$

$$\frac{\partial \ln P_k}{\partial \nu} = \ln(1-\rho) + \frac{1}{P_k}\sum_{r=0}^{k-1}\frac{\rho^{k-r}}{k-r}P_r, \quad \frac{\partial \ln P_0}{\partial \nu} = \ln(1-\rho)$$

We evaluate I by summing each entry of (5.4), which is an infinite series with a relative error of 10^{-20}.

5.2.3 Efficiencies of the Methods of Estimation

Let G be the generalized variance of $\tilde{\lambda}, \tilde{\rho}$ and $\tilde{v}$. Then G is defined in terms of the variance-covariance matrix and Jacobian matrix,

$$\mathrm{G} = \left|V(\bar{X}, S^2, m_3)\right| / \left|J\right|^2$$

The variances and covariances of sample moments and proportions (Griffiths, 1977) are

$$\mathrm{var}(\bar{X}) = \frac{\kappa_2}{n}, \quad \mathrm{var}(S^2) = \frac{\kappa_4 + 2\kappa_2^{\ 2}}{n},$$

$$\mathrm{var}(m_3) = \frac{\kappa_6 + 9\kappa_2\kappa_4 + 9\kappa_3^{\ 2} + 6\kappa_2^{\ 3}}{n}, \quad \mathrm{var}(f_0) = \frac{P_0(1-P_0)}{n};$$

$$\mathrm{cov}(\bar{X}, S^2) = \frac{\kappa_3}{n}, \quad \mathrm{cov}(\bar{X}, m_3) = \frac{\kappa_4}{n},$$

$$\mathrm{cov}(S^2, m_3) = \frac{\kappa_5 + 6\kappa_2\kappa_3}{n}, \quad \mathrm{cov}(\bar{X}, f_0) = \frac{-P_0\kappa_1}{n},$$

where $\bar{X}, S^2, m_3, f_0$ and κ_j are the sample mean, variance, third moment, proportion of zeros, and j^{th} population cumulant respectively.

The first cumulant (mean) for the Delaporte distribution is given by

$$\kappa_1 = \mu = \lambda + \frac{v\rho}{1-\rho},$$

The cumulants for the Delaporte distribution are easily obtained as the sum of the cumulants for the Poisson and NB distributions. The cumulants for the Poisson distribution are $\kappa_{\mathrm{r}} = \lambda$ for all $r \geq 1$. Since NB distribution is a power series distribution with series parameter ρ, the cumulants satisfy (see Johnson et al., 2005, p. 216),

$$\kappa_{r+1} = \rho \frac{\partial \kappa_r}{\partial \rho}, \quad r = 1, 2, \ldots \qquad \kappa_1 = \kappa_{[1]} = \frac{\nu\rho}{1-\rho}$$

Thus the cumulants for Delaporte distribution are given by

$$\kappa_2 = \lambda + \frac{\nu\rho}{(1-\rho)^2}, \quad \kappa_3 = \lambda + \frac{\nu\rho(1+\rho)}{(1-\rho)^3},$$

$$\kappa_4 = \lambda + \frac{\nu\rho(1+4\rho+\rho^2)}{(1-\rho)^4},$$

$$\kappa_5 = \lambda + \frac{\nu\rho(1+11\rho+11\rho^2+\rho^3)}{(1-\rho)^5} = \lambda + \frac{\nu\rho(1+\rho)(1+10\rho+\rho^2)}{(1-\rho)^5},$$

$$\kappa_6 = \lambda + \frac{\nu\rho(1+26\rho+66\rho^2+26\rho^3+\rho^4)}{(1-\rho)^6}$$

By direct computation, we have

$$\left|V(\overline{X}, S^2, m_3)\right| = \begin{vmatrix} \operatorname{var}(\overline{X}) & \operatorname{cov}(\overline{X}, S^2) & \operatorname{cov}(\overline{X}, m_3) \\ \operatorname{cov}(\overline{X}, S^2) & \operatorname{var}(S^2) & \operatorname{cov}(S^2, m_3) \\ \operatorname{cov}(\overline{X}, m_3) & \operatorname{cov}(S^2, m_3) & \operatorname{var}(m_3) \end{vmatrix}$$

$$= \begin{vmatrix} \frac{\kappa_2}{n} & \frac{\kappa_3}{n} & \frac{\kappa_4}{n} \\ \frac{\kappa_3}{n} & \frac{\kappa_4 + 2\kappa_2{}^2}{n} & \frac{\kappa_5 + 6\kappa_2\kappa_3}{n} \\ \frac{\kappa_4}{n} & \frac{\kappa_5 + 6\kappa_2\kappa_3}{n} & \frac{\kappa_6 + 9\kappa_2\kappa_4 + 9\kappa_3{}^2 + 6\kappa_2{}^3}{n} \end{vmatrix}$$

$$= \frac{1}{n^3}(12\kappa_2{}^6 - 9\kappa_3{}^4 + 24\kappa_2{}^4\kappa_4 - \kappa_4{}^3 + 2\kappa_3\kappa_4\kappa_5 + \kappa_2{}^2(7\kappa_4{}^2 - 12\kappa_3\kappa_5) - \kappa_3{}^2\kappa_6$$
$$+ \kappa_2{}^3(-24\kappa_3{}^2 + 2\kappa_6) + \kappa_2(12\kappa_3{}^2\kappa_4 - \kappa_5{}^2 + \kappa_4\kappa_6)),$$

$$|J| = \left|\frac{\partial(\mu_1', \sigma^2, \mu_3')}{\partial(\lambda, \rho, \nu)}\right| = \begin{vmatrix} \frac{\partial\kappa_1}{\partial\lambda} & \frac{\partial\kappa_1}{\partial\rho} & \frac{\partial\kappa_1}{\partial\nu} \\ \frac{\partial\kappa_2}{\partial\lambda} & \frac{\partial\kappa_2}{\partial\rho} & \frac{\partial\kappa_2}{\partial\nu} \\ \frac{\partial\kappa_3}{\partial\lambda} & \frac{\partial\kappa_3}{\partial\rho} & \frac{\partial\kappa_3}{\partial\nu} \end{vmatrix} = -\frac{2\nu\rho^4}{(1-\rho)^6}$$

where

$$\frac{\partial\kappa_1}{\partial\lambda}=\frac{\partial\kappa_2}{\partial\lambda}=\frac{\partial\kappa_3}{\partial\lambda}=1,$$

$$\frac{\partial\kappa_1}{\partial\rho}=\frac{\nu}{(1-\rho)^2},\qquad \frac{\partial\kappa_1}{\partial\nu}=\frac{\rho}{1-\rho},$$

$$\frac{\partial\kappa_2}{\partial\rho}=\frac{\nu\left(1-\rho^2\right)}{(1-\rho)^4}=\frac{\nu\left(1+\rho\right)}{(1-\rho)^3},\qquad \frac{\partial\kappa_2}{\partial\nu}=\frac{\rho}{(1-\rho)^2},$$

$$\frac{\partial\kappa_3}{\partial\rho}=\frac{\nu\left(1+4\rho+\rho^2\right)}{(1-\rho)^4},\qquad \frac{\partial\kappa_3}{\partial\nu}=\frac{\rho\left(1+\rho\right)}{(1-\rho)^3}.$$

Therefore,

$$\begin{aligned}\mathrm{G}\left(\tilde{\lambda},\tilde{\rho},\tilde{\nu}\right)=\frac{(1-\rho)^{12}}{4\nu^2\rho^8n^3}&(12\kappa_2{}^6\text{-}\,9\kappa_3{}^4+24\kappa_2{}^4\kappa_4\text{-}\kappa_4{}^3+2\kappa_3\kappa_4\kappa_5+\kappa_2{}^2(7\kappa_4{}^2\text{-}12\kappa_3\kappa_5)\text{-}\kappa_3{}^2\kappa_6\\ &+\kappa_2{}^3(\text{-}24\kappa_3{}^2+2\kappa_6)+\kappa_2(12\kappa_3{}^2\kappa_4\text{-}\kappa_5{}^2+\kappa_4\kappa_6)).\end{aligned}$$

The efficiency of estimation for method of moments is computed using (5.3).

