Asymptotic Likelihood Inference for Sharpe Ratio

Ji Qi

A Dissertation Submitted To
The Faculty Of Graduate Studies
In Partial Fulfillment Of The Requirements
For The Degree Of
Doctor Of Philosophy

Graduate Program in Economics
York University
Toronto, Ontario

August, 2016

Abstract

The Sharpe ratio is one of the most widely used measures of the performance of an investment with respect to its return and risk. Since William Sharpe (1966) defined the ratio, as the fund's excess return per unit of risk measured by standard deviation, investments have been often ranked and evaluated on the basis of Sharpe ratio by both private as well as institutional investors. Our study on Sharpe ratio estimator is focused on its finite sample statistical properties which have being given less attention in practice.

Approximations aimed at improving the accuracy of likelihood method have been proposed over the past three decades. Among them, Lugannani and Rice (1980) and Barndorff-Nielsen (1986a) introduced two widely used tail area approximations with order of convergence $O(n^{-\frac{3}{2}})$. Furthermore, Fraser(1988; 1990), Fraser and Reid (1995), Fraser, Reid and Wu (1999) improved their methods and developed a general tail probability methodology, based on the tangent exponential model.

The objective of this paper is to use the third order asymptotic likelihood-based statistical method to obtain highly accurate inference on Sharpe ratio. Since the methodology is demonstrated to work well generally for any parametric distribution, our study will assume the market log returns are

independent identically distributed (IID) normal, or follow an autoregressive process of order one (AR(1)) with Gaussian white noise.

While most literature address large sample properties of the Sharpe ratio statistic (Lo 2002, Mertens 2002, Christie 2005, Bailey and Lopez de Prado 2012); it is important to compare the performance of investments when only small sample observations are available, especially before and after markets change direction. Our research would address this issue. New tests are developed for testing hypothesis on the Sharpe ratio calculated from one sample and on the difference of two Sharpe ratios. Comparison between our method and the currently existing methods in the literature are conducted by simulations. The p-values and confidence intervals for Sharpe ratio are calculated and various applications are illustrated.

Dedicated to KangKang

Acknowledgments

It is hard to believe I am running out of five years in my PhD life. Sometimes, I do dream I can stay in school for my whole life being like a kid or pretending to be one.

I would like to thank Professor Augustine Wong. A few days ago when I checked my email account, I found we had over 400 emails about my research during the past 3 years, with some emails long enough to be a letter size; and some immediate reply to my questions were sent even at midnight.

I would like to thank Professor Joann Jasiak. You are the one who lead me to the field of Financial Econometrics. Thank you for your acceptance to be my supervisor when I can hardly find any one right before my comprehensive exam.

I would like to thank Professor Shin-Hwan Chiang. Thank you for introducing me to York and encouraging me to pursue the PhD degree. Thank you for your help during my time in Canada.

I would like to thank Professor Barry Smith, Jianping Sun, Regina Pinto, Professor Sam Bucovetsky, Professor Kin Chung Lo, Professor Xianghong Li, Professor Tasso Adamopoulos, Professor Xueda Song, Professor John Paschakis, Xiang Li, Yi Zhang, Haokai Ning, Jing Zhang for all their sup-

port in my PhD program.

Finally, but most importantly, I am grateful to Hang Zheng.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Performance measurement is an integral part of investment analysis and risk management. Its goal is to build a ranking of different investments on the basis of risk-adjusted returns in order to evaluate the relative success of the investments. The Sharpe ratio is one of the most widely used measures of the performance of an investment with respect to its return and risk. Since William Sharpe (1966) defined the ratio, as the fund's excess return per unit of risk measured by standard deviation, investments have been often ranked and evaluated on the basis of Sharpe ratio by both private as well as institutional investors. The dominance of this performance measure is obvious and in literature the Sharpe ratio is referred as "the most common measure of risk-adjusted return" (Modigliani and Modigliani 1997).

Given its importance, the Sharpe ratio has been extensively studied in the literature. The study on Sharpe ratio can be classified into two main stream. One is the study regarding the structure of Sharpe ratio and the suitability of the Sharpe ratio as a suitable benchmark for portfolio performance evaluation. The other, which is also our goal of this study, is focused on the statistical properties of the Sharpe ratio estimator when it is used to measure risk and return characteristics of investments.