Similarly, the variance-covariance matrix of moments and proportion is given by

$$\begin{aligned}\left|V(\bar{X},S^2,f_0)\right|&=\begin{vmatrix}\mathrm{var}(\bar{X}) & \mathrm{cov}(\bar{X},S^2) & \mathrm{cov}(\bar{X},f_0)\\ \mathrm{cov}(\bar{X},S^2) & \mathrm{var}(S^2) & \mathrm{cov}(S^2,f_0)\\ \mathrm{cov}(\bar{X},f_0) & \mathrm{cov}(S^2,f_0) & \mathrm{var}(f_0)\end{vmatrix}\\ &=\frac{1}{n^3}(P_0(2\kappa_3P_0\kappa_1{}^3\text{-}4\kappa_2{}^2P_0\kappa_1{}^2\text{-}\kappa_4P_0\kappa_1{}^2+\kappa_3{}^2(P_0\text{-}1)\\ &\quad+\kappa_2{}^3(2\text{-}3P_0)+\kappa_2(\kappa_4\text{-}(\kappa_1{}^4\text{-}2\kappa_3\kappa_1+\kappa_4)P_0)))\end{aligned}$$

where

$$\mathrm{var}(f_0)=\frac{P_0(1-P_0)}{n},$$
$$\mathrm{cov}(\bar{X},f_0)=-\frac{P_0\kappa_1}{n},$$
$$\mathrm{cov}(S^2,f_0)=-\frac{P_0\mu_2{}'}{n}=-\frac{P_0(\kappa_2+\kappa_1{}^2)}{n}$$

so that

$$\mathrm{G}\left(\tilde{\lambda},\tilde{\rho},\tilde{v}\right)=\frac{(1-\rho)^{12}}{4v^{2}\rho^{8}n^{3}}(P_{0}(2\kappa_{3}P_{0}\kappa_{1}^{3}\text{-}4\kappa_{2}^{2}P_{0}\kappa_{1}^{2}\text{-}\kappa_{4}P_{0}\kappa_{1}^{2}+\kappa_{3}^{2}(P_{0}\text{-}1)$$
$$+\kappa_{2}^{3}(2\text{-}3P_{0})+\kappa_{2}(\kappa_{4}\text{-}(\kappa_{1}^{4}\text{-}2\kappa_{3}\kappa_{1}+\kappa_{4})P_{0}))).$$

The efficiencies of the two methods of estimations are given in Tables 5.1 and 5.2 for the case where all three parameters are unknown.

Table 5.1: Efficiency of Moments estimation.

λ	ρ	υ					
		0.2	0.5	1.0	2.0	5.0	10.0
0.2	0.1	0.9203	0.9860	0.9958	0.9834	0.9653	0.9647
	0.3	0.7887	0.7236	0.6924	0.2334	0.7373	0.8270
	0.5	0.3803	0.3431	0.1654	0.2838	0.4509	0.5594
	0.7	0.1087	0.0720	0.1301	0.1577	0.3043	0.4414
	0.9	0.0067	0.0107	0.0187	0.0472	0.0798	0.0015
0.5	0.1	0.8665	0.9248	0.9682	0.9946	0.9963	0.9877
	0.3	0.9146	0.9262	0.3990	0.8425	0.8276	0.8591
	0.5	0.5469	0.3600	0.5092	0.4154	0.3126	0.3097
	0.7	0.1933	0.2157	0.1475	0.1868	0.2994	0.4402
	0.9	0.0179	0.0164	0.0247	0.0535	0.0792	0.0014
1.0	0.1	0.9082	0.9283	0.9513	0.9768	0.9981	0.9988
	0.3	0.8030	0.9399	0.9744	0.7330	0.2910	0.6087
	0.5	0.7926	0.7555	0.3796	0.4863	0.6399	0.7369
	0.7	0.3407	0.2411	0.2213	0.2481	0.3517	0.4367
	0.9	0.0312	0.0280	0.0346	0.0622	0.0783	0.0014
2.0	0.1	0.9488	0.9540	0.9615	0.9727	0.9902	0.9986
	0.3	0.6798	0.8378	0.9384	0.4777	0.5979	0.9368
	0.5	0.8103	0.8534	0.6347	0.4694	0.6517	0.7559
	0.7	0.4949	0.4642	0.3191	0.3016	0.3905	0.4557
	0.9	0.0617	0.0488	0.0522	0.0778	0.0761	0.0013
5.0	0.1	0.9784	0.9793	0.9806	0.9829	0.9884	0.9939
	0.3	0.7447	0.8035	0.8670	0.2395	0.9955	0.9961
	0.5	0.6099	0.6914	0.8870	0.6354	0.5252	0.6856
	0.7	0.6609	0.7404	0.6581	0.5337	0.4608	0.5050
	0.9	0.1426	0.1096	0.0993	0.1166	0.0679	0.0012

From Table 5.1, we observe that efficiency of Moments estimation is generally low for small ρ (<0.5) but improves as λ increases compared to method of Maximum likelihood estimation. However, Moments estimation is not efficient when the value of ρ approaches

1 even with the greater value of λ. Next, we have presented the results of efficiency of Moments and Zero Frequency estimation in Table 5.2. Notice that the efficiency of Moments and Zero Frequency estimation is high for most of the combination values of the parameters. We see that the efficiency increases significantly as λ increases.

Table 5.2: Efficiency of Moments and Zero Frequency estimation.

λ	ρ	υ				
		0.2	0.5	1.0	2.0	5.0
0.1	0.1	0.0997	0.1132	0.1435	0.2083	0.3969
	0.3	0.2038	0.2426	0.3054	0.4372	0.1967
	0.5	0.3572	0.3785	0.3286	0.5819	0.5705
	0.7	0.4191	0.3643	0.6569	0.6510	0.2470
	0.9	0.2852	0.4342	0.6130	0.5827	0.0088
0.2	0.1	0.2153	0.2027	0.2231	0.2826	0.4481
	0.3	0.2040	0.2590	0.3314	0.1546	0.6145
	0.5	0.3561	0.4033	0.2310	0.4508	0.4608
	0.7	0.4536	0.3369	0.6219	0.6144	0.2214
	0.9	0.3214	0.5012	0.6349	0.5512	0.0079
0.5	0.1	0.4726	0.4502	0.4428	0.4610	0.5335
	0.3	0.2311	0.2967	0.1685	0.4836	0.5700
	0.5	0.2735	0.2690	0.5071	0.5004	0.2475
	0.7	0.4129	0.6267	0.4895	0.5377	0.1662
	0.9	0.4699	0.5045	0.5981	0.4614	0.0058
1.0	0.1	0.4970	0.4995	0.5033	0.5103	0.5207
	0.3	0.2185	0.2917	0.3618	0.3372	0.1464
	0.5	0.2438	0.3489	0.2419	0.3841	0.3334
	0.7	0.3352	0.3729	0.4284	0.4398	0.1220
	0.9	0.3880	0.4645	0.4922	0.3259	0.0035
2.0	0.1	0.3844	0.3851	0.3857	0.3849	0.3725
	0.3	0.1388	0.1879	0.2344	0.1336	0.1606
	0.5	0.1269	0.1932	0.1913	0.1696	0.1487
	0.7	0.1662	0.2615	0.2373	0.2111	0.0530
	0.9	0.2367	0.2740	0.2676	0.1510	0.0013
5.0	0.1	0.1133	0.1120	0.1099	0.1057	0.0933
	0.3	0.0405	0.0440	0.0474	0.0126	0.0388
	0.5	0.0151	0.0213	0.0323	0.0240	0.0101
	0.7	0.0178	0.0315	0.0351	0.0251	0.0037
	0.9	0.0278	0.0320	0.0266	0.0117	0.0001

5.3 Interval Estimation

In most of the standard classical estimation methods, we always assume that the probability model is "correctly specified". However, if the model is misspecified, the standard tests are invalid. Two issues that are to be considered in the presence of misspecification will be the consistency and the asymptotic normality properties of the estimators. White (1982) had treated the asymptotic normality question and some assumptions are given in order to obtain consistency. The asymptotic covariance matrix of his proposed method is no longer equal to the inverse of the information matrix in general but it can be estimated consistently. In the absence of misspecification, the suggested asymptotic covariance matrix will simplify to the usual form. The proposed estimator is expected to be more "robust" even if the model is not correctly specified.

This section is organized as follows. Section 5.3.1 discusses about the confidence interval under the assumption that the probability model is a true model. White's results are exploited to construct the confidence interval under model misspecification in section 5.3.2. For illustrative purpose, we compute the confidence interval based on the two approaches for Delaporte distribution. Some random samples are generated from NB model to compare the robustness of the two approaches in the presence of misspecification. The results of the comparison are reported in section 5.3.3.

5.3.1 Confidence Interval for Correctly Specified Model

The confidence interval is obtained based upon the following asymptotic normality property.

Theorem 5.3.1.1. (Hogg, Craig and McKean, 2005) *Assume $X_1,...,X_n$ are iid with pdf $f\left(x;\theta_0\right)$ for $\theta_0 \in \Omega$ such that the regularity conditions are satisfied. Suppose further that Fisher information satisfies $0 < I\left(\theta_0\right) < \infty$. Then any consistent sequence of solutions of the mle equations satisfies*

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right)\xrightarrow{D} N\left(0,\frac{1}{I(\theta_0)}\right)$$

Based on Theorem 5.3.1.1, an approximate $(1-\alpha)100\%$ confidence interval for θ,

$$\left(\hat{\theta}_n - z_{\alpha/2}\frac{1}{\sqrt{nI\left(\hat{\theta}_n\right)}},\hat{\theta}_n + z_{\alpha/2}\frac{1}{\sqrt{nI\left(\hat{\theta}_n\right)}}\right)$$

where $\hat{\theta}_n$ is the ML estimator and $I\left(\hat{\theta}_n\right)$ for count data is given by

$$I\left(\hat{\theta}_n\right)=E\left[\left(\frac{\partial \ln P(x)}{\partial \hat{\theta}_n}\right)^2\right]=\sum_{x=0}^{\infty}\left(\frac{\partial \ln P(x)}{\partial \hat{\theta}_n}\right)^2 P(x).$$

5.3.2 Confidence Interval for Misspecified Model

In White's approach, the Hessian and outer product forms for the information matrix as given in equations (3.3) and (3.4) have been used. The asymptotic normality property is given by the following theorem,

Theorem 5.3.2.1. (White, 1982, p. 6) *Given Assumptions A1-A6* (see White, 1982)

$$\sqrt{n}\left(\hat{\theta}-\theta_*\right)\xrightarrow{D} N\left(0,C(\theta_*)\right).$$

Moreover, $C_n\left(\hat{\theta}_n\right)\xrightarrow{a.s.} C\left(\hat{\theta}_*\right)$ *element wise.*

Thus, an approximate $(1-\alpha)100\%$ confidence interval for θ is given by

$$\left(\hat{\theta}_n - z_{\alpha/2}C_n\left(\hat{\theta}_n\right),\hat{\theta}_n + z_{\alpha/2}C_n\left(\hat{\theta}_n\right)\right)$$

where $C_n\left(\hat{\theta}_n\right)=A_n\left(\hat{\theta}_n\right)^{-1}B_n\left(\hat{\theta}_n\right)A_n\left(\hat{\theta}_n\right)^{-1}$, $A_n\left(\hat{\theta}_n\right)$ and $B_n\left(\hat{\theta}_n\right)$ are defined as in equations (3.3) and (3.4) respectively.

5.3.3 Results and Discussions

To compare between the two approaches, we have generated five random samples and obtained the confidence intervals. All the random samples are generated by the inverse transform method and using the uniform random number generator of *Matlab*. The results are labeled as **IM** if the asymptotic variance is computed based on the inverse of Fisher Information Matrix and **WH** if the asymptotic variance is computed using White's approach given in section 5.3.2. The results are tabulated in Table 5.3. The first three random samples are generated from the Delaporte distribution with different values of parameters, λ, ρ and υ. In order to construct the confidence intervals for the three parameters, λ, ρ and υ are estimated using the ML method. From the simulation results, the WH confidence intervals have wider expected length. Next, we generate a random sample from the Delaporte distribution with outliers to examine the robustness of White's method. We found that the confidence interval is more robust to outliers. A random sample is generated from the negative binomial distribution, a special case of Delaporte distribution when $\lambda = 0$. Note that the WH confidence interval for λ contains the value zero while the IM interval does not. This implies that the WH confidence interval is able to indicate particular case of the more general model.

Table 5.3: 95% confidence interval for parameters of Delaporte distribution based on **IM** and **WH** approaches.
(The maximum likelihood estimators, $\hat{\lambda}, \hat{\rho}$ and $\hat{\upsilon}$ in bracket)

Model	*n*	λ	ρ	υ	95% Confidence Interval for	
					IM	**WH**
Delaporte	500	4.0 (4.3154)	0.6 (0.6345)	4.0 (3.4587)	$\rho = (0.6294, 0.6397)$ $\lambda = (4.0979, 4.4328)$ $\upsilon = (3.3182, 3.5991)$	$\rho = (0.4906, 0.7785)$ $\lambda = (0.5454, 7.4627)$ $\upsilon = (0.9668, 7.6639)$
Delaporte	1000	4.0 (4.0559)	0.6 (0.5967)	4.0 (4.0541)	$\rho = (0.5932, 0.6001)$ $\lambda = (3.9787, 4.1330)$ $\upsilon = (3.9472, 4.1610)$	$\rho = (0.5702, 0.6232)$ $\lambda = (3.1651, 4.9466)$ $\upsilon = (3.1651, 4.8454)$
Delaporte	1000	3.0 (3.1058)	0.4 (0.3952)	4.0 (3.9627)	$\rho = (0.3990, 0.4188)$ $\lambda = (3.1399, 3.3530)$ $\upsilon = (3.2195, 3.8609)$	$\rho = (0.3807, 0.4370)$ $\lambda = (2.5400, 3.9529)$ $\upsilon = (2.7318, 4.3485)$
Delaporte (Contaminated with outliers)	1000	4.0 (5.2577)	0.6 (0.6682)	4.0 (2.4056)	$\rho = (0.6657, 0.6707)$ $\lambda = (5.2097, 5.3057)$ $\upsilon = (2.3562, 2.4550)$	$\rho = (0.5622, 0.7741)$ $\lambda = (3.2062, 7.3093)$ $\upsilon = (0.2888, 4.5225)$
Negative Binomial	1000	(0.5673)	0.6 (0.6167)	4.0 (3.4043)	$\rho = (0.6147, 0.6191)$ $\lambda = (0.5274, 0.6073)$ $\upsilon = (3.3494, 3.4591)$	$\rho = (0.5499, 0.6838)$ $\lambda = (-0.5840, 1.7186)$ $\upsilon = (1.7728, 5.0357)$

5.4 Parameter Estimation via a Quadratic Distance Statistic

5.4.1 Introduction

White's IM test has been derived based on Bartlett's First Identity (BFI) which states that when the model is correctly specified the Hessian form of the information matrix $-A(\theta)$ and its outer product form $B(\theta)$ satisfy $A(\theta) + B(\theta) = 0$. The failure of this equality implies misspecification of the model.