Approximations aimed at improving the accuracy of likelihood method have been proposed over the past three decades. Among them, Lugannani and Rice (1980) and Barndorff-Nielsen (1986a) introduced two widely used tail area approximations with order of convergence $O(n^{-\frac{3}{2}})$. Furthermore, Fraser(1988; 1990), Fraser and Reid (1995), Fraser, Reid and Wu (1999) improved their methods and developed a general tail probability methodology, based on the tangent exponential model. The objective of this paper is to use the third order asymptotic likelihood-based statistical method to obtain highly accurate inference on Sharpe ratio. Since the methodology is demonstrated to work well generally for any parametric distribution, our study will assume the market log returns are independent identically distributed (IID) normal, or follow an autoregressive process of order one (AR(1)) with

Gaussian white noise. Once we know the distributional assumption with even small sample, we can make extremely accurate inference on Sharpe Ratio based on our proposed method.

While most literature address large sample properties of the Sharpe ratio statistic (Lo 2002, Mertens 2002, Christie 2005, Bailey and Lopez de Prado 2012); it is important to compare the performance of investments when only small sample observations are available. Our research would address this issue. New tests are developed for testing hypothesis on the Sharpe ratio calculated from one sample and on the difference of two Sharpe ratios. Comparison between our method and the currently existing methods in the literature are conducted by simulations. The p-values and confidence intervals for Sharpe ratio are calculated and various applications are illustrated.

The organization of the dissertation is as follows: Chapter 1 provides a review of statistical properties on Sharpe ratio estimator. In particular, when the underlying log returns follow a normal distribution, the corresponding estimated Sharpe ratio would follow a noncentral t distribution exactly and follow a normal distribution asymptotically; in addition when the underlying returns follow an autoregressive process, we can obtain the corresponding distribution of estimated Share ratio asymptotically by maximum likelihood estimator and Delta's method. Chapter 2 details the mechanics of the likelihood-based third-order methods and the methodology can be applied to general statistical model. Chapter 3 studies risk-adjusted behaviors of investments by calculating highly accurate confidence intervals of the Sharpe ratio under normal assumption of underlying log returns. Applications on examples of small sample size shows difference between first-order results and third-order results, and the simulation studies testify the accuracy and stability of the third-order method. In Chapter 4, we set the assumption of log return to a more powerful and popular base, that is, the autoregressive time series model. The performance of the proposed method for time series data is also examined through both real-life data set and simulated small or medium data sets. Some discussion and future work are presented in Chapter 5.

1.1 Definition of Sharpe Ratio

The Sharpe ratio is one of the most popular measures available to money managers to examine the risk-adjusted performance of investments in Finance. In an investment asset or a trading strategy, the Sharpe ratio, named after William Sharpe (1966), measures the excess expected return or risk premium relative to its volatility. Expressed in functional form, the Sharpe ratio for an asset with an expected return μ and return standard deviation σ is given by the following:

$$SR = \frac{\mu - r_f}{\sigma} \ , \tag{1.1.1}$$

where r_f is the return on a benchmark asset, such as the risk-free rate of return.

This Sharpe ratio can be shown as the slope between risky and risk-free asset at (σ, μ) space. Also it measures the slope of the indifference curve in that space and a higher value implies higher mean variance expected utility. According to the mean-variance theory developed by Markowitz (1952) and the Capital Asset Pricing model (CAPM) developed by Sharpe (1964) and Lintner (1965), portfolios with highest Sharpe ratio are mean-variance efficient with this highest ratio being the slope of the Capital Market Line, and in equilibrium the market portfolio is one of those mean-variance efficient portfolios.

The (natural) estimator of the Sharpe ratio ($\widehat{SR}$) is:

$$\widehat{SR} = \frac{\hat{\mu} - r_f}{\hat{\sigma}} \,, \tag{1.1.2}$$

where $\hat{\mu}$ is the historical, or sample, mean return of funds, and $\hat{\sigma}$ is the sample standard deviation. Sharpe admits that one would ideally use predictions of return and volatility, but that "the predictions cannot be obtained in any satisfactory manner... Instead, ex post values must be used." (Sharpe 1966).

1.2 Literature Review of Sharpe Ratio

Before the distribution of the Sharpe ratio can be derived, assumptions must be made about the distribution of the underlying random variable of which the Sharpe ratio is a function: the return.