In general, the Hessian matrix $A_n(\theta) = \left(A_{ij}(\theta,n)\right)$ and outer product matrix $B_n(\theta) = \left(B_{ij}(\theta,n)\right)$ for the information matrix for grouped frequency data are defined as (White, 1982)

$$A_{ij}(\theta,n) = n^{-1}\sum_{x=0}^{k} f_x \frac{\partial^2 \ell nP(x;\theta)}{\partial\theta_i \partial\theta_j},$$

$$B_{ij}(\theta,n) = n^{-1}\sum_{x=0}^{k} f_x \left[\frac{\partial \ell nP(x;\theta)}{\partial\theta_i}.\frac{\partial \ell nP(x;\theta)}{\partial\theta_j}\right].$$

Suppose that $i, j = 1,2,3, \ldots, p$. Consider the quadratic distance statistic

$$Q_n(\theta) = \sqrt{n}D_n(\theta)' V^{-1}\sqrt{n}D_n(\theta), \tag{5.5}$$

where $\sqrt{n}D_n(\theta) = \sqrt{n}\left(\sum_{i=1}^{p} D_{ii}(\theta,n)\right)$, $i = 1,2, \ldots, p$, $D_{ij}(\theta,n) = A_{ij}(\theta,n) + B_{ij}(\theta,n)$ and $V = I$ is the $p \times p$ identity matrix. In defining (5.5), we have only considered $D_n(\theta)$ as a $p \times 1$ vector of diagonal elements of $A_n(\theta) + B_n(\theta)$. Statistic (5.5) is identical in form to White's information matrix test statistic (White, 1982, Eq. (4.1)) but differs in two respects (1) V is not a covariance matrix but an identity matrix; (2) elements of $D_n(\theta)$ are selected from the diagonal of $A_n(\theta) + B_n(\theta)$, whereas in White's statistic they are selected from the $p(p+1)/2$ elements in the upper triangular part of $A_n(\theta) + B_n(\theta)$. Of course if model parameters are orthogonal, we have a diagonal matrix.

Equation (5.5) may be written as

$$Q_n(\theta) = n\left(D_n(\theta)\right)^2 \tag{5.6}$$

In fact, equation (5.6) is in the form of an estimating equation in M-estimation. Therefore, the basic theory and results in M-estimation applies. Since $D_n(\theta)$ is the trace of the diagonal matrix, we have to solve for the roots for each of the $p \times 1$ vector of diagonal elements of $A_n(\theta) + B_n(\theta)$ in order to satisfy (5.6). The roots found will be the estimated parameters.

Gan and Jiang (1999), based on the assumption that the model is correctly specified, gave a criterion to determine whether the global maximum has been achieved in maximum likelihood estimation of the model. This criterion essentially says that Bartlett's First Identity, BFI, (information matrix equality) holds asymptotically at the global maximum. Therefore, the global maximum is expected to be attained with the application of the proposed quadratic distance statistic.

For illustration, the estimation procedure is applied to the Delaporte distribution with orthogonal parameters (see Eq. 4.11). Let $P_k = P(X = k) = P(x; \mu, \phi_2, \xi)$. The first and second derivatives of the probabilities with respect to the three orthogonal parameters of Delaporte distribution are obtained as follows with the help of *Mathematica*:

$$\frac{\partial \ln P_k}{\partial \mu} = \begin{cases} \dfrac{\mu - \phi_2 - \xi - \Phi_1}{2\Phi_1}, & k = 0; \\ \dfrac{(\mu - k)(\phi_2 - \mu - \xi + \Phi_1)}{\Phi_1(\mu + \phi_2 + \xi - \Phi_1)}, & k = 1, 2, 3, ... \end{cases}$$

$$\frac{\partial^2 \ln P_k}{\partial \mu^2} = \begin{cases} \dfrac{\xi(\mu + \phi_2 + \xi)}{\Phi_1^{\,3}}, & k = 0; \\ \dfrac{\left\{\begin{array}{l} 2\Big[\xi(\mu + \phi_2 + \xi)\big(\mu^2 + \phi_2^{\,2} + \xi^2 + 2\xi(\mu + \phi_2) - \xi\Phi_3 - \mu\Phi_3 - \phi\Phi_3\big) \\ -k\Big\langle -\xi^3 + \xi^2(-3\mu + \phi_2 + \Phi_3) + (\mu - \phi_2)^2(-\mu + \phi_2 + \Phi_3) \\ +\xi\big(-3\mu^2 + 4\mu\phi_2 + 3\phi_2^{\,2} + 2\mu\Phi_3 - 2\phi_2\Phi_3\big)\Big\rangle\Big] \end{array}\right\}}{\Phi_1^{\,3}(\mu + \phi_2 + \xi - \Phi_1)^2}, & k = 1, 2, 3, ... \end{cases}$$

$$\frac{\partial \ln P_k}{\partial \phi_2}=\begin{cases} \dfrac{\xi^2+\xi\mu+2\xi\phi_2-\mu\phi_2+{\phi_2}^2-\xi\Phi_1-\phi_2\Phi_1}{2\phi_2\Phi_1}, & k=0; \\ \dfrac{1}{2}\left(1-\dfrac{\phi_2+\xi-\mu}{\Phi_1}\right)\left(\dfrac{P_{k-1}}{P_k}-1\right)+\Phi_4.\phi_2.\dfrac{P_{k-1}}{P_k}-\dfrac{\Phi_4}{\Phi_2}k+\left\{\dfrac{1}{1-\Phi_2}\right\}\Phi_4.\xi, & k=1,2,3,.... \end{cases}$$

Let $\Pr=\dfrac{P_{k-1}}{P_k}\left(\dfrac{\partial \ln P_{k-1}}{\partial \phi_2}-\dfrac{\partial \ln P_k}{\partial \phi_2}\right),$