1.2.1 IID Normality Assumption on Return

Most financial studies involve returns instead of prices of asset. Campbell, Lo, and MacKinlay(1997) stated two reasons for using returns. First, for the average investor, financial markets may be considered close to be perfectly competitive, so that the size of the investment does not affect price changes. Therefore the return is a complete and scale-free summary of the investment opportunity. Second, for theoretical and empirical reasons, returns have more attractive statistical properties than prices, like stationarity, heavy tails, gain/loss asymmetry, volatility clustering, and so on.

We denote P_t as the price of an asset at date t and assume that this asset pays no dividends. The net return, R_t, of this asset between $t-1$ and t is defined as:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}} = \frac{P_t}{P_{t-1}} - 1 = g_t - 1 \,, \tag{1.2.1}$$

where g_t is the gross return or relative price of that asset. In addition,

the natural logarithm of the gross return is called the continuously compounded return or log return, denoted by r_t:

$$r_t = \log \frac{P_t}{P_{t-1}} = \log(g_t) \ . \tag{1.2.2}$$

While unequal at most conditions, net return and log return are approximately equal to each other when they are close to zero, which can be proved by Taylor approximation. Net return is ordinary in everyday use while log returns are useful for mathematical finance. One of the advantages for log return is its symmetry: positive and negative percentage ordinary net returns of equal magnitude do not cancel each other, but log returns of equal magnitude with opposite signs will cancel each other out and result in no change.

IID normal distribution has been widely proposed in the literature for the marginal distribution of asset's log returns. To specify, the log returns, r_t, are independent and identically distributed as normal distribution $r_t \sim N(\mu, \sigma^2)$. Or equivalently, the gross return g_t will follow an IID lognormal distribution, $g_t = (1 + R_t) \sim LN(\mu, \sigma^2)$.

The advantages of IID normal assumption on log return can be summarized into the following points. First, relative prices under lognormal assumption are always positive so that the net returns are well bounded below by -1. On the contrary, other assumptions like normal assumption on net return would make net returns unbounded from below. Second, the sum of finite number of IID normal random variables is still normal, so the multi-period log-return will still be normally distributed. Third, this assumption can be implied by stochastic process dynamics that underpins the option pricing theory. We will implement the geometric Brownian motion to derive the validity of normal assumption at the example below. Other arguments as to the plausibility of log-normal prices for the market portfolio appeared in He and Leland (1993).

Example. Consider asset price P_t, which follows the standard geometric Brownian motion.

$$d \log P_t = \mu dt + \sigma d\omega_t \ ,$$

where $\{\omega_t\}$ is a Brownian motion process. By Ito Lemma from stochastic calculus, we can have the following general expression for P_t and its one period lag P_{t-1}:

$$P_t = P_0 e^{\mu t + \sigma w_t} \ ,$$

$$P_{t-1} = P_0 e^{\mu(t-1) + \sigma w_{t-1}} \ .$$

Then we divide these two equations and get:

$$\frac{P_t}{P_{t-1}} = e^{\mu + \sigma(w_t - w_{t-1})} \ .$$

According to the definition of Brownian motion, $(w_T - w_t)$ are IID normal

as $N(0, T-t)$. Hence, we will end up with the outcome for gross return as well as log return:

$$g_t = \frac{P_t}{P_{t-1}} = e^{\mu+\sigma z} \ ,$$

$$r_t = \log \frac{P_t}{P_{t-1}} = \mu + \sigma z \ ,$$

with $Z \sim N(0,1)$. Therefore, our assumption on the return distribution has solid base: $r_t \sim N(\mu, \sigma^2)$ and $g_t \sim LN(\mu, \sigma^2)$.

□

1.2.2 Exact Statistical Properties of Estimated Sharpe Ratio under IID Normal Return or Central Limit Theorem

Let $r_1, \ldots, r_n$ be IID draws from a normal distribution $N\left(\mu, \sigma^2\right)$. Then the unbiased sample mean and variance are $\hat{\mu} = \frac{\sum r_t}{n} = \bar{r}$ and $\hat{\sigma}^2 = \frac{\sum (r_t - \hat{\mu})^2}{n-1}$. Then we have the exact distribution of the estimated Sharpe ratio described as follows:

Lemma. *The **exact distribution of the estimated Sharpe ratio** is a noncentral t distribution with degrees of freedom $n-1$ and noncentral parameter $\sqrt{n}\frac{\mu - r_f}{\sigma} = \sqrt{n} \cdot SR$.*

$$\widehat{SR} = \frac{\hat{\mu} - r_f}{\hat{\sigma}} \sim \frac{1}{\sqrt{n}} T_{n-1}\left(\sqrt{n}\frac{\mu - r_f}{\sigma}\right) = \frac{1}{\sqrt{n}} T_{n-1}\left(\sqrt{n} \cdot SR\right) \ .$$

Proof. Generally, a noncentral t distribution, explicitly noted $T_\nu(\delta)$, with degrees of freedom parameter ν and noncentrality value δ, is defined as

$$T_\nu(\delta) = \frac{Z + \delta}{\sqrt{\frac{\chi^2_\nu}{\nu}}} \ . \tag{1.2.3}$$

The $T_\nu(\delta)$ statistic represents the quotient of a standard normal random variable Z displaced by a constant δ over the square root of a Chi-square χ^2_ν random variable divided by its degrees of freedom ν, if these two distributions are independent of each other. In particular, when the noncentral parameter is zero, $T_\nu(0)$ coincides with the standard central t distribution T_ν.

Since $r_1, \ldots, r_n$ are IID normal $N\left(\mu, \sigma^2\right)$, it is well known that $\hat{\mu} = \bar{r} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and $z = \sqrt{n}\frac{\bar{r} - \mu}{\sigma}$; at the same time $\frac{(n-1)\hat{\sigma}^2}{\sigma^2} = \frac{\sum (r_t - \hat{\mu})^2}{\sigma^2} \sim \chi^2_{n-1}$.

Thus,

$$\widehat{SR} = \frac{\hat{\mu} - r_f}{\hat{\sigma}} = \frac{1}{\sqrt{n}} \cdot \frac{\sqrt{n}\left(\hat{\mu} - r_f\right)}{\hat{\sigma}} = \frac{1}{\sqrt{n}} \frac{\frac{\sqrt{n}(\bar{r}-\mu)}{\sigma} + \frac{\sqrt{n}(\mu - r_f)}{\sigma}}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}}$$
$$\sim \frac{1}{\sqrt{n}} T_{n-1}\left(\sqrt{n}\frac{\mu - r_f}{\sigma}\right) .$$

□

Thus, the distribution of estimated Sharpe ratio assuming IID normal returns follows a rescaled noncentral t distribution, where the noncentrality parameter defined with population quantities depends only on the true Sharpe ratio SR and the sample size n. Knowing the distribution of estimated Sharpe ratio is empowering, as interesting facts about the noncentral t distribution or t test can be translated into interesting facts about the true Sharpe ratio: one can construct hypothesis tests for the SR, find the power and sample size of those tests, compute confidence intervals of SR, correct for deviations from assumptions.

Example. There are a number of statistical tests involving the Sharpe ratio or variants. Here are two examples from Scholz (2007).

(1). The classical one-sample test for mean involves (central) t statistic which is like a Sharpe ratio variant. Thus, to test:

$$H_0 : \mu \leq \mu_0 \text{ versus } H_1 \text{: } \mu > \mu_0$$

we reject H_0 if

$$t = \frac{\hat{\mu} - \mu_0}{\sqrt{\frac{\hat{\sigma}^2}{n}}} > t_{1-\alpha,\nu=n-1} ,$$

where $t_{1-\alpha,\nu=n-1}$ is the $1-\alpha$ quartile of the central t distribution with $n-1$ degrees of freedom. Now if $\mu = \mu_1 > \mu_0$, then the power of this test is

$$1 - F\left(t_{1-\alpha,\nu=n-1}; n-1; \sqrt{n}\frac{\mu_1 - \mu_0}{\sigma}\right) ,$$

where $F\left(x; \nu; \delta\right)$ is the cumulative distribution function of the noncentral t distribution with noncentrality parameter δ and degrees of freedom ν.

(2). A one-sample test for the population Sharpe ratio involves the noncentral t statistic. To test:

$$H_0 : SR \leq SR_0 \text{ versus } H_1 \text{: } SR > SR_0$$

we reject H_0 if

$$t = \sqrt{n}\widehat{SR} > t_{1-\alpha,\nu=n-1}\left(\sqrt{n} \cdot SR_0\right) .$$

Now if $SR = SR_1 > SR_0$, then the power of this test is

$$1 - F\left(t_{1-\alpha,\nu=n-1}\left(\sqrt{n}SR_0\right); n-1; \sqrt{n}\cdot SR_1\right) \ .$$

(3). We can get confidence intervals on population SR by inversion of the cumulative distribution function of the noncentral t distribution (e.g., by Brent's method 2013), which is computationally slower than approximations based on asymptotic normality to be introduced later in subsection (1.2.3). A $(1-\alpha)\times 100\%$ symmetric confidence interval on population SR has endpoints (SR_l, SR_u) defined implicitly by $1-\alpha/2 = F\left(\widehat{SR}; n-1; \sqrt{n}SR_l\right)$ and $\alpha/2 = F\left(\widehat{SR}; n-1; \sqrt{n}SR_u\right)$.

□

Before we end this subsection, we look into the moments of the estimated Sharpe ratio. According to Hogben, Pinkham and Wilk (1961), the k th raw moment of the noncentral t-distribution is generally

$$E\left[\left(T_\nu\left(\delta\right)\right)^k\right] = \begin{cases} \left(\frac{\nu}{2}\right)^{\frac{k}{2}} \frac{\Gamma\left(\frac{\nu-k}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \exp\left(-\frac{\delta^2}{2}\right) \frac{d^k}{d\delta^k}\exp\left(\frac{\delta^2}{2}\right), & \text{if } \nu > k; \\ \text{Does not exist}, & \text{if } \nu \le k. \end{cases}$$

In particular, the mean and variance of the noncentral t distribution are

$$E\left[T_\nu\left(\delta\right)\right] = \begin{cases} \delta\sqrt{\frac{\nu}{2}}\frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)} = \delta d_{\nu+1}, & \text{if } \nu > 1; \\ \text{Does not exist}, & \text{if } \nu \le 1, \end{cases}$$

and

$$Var\left[T_\nu\left(\delta\right)\right] = \begin{cases} \frac{\nu(1+\delta^2)}{\nu-2} - \frac{\delta^2\nu}{2}\left(\frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}\right)^2 = \frac{\nu(1+\delta^2)}{\nu-2} - \left\{E\left[T_v\left(\delta\right)\right]\right\}^2, & \text{if } \nu > 2; \\ \text{Does not exist}, & \text{if } \nu \le 2. \end{cases}$$

With regard to the third moment, the noncentral t-distribution is asymmetric unless δ is zero, i.e., a central t-distribution. The right tail will be heavier than the left when $\delta > 0$, and vice versa. However, the usual skewness is not generally a good measure of asymmetry for this distribution, because if the degrees of freedom is not larger than 3, the third moment does not exist at all. Even if the degrees of freedom is greater than 3, the sample estimate of the skewness is still very unstable unless the sample size is very large.

The moments of the noncentral t distribution can be trivially translated into equivalent facts regarding the estimated Sharpe ratio:

$$E\left[\widehat{SR}\right] = \begin{cases} SR\cdot d_n, & \text{if } n > 2; \\ \text{Does not exist}, & \text{if } n \le 2, \end{cases}$$

$$Var\left[\widehat{SR}\right] = \begin{cases} \frac{(n-1)(1+n\cdot SR^2)}{n(n-3)} - \left\{E\left[\widehat{SR}\right]\right\}^2, & \text{if } n > 3; \\ \text{Does not exist}, & \text{if } n \leq 3. \end{cases}$$

Thus, we can see that $E\left[\widehat{SR}\right] \neq SR$; rather there is a systematic geometric bias $d_n > 1$, implying that the estimated Sharpe ratio will overestimate the population Sharpe ratio when the latter is positive, and underestimate it when it is negative (Miller and Gehr 1978; Jobson and Korkie 1981). The bias term $d_{\nu+1} = \sqrt{\frac{\nu}{2}}\frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}$ is a function of sample size only with reasonable asymptotic approximation $1 + \frac{3}{4\nu} + \frac{25}{32\nu^2} + \frac{105}{128\nu^3} + O\left(n^{-4}\right)$. [1]

1.2.3 Asymptotic Statistical Properties of Estimated Sharpe Ratio

1.2.3.1 General Setting on Asymptotic Statistical Properties of Estimated Sharpe Ratio

The asymptotic distribution of estimated Sharpe ratio is derived in this subsection. The following derivations are based on Jobson and Korkie(1981), Lo(2002), Mertens (2002), Leung and Wong (2006), Ledoit and Wolf (2008) and Wright, Yam and Yung (2011). Their methodology are mainly Delta method and Central Limit Theorem, and thus, imply only first order asymptotic results. We will set them as benchmark to compare with our third order likelihood-based method later.