$$\frac{\partial^2 \ln P_k}{\partial {\phi_2}^2}=\begin{cases} -\dfrac{1}{2{\phi_2}^2\Phi_1^3}\Big\{\xi\Big[\xi^3+\mu^3-3\mu^2\phi_2-3\mu{\phi_2}^2+{\phi_2}^3+3\xi^2\left(\mu+\phi_2\right)+3\xi\left(\mu^2+{\phi_2}^2\right) & \\ -\xi^2\Phi_1-2\mu\xi\Phi_1-\mu^2\Phi_1-2\xi\phi_2\Phi_1+2\mu\phi_2\Phi_1-{\phi_2}^2\Phi_1\Big], & k=0; \\ & \\ -\dfrac{2\mu\xi}{\Phi_1^3}\left(\dfrac{P_{k-1}}{P_k}-1\right)+\dfrac{1}{2}\left(1-\dfrac{\phi_2+\xi-\mu}{\Phi_1}\right)\Pr+\Big\{\dfrac{1}{2{\phi_2}^2\Phi_1^3}\Big[\xi^4+\xi^3\left(4\mu+3\phi_2-\Phi_3\right) & \\ -\mu\left(\mu-\phi_2\right)^2\left(-\mu+\phi_2+\Phi_3\right)+\xi^2\left(6\mu^2+3\mu\phi_2+3{\phi_2}^2-3\mu\Phi_5-2\phi_2\Phi_5\right) & \\ +\xi\left(4\mu^3+2\mu{\phi_2}^2+{\phi_2}^2\left(\phi_2-\Phi_3\right)-3\mu^2\left(\phi_2+\Phi_3\right)\right)\Big]\Big\}\left(\dfrac{P_{k-1}}{P_k}\right) & \\ +\phi_2\Phi_4\Pr-k\Big\{-\dfrac{1}{4{\phi_2}^4\Phi_2^2}\Big[-\dfrac{8\mu{\phi_2}^3\xi\Phi_2}{\Phi_1^3}+\phi_2\left(1-\dfrac{\phi_2+\xi-\mu}{\Phi_1}\right) & \\ \left(2\phi_2\Phi_2-\phi_2\left(1-\dfrac{\phi_2+\xi-\mu}{\Phi_1}\right)\right)+2\phi_2\Phi_2\left(2\phi_2\Phi_2-\phi_2\left(1-\dfrac{\phi_2+\xi-\mu}{\Phi_1}\right)\right)\Big]\Big\} & \\ +\xi\Big\{\dfrac{1}{2{\phi_2}^2\Phi_1^3}\Big[-\xi^3-\mu^3+3\mu^2\phi_2-\mu{\phi_2}^2-{\phi_2}^3-3\xi^2\left(\mu+\phi_2\right)-3\xi\left(\mu^2+{\phi_2}^2\right) & \\ +\xi^2\Phi_1+2\xi\mu\Phi_1+\mu^2\Phi_1+2\xi\phi_2\Phi_1-2\mu\phi_2\Phi_1+{\phi_2}^2\Phi_1\Big]\Big\}, & k=1,2,3.... \end{cases}$$

$$\frac{\partial \ln P_k}{\partial \xi}=\begin{cases} \dfrac{\mu}{\Phi_1}+\ln\left(\dfrac{\phi_2-\xi-\mu+\Phi_1}{2\phi_2}\right), & k=0; \\ & \\ -\dfrac{1}{2}\left(1-\dfrac{\mu+\phi_2+\xi}{\Phi_1}\right)+k\dfrac{\left(1-\dfrac{\mu+\phi_2+\xi}{\Phi_1}\right)}{2\phi_2\Phi_2}-\dfrac{\xi\left(1-\frac{\mu+\phi_2+\xi}{\Phi_1}\right)}{2\phi_2\left(1-\Phi_2\right)} & \\ +\ln\left(1-\Phi_2\right)+\dfrac{1}{P_k}\displaystyle\sum_{j=0}^{k-1}\dfrac{\left(\Phi_2\right)^{j+1}}{j+1}P_{k-j-1}, & k=1,2,3,.... \end{cases}$$

$$\frac{\partial^2 \ln P_k}{\partial \xi^2} = \begin{cases} \frac{1}{2\xi\Phi_1^3}\left\{-\xi^3 + \xi^2\left(-3\mu - 3\phi_2 + \Phi_3\right) + \left(\mu - \phi_2\right)^2\left(\mu - \phi_2 + \Phi_3\right) \right. \\ \left. + \xi\left[-\mu^2 + 2\mu\Phi_3 + \phi_2\left(-3\phi_2 + 2\Phi_3\right)\right]\right\}, & k = 0; \\ -\frac{1}{2}\left(\frac{4\mu\phi_2}{\Phi_1^3}\right) + k\left(\frac{\mu + \phi_2 + \xi}{\Phi_1^3}\right) - \left(\frac{\mu\left(\mu + \xi - \phi_2\right)}{\Phi_5^3}\right) \\ -\frac{\left(1 - \frac{\mu + \phi_2 + \xi}{\Phi_1}\right)}{2\phi_2\left(1 - \Phi_2\right)} - \frac{1}{P_k}\frac{\partial \ln P_k}{\partial \xi}\sum_{j=0}^{k-1}\frac{\Phi_2^{j+1}}{j+1}P_{k-j-1} \\ + \frac{1}{P_k}\sum_{k=0}^{j-1}\left[\frac{1}{2}\frac{\left(1 - \frac{\mu + \phi_2 + \xi}{\Phi_1}\right)\Phi_2^j}{\phi_2}P_{k-j-1} + \frac{\Phi_2^{j+1}}{j+1}P_{k-j-1}\frac{\partial \ln P_{k-j-1}}{\partial \xi}\right], & k = 1, 2, 3, ... \end{cases}$$

where

$$\Phi_1 = \sqrt{\left(\mu + \phi_2 + \xi\right)^2 - 4\mu\phi_2}, \Phi_2 = \frac{\left(\mu + \phi_2 + \xi\right) - \Phi_1}{2\phi_2}, \Phi_3 = \sqrt{\left(\mu + \xi\right)^2 + 2\left(\xi - \mu\right)\phi_2 + \phi_2{}^2},$$

$$\Phi_4 = \frac{\left(\mu + \phi_2 + \xi\right) - \sqrt{\left(\mu + \phi_2 + \xi\right)^2 - 4\mu\phi_2}}{2\phi_2{}^2} - \frac{1 - \frac{\phi_2 + \xi - \mu}{\sqrt{\left(\mu + \phi_2 + \xi\right)^2 - 4\mu\phi_2}}}{2\phi_2} \text{ and}$$

$$\Phi_5 = \sqrt{\beta^2 + \left(\mu - \phi_2\right)^2 + 2\beta\left(\mu + \phi_2\right)}.$$

5.4.2 Results and Discussions

For parameter estimation based on the proposed quadratic distance statistic (5.6), we consider two simulated random samples from the Delaporte distribution and the data on Malayan butterflies, which was fitted with the Poisson-lognormal distribution by Blumer (1974). The data has also been fitted to Poisson-negative binomial distribution by Gupta, Gupta and Ong (2004) using maximum likelihood estimation. Table 5.4 has reported some results regarding parameter estimation by using the proposed quadratic distance statistic (Q_n) and the results are compared to the method of maximum likelihood estimation

(MLE). We observe that MLE and Q_n give similar parameter estimates for the three data sets. However, the value of *BFI* for Q_n is very small and closer to zero compared to MLE.

Table 5.4: Parameter estimation for Delaporte distribution based on MLE and Q_n.

Model	MLE			Q_n			*BFI*	
	$\hat{\mu}$	$\hat{\phi}_2$	$\hat{\xi}$	$\tilde{\mu}$	$\tilde{\phi}_2$	$\tilde{\xi}$	MLE	Q_n
Delaporte ($n = 1000, \mu = 4.15,$ $\phi_2 = 2.69,\ \xi = 0.5$)	4.1060	2.4999	0.5220	3.8816	2.6412	0.5643	-0.0522	2.473E-4
Delaporte ($n = 1000, \mu = 13.0,$ $\phi_2 = 5.56,\ \xi = 3.5$)	12.8910	3.5807	4.7137	13.3680	5.3221	3.4628	6.975E-3	5.927E-4
Distribution Of Corbet's Butterflies with zeros. (Blumer,1974)	10.2159	0.1550	0.2623	11.1350	0.0923	0.2880	-0.1974	1.865E-7

Specifically, we have considered the data fitting on Malayan butterflies by Poisson-lognormal (Blumer, 1974), Poisson-negative binomial (Gupta, Gupta and Ong, 2004) and Delaporte models. The fits of the models is given by Table 5.5. It is known that the additive Poisson-negative binomial model has comparable better fit (see Gupta, Gupta and Ong, 2004, p. 563). We observe that the chi-square value for the fits by the Delaporte and Poisson-negative binomial distributions are almost the same.

5.5 Conclusion

In this Chapter, some statistical inferences for the Delaporte distribution have been studied. First, the efficiency of the methods of estimation has been reported. Besides, we have considered the confidence interval under model misspecification for Delaporte distribution; some examples are given as illustration purpose. We have proposed to estimate the parameters using the statistic Q_n based on Bartlett's First Identity which is identical in form to White's information matrix test statistic. The advantage of the proposed quadratic distance statistic is that the global maximum of the estimation of model parameters can be achieved. Furthermore, the statistic is easier to compute after orthogonalization of parameters. We have compared the two estimation methods, MLE and Q_n by using generated random samples and a real data set. We found that the two methods give similar performance but the *BFI*'s value for Q_n is much smaller. Lastly, we have considered the goodness of fit of the Delaporte model to the Malayan butterfly data.

Table 5.5: Data fitting for distribution of Corbet's Butterflies with zeros (Blumer, 1974).

	Observation	Poisson-lognomal (Blumer)	Additive Poisson-negative binomial (Gupta and Ong)	Delaporte
0	304	295.0	303.10	303.14
1	118	127.4	123.28	123.29
2	74	74.6	62.83	62.83
3	44	50.7	43.29	43.29
4	24	37.5	33.73	33.74
5	29	29.3	27.77	27.77
6	22	23.7	23.61	23.61
7	20	19.7	20.51	20.51
8	19	16.7	18.10	18.10
9	20	14.4	16.16	16.16
10	15	12.6	14.57	14.57
11	12	11.1	13.23	13.23
12	14	9.9	12.10	12.09
13	6	8.9	11.10	11.10
14	12	8.1	10.24	10.24
15	6	7.3	9.49	9.49
16	9	6.7	8.81	8.81
17	9	6.2	8.21	8.21
18	6	5.7	7.67	7.67
19	10	5.3	7.19	7.19
20	10	4.9	6.74	6.74
21	11	4.6	6.34	6.34
22	5	4.3	5.97	5.97
23	3	4.0	5.63	5.63
24	3	3.8	5.32	5.32
25+	119	131.3	118.99	118.95
Total	924	923.7	923.98	923.99
Chi.sq.		36.8	19.46	19.47
df		23	22	22

df – degree of freedom

CHAPTER 6

CONCLUSION AND FURTHER RESEARCH WORK

This thesis proposed a goodness-of-fit test based on Bartlett's First Identity. The application of Bartlett's First Identity for goodness-of-fit test does not seem to have been widely reported in the statistical literature. This identity is the basis of White's (1982) Information Matrix (IM) test for model misspecification. However, the proposed test differs from the IM test as follows. The goodness-of-fit test statistic has been considered under orthogonality of the parameters with the bootstrapped critical values adjusted for bias by using the bias-corrected accelerated (BC_a) method. When parameters are orthogonal, the proposed test statistic is simplified and this reduces computation. Besides, the direct application of Bartlett First Identity as a goodness-of-fit test avoids the evaluation of the complicated covariance matrix in the IM test. The consistency and asymptotic normality properties of the proposed goodness-of-fit test have been proved. The empirical distribution function tests have been considered for comparison purpose. For future work, we will consider the proposed goodness-of-fit test for the class of distributions with orthogonal parameters.

The consequences of orthogonal parameters in statistical inference have been examined by Cox and Reid (1987). One of the important roles of orthogonal parameterization is to speed up the convergence of the numerical method employed in maximum likelihood estimation. This is especially significant for models that have complicated probability functions with many parameters. Willmot (1988, 1990) have presented some results of orthogonal parameterization for a class of discrete models with two unknown parameters. We have extended the work of Willmot by examining the orthogonality of the mean μ for models with more than two parameters, where the two remaining parameters are regarded as nuisance parameters since they are orthogonal to μ. We considered orthogonality for a class of Poisson convolution models and since this convolution can be treated as a signal-plus-noise model that is of practical importance. The condition for orthogonal parameters has been derived. In particular orthogonal parameters have been derived for the Delaporte distribution, a model which is commonly used in actuarial studies. A uniformly most powerful test for mean has also been developed as an application of the orthogonality results. Future study may consider orthogonality between the two nuisance parameters as well. We will also consider other applications of this result in statistical inference.

Since the Delaporte distribution is of independent interest, some statistical inference and practical applications has been considered. The efficiency of the method of moments and moment and zero frequency relative to maximum likelihood estimation has been examined. Further work on other recent methods of parameter estimation and a comparative study will be of interest. A conservative interval estimation method under model misspecification for Delaporte model has also been presented. Besides, we have proposed a quadratic distance statistic for parameter estimation. The estimation results of the newly proposed statistic are compared to the maximum likelihood estimation. For further work a Monte Carlo simulation study will be conducted on the performance and property of the quadratic distance estimation.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the likelihood principle. *Proc. of the Second International Symposium of Information Theory*, ed. B.N. Petrov and F. Csaki. Budapest: Akademiai Kiado, 1973.

Baringhaus, L. and Henze, N. (1992). A goodness-of-fit test for the Poisson distribution based on the empirical generating function. *Statist. Probab. Lett.*, **13**, 269-274.

Barnett, V.D. (1966). Evaluation of the Maximum Likelihood estimator where the Likelihood Equation has Multiple Roots. *Biometrika*, **53**, 151-166.

Beran, R. (1988). Prepivoting test statistics: A bootstrap view of asymptotic refinements. *J. Amer. Statist. Assoc.*, **83**, 687-697.

Berk, R. (1967). Review 1922 of 'Invariance of Maximum Likelihood Estimators' by Peter W. Zehna'. *Mathematical Reviews*, **33**, 342-343.

Berk, R.H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Ann. of Math. Statist.*, **37**, 57-58.

Berk, R. H. (1970). Consistency a posteriori. *Ann. of Math. Statist.*, **41**, 894-906.

Bertsimas, D. and Tsitsiklis, J. (1993). Simulated annealing. *Statist. Sci.*, **8,** 10-15.

Blumer, M.G. (1974). On fitting the Poisson-lognormal distribution to species-abundance data. *Biometrics*, **30**, 101-110.

Bohachevsky, I.O., Johnson, M.E. and Stein, M.L. (1986). Generalized simulated annealing for function optimization. *Technometrics*, **28**, 209-217.

Bowman, K.O. and Shenton, L.R. (1975). Omnibus contours for departures from normality based on $\sqrt{b_1}$ and b_2. *Biometrika*, **62**, 243-250.

Brooks, S.P. and Morgan, J.T. (1995). Optimization using simulated annealing. *The Statistician*, **44**, 2, 241-257.

Casella, G. and Berger, R.L. (2002). *Statistical Inference*, 2nd ed., Duxbury, Belmont, CA.

Chesher, A. and Spady, R. (1991). Asymptotic expansions of the information matrix test statistics. *Econometrica*, **59**, 787-815.

Cox, D.R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Statist. Soc. (B)*, **49**, 1-39.

Croux, C., Dhaene, G. and Hoorelbeke, D. (2006). Testing the information matrix equality with robust estimators. *J. Statist.Plann. Inference*, **136**, 10, 3583-3613.

D' Agostino, R. and Stephens, M. (1986). *Goodness of Fit Techniques*. Marcel Dekker, New York.