Consider a general case of p possibly correlated daily return streams

[1] According to Tricomi and Erdélyi (1951) and Olver, Lozier, Boisvert, and Clark (2010 eq5.11.13), we can obtain the asymptotic expansion of the quotient of two gamma function by the following method: Firstly, we establish a infinite series $\{A_n(\alpha)\}$ with recurrence relation $A_n(\alpha) = \frac{1}{n}\sum_{m=0}^{n-1}\binom{\alpha-m}{n-m+1}A_m(\alpha)$ $(n = 1, 2, \ldots$ and $\forall\alpha)$ and its initial conditions $A_0(\alpha) = 1$, $A_1(\alpha) = \binom{\alpha}{2}$, $A_2(\alpha) = \frac{3\alpha-1}{4}\binom{\alpha}{3}$, $A_3(\alpha) = \binom{\alpha}{2}\binom{\alpha}{4}$,.... Now, if we put $C_n(\alpha' = \alpha - \beta, \beta) = \sum_{m=0}^{n}\binom{\alpha'-m}{n-m}A_m(\alpha')\beta^{n-m}$ $(n = 0, 1, 2, \ldots)$ and its initial conditions $C_0 = 1$, $C_1 = \frac{1}{2}\alpha'(\alpha' + 2\beta - 1)$, $C_2 = \frac{1}{12}\binom{\alpha'}{2}[(\alpha'-2)(3\alpha'-1) + 12\beta(\alpha'+\beta-1)]$, ... on the whole ν-plane cut along any curve connecting $\nu = 0$ with $\nu = \infty$, we have $\frac{\Gamma(\nu+\alpha)}{\Gamma(\nu+\beta)} = \sum_{n=0}^{\infty}C_n(\alpha', \beta)\nu^{\alpha'-n}$, provided that ν avoids the points $\nu = -\alpha, -\alpha-1, -\alpha-2, \ldots$ and $\nu = -\beta, -\beta-1, -\beta-2, \ldots$.

On the other hand, people can get this expansion result directly at http://www.wolframalpha.com/ by the code " Series[Sqrt[n/2] Gamma[(n-1)/2]/Gamma[n/2], {n, \[Infinity], 5} ".

Another important expansion that can be derived by the same method and will be used later is

$$b_\nu = \sqrt{\frac{2}{\nu}}\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} = 1 - \frac{1}{4\nu} + \frac{1}{32\nu^2} + \frac{5}{128\nu^3} + O\left(\nu^{-4}\right)$$

over the past n days, denoted by $\mathbf{r} = [r_{i,j}]$ where $i \in \{1, 2, \ldots, p\}$ and $j \in \{1, 2, \ldots, n\}$. For return stream r_i, μ_i is the population mean $E[r_i]$ and $\hat{\mu}_i = \frac{\sum_{j=1}^{n} r_{i,j}}{n} = \bar{r}_i$ is the sample mean and unbiased estimator for the first raw moment; $m'_{2,i}$ is the uncentered second raw moment $E\left[r_i^2\right]$ and $\hat{m}'_{2,i} = \frac{\sum_{j=1}^{n} r_{i,j}^2}{n} = \bar{r_i^2}$ is the sample mean of the squared returns and unbiased estimator for the second raw moment.[2] We shall assume that the daily p variate vectors of returns are IID and that the daily return for each fund has a finite fourth moment. These conditions are general enough to include a wealth of return models (including Levy processes) and are required in order for the Central Limit Theorem to be used. Under the multivariate Central Limit Theorem (Wasserman 2013) [3], we have

$$\sqrt{n}\left(\begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_p \\ \hat{m}'_{2,1} \\ \vdots \\ \hat{m}'_{2,p} \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \\ m'_{2,1} \\ \vdots \\ m'_{2,p} \end{pmatrix}\right) = \sqrt{n}\left(\begin{pmatrix} \bar{r}_1 = \frac{\sum_{j=1}^{n} r_{1,j}}{n} \\ \vdots \\ \bar{r}_p = \frac{\sum_{j=1}^{n} r_{p,j}}{n} \\ \bar{r_1^2} = \frac{\sum_{j=1}^{n} r_{1,j}^2}{n} \\ \vdots \\ \bar{r_p^2} = \frac{\sum_{j=1}^{n} r_{p,j}^2}{n} \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \\ m'_{2,1} \\ \vdots \\ m'_{2,p} \end{pmatrix}\right)$$

$$\xrightarrow{d} N\left(\mathbf{0}, \mathbf{\Omega} = Var\begin{pmatrix} r_1 \\ \vdots \\ r_p \\ r_1^2 \\ \vdots \\ r_p^2 \end{pmatrix} = \begin{pmatrix} Var\begin{pmatrix} r_1 \\ \vdots \\ r_p \end{pmatrix} & \cdot \\ \left(Cov\begin{pmatrix} r_1 & r_1^2 \\ \vdots & \vdots \\ r_p & r_p^2 \end{pmatrix}\right)' & Var\begin{pmatrix} r_1^2 \\ \vdots \\ r_p^2 \end{pmatrix} \end{pmatrix}\right),$$

where "$\xrightarrow{d}$" represents asymptotic convergence in distribution. Since gener-

[2] In this thesis, we use m_n to denote the nth central moment and m'_n to denote the nth uncentered raw moment and $\alpha_n = \frac{m_n}{\sigma^n}$ to denote the nth standardized moment. For more information, see subsection (2.4.1.1)

[3] The multivariate Central Limit Theorem states that if $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are vectors of independent observations from any population with mean vector $\boldsymbol{\mu}$ and finite covariance matrix $\mathbf{\Omega}$, then the sample mean vector $\bar{\mathbf{Y}}$ follows

$$\sqrt{n}\left(\bar{\mathbf{Y}} - \boldsymbol{\mu}\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{\Omega}\right)$$

Note that, our result differs from Leung and Wong (2006) in the way that the authors in their paper derive the asymptotic distribution for the $2p \times 1$ vector $\left(\hat{\mu}_1, \ldots, \hat{\mu}_p, \hat{\sigma}_1^2 = \hat{m}_{2,1}, \ldots, \hat{\sigma}_p^2 = \hat{m}_{2,p}\right)'$ while we derive that of $\left(\hat{\mu}_1, \ldots, \hat{\mu}_p, \hat{m}'_{2,1}, \ldots, \hat{m}'_{2,p}\right)'$.

ally $Var(r_i) = E\left(r_i^2\right) - \left(E(r_i)\right)^2$, or $\sigma_i^2 = m'_{2,i} - \mu_i^2$, we have $SR_i = \frac{\mu_i - r_f}{\sigma_i} = \frac{\mu_i - r_f}{\sqrt{m'_{2,i} - \mu_i^2}}$. By the multivariate Delta method, we can find the asymptotic distribution of the $p \times 1$ vector of Sharpe ratio estimates $\widehat{\mathbf{SR}} = \left[\widehat{SR}_i\right]$.

$$
\begin{aligned}
& \sqrt{n}(\widehat{\mathbf{SR}} - \mathbf{SR}) \\
\xrightarrow{d} \quad & N(\mathbf{0}, \mathbf{\Sigma} = \nabla \mathbf{SR} \cdot \mathbf{\Omega} \cdot (\nabla \mathbf{SR})') \\
\xrightarrow{d} \quad & N\left(\mathbf{0}, \left(\frac{\partial \mathbf{SR}}{\partial\left(\mu_1, \ldots, \mu_p, m'_{2,1}, \ldots, m'_{2,p}\right)}\right) \mathbf{\Omega} (\cdot)'\right) \\
\xrightarrow{d} \quad & N\left(\mathbf{0}, \left(\text{diag}\left(\frac{\sigma_i + \mu_i SR_i}{\sigma_i^2}\right) \vdots \text{diag}\left(-\frac{SR_i}{2\sigma_i^2}\right)\right) \mathbf{\Omega} (\cdot)'\right) \\
\xrightarrow{d} \quad & N\left(\mathbf{0}, \left(\frac{\partial \widehat{\mathbf{SR}}}{\partial\left(\hat{\mu}_1, \ldots, \hat{\mu}_p, \hat{m}'_{2,1}, \ldots, \hat{m}'_{2,p}\right)}\right) \hat{\mathbf{\Omega}} (\cdot)'\right) \ .
\end{aligned}
$$