Davidson, R. and MacKinnon, J.G. (1992). A new form of the information matrix test statistics. *Econometrica*, **60**, 145-157.

Delaporte, P. (1959). Quelques problèmes de statistique mathématique posés par l'assurance automobile et le bonus non sinistre. *Bulletin Trimestriel de l'Institut des Actuaires Francais*, **227**, 87-102.

Delaporte, P. (1960). Un problème de tarification de l'assurance accidents d'automobiles examine par la statistique ì mathématique. *Transactions of the XVIth International Congress of Actuaries* **II**, 121-135.

Delaporte, P. (1972a). Le mathématique de l'assurance automobile. *ASTIN Bulletin*, **6**, 185-190.

Delaporte, P. (1972b). Construction d'un tarif d'assurance automobile base sur le principe de la prime modelée sur le risque. *Mitteilungen der Vereinigung schweizerischer Versicherungs-mathematiker*, **72**, 101-113.

Dhaene, G. and Hoorelbeke, D. (2004). The information matrix test with bootstrap-based covariance matrix estimation. *Econ. Lett.*, **82**, 341-347.

Douglas, J.B. (1980). *Analysis with Standard Contagious Distributions*. Fairland, Maryland: International Co-operative Publishing House.

Efron, B. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC, p. 184-188.

Famoye, F. (2000). Goodness-of-fit tests for generalized logarithmic series distribution. *Comput. Statist. Data Anal.*, **33**, 59-67.

Fisher, R.A. (1925). Theory of Statistical Estimation. *Proc. Camb. Philos. Soc.*

Fomby, T.B. and Carter Hill, R. (2003). *Maximum likelihood estimation of Misspecified Models: Twenty years later.* Advances in Econometrics, Volume 17, Elsevier Ltd.

Fouskakis, D. and Drapper, D. (2002). Stochastic Optimization: a Review. *International Statistical Review*, *International Statistical Institute*, **70**, 3, 315-349.

Gan, L. and Jiang, J. (1999). A test for global maximum. *J. Amer. Statist. Assoc.*, **94**, 847-854.

Goffe, W. (1996). *SIMANN: A Global Optimization Algorithm using Simulated Annealing. Studies in Nonlinear Dynamics & Econometrics*, Berkeley Electronic Press, 1(3), 169-176.

Gourieroux, C.S., Monfort, A. and Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica*, 1984, **52**, 3, 681-700.

Griffiths, D. (1977). Avoidance-modified generalized distributions and their application to studies of superparasitism. *Biometrics*, **33**, 103-112.

Gupta, P.L., Gupta, R.C. and Ong, S.H. (2004). Modelling count data by random effect poisson model. *Sankhyā*, **66**, 3, 1-18.

Habibullah, M. and Katti, S.K. (1991). A modified steepest descent method with applications to maximizing likelihood functions. *Ann. Inst. Statist. Math.*, **43**, 2, 391-404.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals (with discussion). *Ann. Statist.*, **16**, 927-953.

Hogg, R., Craig, A. and McKean, J. (2005). *Introduction to Mathematical Statistics*. Pearson Prentice Hall, Upper Saddle River.

Horowitz, J.L. (1994). Bootstrap-based critical values for the information matrix test. *J. Econometrics*, **61**, 395-411.

Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions., *Proc. of the fifth Berkeley symposium in Mathematical Statistics and Probability*. Berkeley: University of California Press.

Huzurbazar, V. (1950). Probability distributions and orthogonal parameters. *Proc. Camb. Phil. Soc.*, **46**, 281-284.

Huzurbazar, V. (1956). Sufficient statistics and orthogonal parameters. *Sankhya*, **17**, 217-220.

Jarque, C.M. and Bera, A.K. (1987). A test for normality of observations and regression residuals. *International Statistical Review.*, **55**, 163-172.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed., Oxford: Clarendon Press.

Johnson, N.L., Kemp, A.W. and Kotz, S. (2005). *Univariate Discrete Distributions*. John Wiley & Sons, Inc., Hoboken: New Jersey.

Kalbfleisch, J. D. and Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters. *J.R.Statist. Soc. Ser. B*, **32**, 175-208.

Kale, B.K. (1961). On the solution of the likelihood equation by iteration processes. *Biometrika*, **48**, 452-456.

Kale, B.K. (1962). On the solution of the likelihood equation by iteration processes – the multiparametric case. *Biometrika*, **49**, 479-486.

Katti, S.K. and Gurland, J. (1962). Efficiency of certain methods of estimation for the negative binomial and the Neyman Type A distribution. *Biometrika*, **49**, 215-226.

Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. (1983). Optimization by simulated annealing. *Science*, **220**, 671-680.

Lehmann, E. L. (1959). *Testing Statistical Hypotheses*. A Wiley publication in mathematical statistics. New York: Wiley.

Lehmann, E.L. (1999). *Elements of Large Sample Theory*. Springer, New York.

LeCam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related bayes's estimates. *University of California Publications in Statistics*, **1**, 277-330.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M., Teller, A. H. and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.*, **21**, 1087-1092.

Orme, C. (1990). The small-sample performance of the information-matrix test. *J. Econometrics*, **46**, 309-331.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Phil. Mag.*, **50**, 157-172.

Ruohenen, M. (1988). A model for the claim number process. *ASTIN Bulletin*, **18**, 57-68.

Samaniego, F.J. (1976). A Characterization of Convoluted Poisson Distributions with Applications to Estimation. *J. Amer. Statist. Assoc.*, **71**, 475-479.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. of Statist*, **6**, 461-464.

Shenton, L.R. (1950). Maximum likelihood and the efficiency of the methods of moments. *Biometrika*, **37**, 111-116.

Shonick, W. (1970). A stochastic model for occupancy related random variables in general acute hospitals. *J. Amer. Statist. Assoc.*, **65**, 1474-1500.

Smyth, G. K. (2002). Optimization. *Encyclopedia of Environmetrics*, A. H. El-Shaarawi and W. W. Piegorsch (eds.), Wiley, Chichester, Vol. 3, 1481-1487.

Sprott, D. (1980). Maximum likelihood in small samples: Estimation in the presence of a nuisance parameter. *Biometrika*, **56**, 515-523.

Sprott, D. (1983). Estimating the parameters of a convolution by maximum likelihood. *J. Amer. Statist. Assoc.*, **78**, 457-460.

Stein, G., Zucchini, W. and Juritz, J. (1987). Parameter estimation for the Sichel distribution and its multivariate extension. *J. Amer. Statist. Assoc.,* **82**, 938-944.

Stephens, M.A. (1974). EDF Statistics for goodness of fit and some comparisons. *J. Amer. Statist. Assoc.*, **69**, 730-737.

Stuart, A. and Ord, K. (1991). *Kendall's Advanced Theory of Statistics*, Vol. II. Clarendon Press, New York.

Taylor, L.W. (1987). The size bias of White's information matrix test. *Economics Letters*, **24**, 63-67.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, **54**, 426-482.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1-25.

Willmot, G. (1987). The Poisson-Inverse Gaussian Distribution as an Alternative to the Negative Binomial. *Scand. Actuarial J.*, **41**, 113-127.

Willmot, G. (1988). Parameter orthogonality for a family of discrete distributions. *J. Amer. Statist. Assoc.*, **83**, 517-521.

Willmot, G. (1989) Limiting tail behaviour of some discrete compound distributions. *Insurance: Mathematics and Economics*, **8**, 175-185.

Willmot, G. (1990). On the construction of a parameter orthogonal to the mean. *Biometrika.*, **77**, 424-428.

Willmot, G. and Sundt, B. (1989) On evaluation of the Delaporte distribution and related distributions. *Scand. Actuar. J.*, 101-113.

Yanagimoto, T. (1991). Estimating a model through the conditional MLE. *Ann. Inst. Statist. Math.*, **43**, 4, 735-746.

Zehna, P.W. (1966). Invariance of maximum likelihood estimators. *Annals of Mathematical Statictics*, **37**, 744.

APPENDIX A

WHITE'S (1982) RESULTS

The following assumptions and theorems have been directly abstracted from White (1982) for easy reference.

Assumption A1: The independent random $1 \times M$ vectors $X_t, t = 1,...,n,$ have common joint distribution function G on Ω, a measurable Euclidean space, with measurable Radon-Nikodym density $g = dG/dv$.

Assumption A2: The family of distribution functions $F(x,\theta)$ has Radon-Nikodym densities $f(x,\theta) = dF(x,\theta)/dv$ which are measurable in *x* for every θ in Θ, a compact subset of a *p*-dimensional Euclidean space, and continuous in θ for every *u* in Ω.

Assumption A3: (a) $E(\ln g(X_t))$ exists and $|\ln f(x,\theta)| \leqq m(x)$ for all θ in Θ, where m is integrable with respect to *G*; (b) $I(g:f,\theta) \equiv E(\ln[g(X_t)/f(X_t,\theta)])$ has a uniques minimum at θ_* in Θ.

Assumption A4: $\partial \ln f(x,\theta)/\partial\theta_i, i = 1,..., p,$ are measurable functions of *u* for each θ in θ in Θ and continuously differentiable functions of θ for each *x* in Ω.

Assumption A5: $|\partial^2 \ln f(x,\theta)/\partial\theta_i\partial\theta_j|$ and $|\partial \ln f(x,\theta)/\partial\theta_i \partial \ln f(x,\theta)/\partial\theta_j|$, *i*, *j* = 1,…, *p*, are dominated by functions integrable with respect to *G* for all *x* in Ω and θ in Θ.

Assumption A6: (a) θ_* is interior to Θ; (b) $B(\theta_*)$ is nonsingular; (c) θ_* is a regular point of $A(\theta)$.

Assumption A7: $\left|\partial\left[\partial f\left(x,\theta\right)/\partial\theta_i . f\left(x,\theta\right)\right]/\partial\theta_j\right|, i, j = 1,..., p,$ are dominated by functions integrable with respect to *v* for all θ_* in Θ, and the minimal support of $f\left(x,\theta\right)$ does not depend on θ.

Given $d_l\left(x,\theta\right) = \partial \ln f\left(x,\theta\right)/\partial\theta_i . \partial \ln f\left(x,\theta\right)/\partial\theta_j + \partial^2 \ln f\left(x,\theta\right)/\partial\theta_i\partial\theta_j,$

$$\left(l = 1,..., p\left(p+1\right)/2; i = 1,..., p; j = i,..., p\right).$$

Assumption A8: $\partial d_l\left(x,\theta\right)/\partial\theta_k, l = 1,..., q, k = 1,..., p,$ exist and are continuous functions of θ for each *x*.

Assumption A9: $\left|d_l\left(x,\theta\right) d_m\left(x,\theta\right)\right|, \left|\partial d_l\left(x,\theta\right)/\partial\theta_k\right|,$ and $\left|d_l\left(x,\theta\right).\partial \ln f\left(x,\theta\right)/\partial\theta_k\right|,$ *k* = 1,…, *p*, $l, m = 1,..., q,$ are dominated by functions integrable with respect to *G* for all *x* and θ in Θ.

Assumption A10: $V\left(\theta_*\right)$ is nonsingular.

Theorem 1: (Information Matrix Test), (White, 1982, p.11) *Given Assumptions A1-A10, if* $g\left(u\right) = f\left(u,\theta_0\right)$ *for* θ_0 *in* Θ *, then (i)* $\sqrt{n}D_n\left(\hat{\theta}_n\right) \overset{A}{\sim} N\left(0, V\left(\theta_0\right)\right)$; *(ii)* $V_n\left(\hat{\theta}_n\right) \overset{\text{a.s.}}{\rightarrow} V\left(\theta_0\right)$, *and* $V_n\left(\hat{\theta}_n\right)$ *is nonsingular almost surely for all n sufficiently large; (iii) the information matrix test statistic*

$$\varsigma_n = nD_n\left(\hat{\theta}_n\right)' V_n\left(\hat{\theta}_n\right)^{-1} D_n\left(\hat{\theta}_n\right)$$

is distributed asymptotically as χ_q^2.

Theorem 2: (Consistency), (White, 1982, p.4) *Given Assumptions A1-A3,* $\hat{\theta}_n \to \theta_*$ *as* $n \to \infty$ *for almost every sequence (* X_t *); i.e.,* $\hat{\theta}_n \overset{\text{a.s.}}{\to} \theta_*$.

APPENDIX B

MEAN VALUE THEOREM

Mean Value Theorem: *Suppose* $f : R^p \to R$ *is defined on an open convex set* $\Theta \subset R^p$ *and is continuously differentiable in* Θ. *Let* ∇ *denote the* $p \times 1$ *vector of derivatives (gradient). Then there exists* θ^* *on a segment formed by any two points* θ *and* θ_0 *in* Θ *so that*

$$f(\theta) = f(\theta_0) + \nabla f(\theta^*)'(\theta - \theta_0).$$

APPENDIX C

PROOF OF RESULT 4.3.1

The proposed UMP test is developed based on Theorem 2, Section 3, page 68, of Lehmann (1959) which asserts that if the probability density $f(x;\mu)$ has a monotone likelihood ratio in $T(x)$, then there exists a UMP test for testing $H_0:\mu\leq\mu_0$ versus $H_1:\mu>\mu_0$. The monotone likelihood ratio is defined as follows:

Definition: The real-parameter family of probability density function (pdf) $f(x;\theta)$ is said to have a monotone likelihood ratio if there exists a real-valued function $T(x)$ such that $f(x;\theta^2)/f(x;\theta^1)$ is a non-decreasing function of $T(x)$ for $\theta^1<\theta^2$.

It suffices to show that the $f(\hat{\theta}_1;\theta_1)$ pdf of $\hat{\theta}_1$ has a monotone likelihood ratio.

Let $T(x)=\hat{\theta}_1$, $f(\hat{\theta}_1;\theta_1)=\dfrac{1}{\hat{\sigma}\sqrt{2\pi}}\exp\left\{-\left(\dfrac{\hat{\theta}_1-\theta_1}{2\hat{\sigma}}\right)^2\right\}$ where $\hat{\sigma}^2$ is the variance of the estimator $\hat{\theta}_1$. If $\theta_1^{'}<\theta_1^{''}$ consider the ratio

$$f(\hat{\theta}_1;\theta_1^{''})/f(\hat{\theta}_1;\theta_1^{'})=\exp\left\{-\left((\hat{\theta}_1-\theta_1^{''})^2-(\hat{\theta}_1-\theta_1^{'})^2\right)/2\hat{\sigma}^2\right\}$$

$$=\kappa\exp\left\{\hat{\theta}_1(\theta_1^{''}-\theta_1^{'})/\hat{\sigma}^2\right\},$$

where $\kappa=\exp\left\{-\left((\theta_1^{''})^2-(\theta_1^{'})^2\right)/2\hat{\sigma}^2\right\}$.

This ratio is a monotone increasing function in $\hat{\theta}_1$ when $\theta_1^{'}<\theta_1^{''}$. The result then follows from Theorem 2, Section 3, page 68, of Lehmann (1959).