Note that $\nabla \mathbf{SR}$ takes the form of two $p \times p$ diagonal matrices augmented together side by side. In practice, population values $\mu_1, \ldots, \mu_p$,$m'_{2,1}, \ldots, m'_{2,p}$, $\mathbf{\Omega}$ are all unknown, and so the asymptotic variance has to be estimated, using the sample estimates (Lo 2002, Mertens 2002).

1.2.3.2 Asymptotic Statistical Properties of Estimated Sharpe Ratio Under IID Return (I)

Based on the content introduced at last subsection, Jobson and Korkie (1981) and Lo (2002) derived the asymptotic distribution of estimated Sharpe ratio given IID returns (1.2.4); Mertens (2002) enhanced the result and obtained (1.2.5).

Lemma. *Given IID returns, the asymptotic distributions of estimated Sharpe ratio are*

$$\sqrt{n}(\widehat{SR} - SR) \xrightarrow{d} N(0, 1 + \frac{1}{2}\widehat{SR}^2) \, , \tag{1.2.4}$$

$$\sqrt{n}(\widehat{SR} - SR) \xrightarrow{d} N(0, 1 + \frac{\widehat{SR}^2}{2} - \hat{\alpha}_3 \widehat{SR} + \frac{\hat{\alpha}_4 - 3}{4}\widehat{SR}^2) \, , \tag{1.2.5}$$

with $(1-\alpha) \times 100\%$ *confidence interval for population Sharpe ratio:* $\widehat{SR} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{1}{n}(1 + \frac{1}{2}\widehat{SR}^2)}$ *and* $\widehat{SR} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{1}{n}(1 + \frac{\widehat{SR}^2}{2} - \hat{\alpha}_3 \widehat{SR} + \frac{\hat{\alpha}_4 - 3}{4}\widehat{SR}^2)}$.

Proof. From the results introduced at last subsection, we study the $p = 1$

case. $\mathbf{\Omega}$ takes the form

$$\mathbf{\Omega} = Var\begin{pmatrix} r \\ r^2 \end{pmatrix} = \begin{pmatrix} E\left[r^2\right] - \left(E\left[r\right]\right)^2 & E\left[r^3\right] - E\left[r\right]E\left[r^2\right] \\ E\left[r^3\right] - E\left[r\right]E\left[r^2\right] & E\left[r^4\right] - \left(E\left[r^2\right]\right)^2 \end{pmatrix}$$
$$= \begin{pmatrix} m_2' - \mu^2 & m_3' - \mu m_2' \\ m_3' - \mu m_2' & m_4' - \left(m_2'\right)^2 \end{pmatrix} = \begin{pmatrix} \sigma^2 & m_3 + 2\sigma^2\mu \\ m_3 + 2\sigma^2\mu & m_4 + 4m_3\mu + 4\sigma^2\mu^2 - \sigma^4 \end{pmatrix}$$
$$= \sigma^2 \begin{pmatrix} 1 & \alpha_3\sigma + 2\mu \\ \alpha_3\sigma + 2\mu & (\alpha_4 - 1)\sigma^2 + 4\alpha_3\sigma\mu + 4\mu^2 \end{pmatrix}.$$

Additionally from $\frac{\partial SR}{\partial\left(\mu, m_2'\right)} = \left(\frac{\sigma + \mu SR}{\sigma^2}, -\frac{SR}{2\sigma^2}\right)$, the asymptotic variance of estimated Sharpe ratio and the asymptotic distribution of estimated Sharpe ratio are obtained

$$\begin{aligned} \sqrt{n}(\widehat{SR} - SR) &\xrightarrow{d} N(0, 1 - \alpha_3 SR + \frac{\alpha_4 - 1}{4} SR^2) \\ &\xrightarrow{d} N\left(0, 1 + \frac{SR^2}{2} - \alpha_3 SR + \frac{\alpha_4 - 3}{4} SR^2\right) \\ &\xrightarrow{d} N\left(0, 1 + \frac{\widehat{SR}^2}{2} - \hat{\alpha}_3\widehat{SR} + \frac{\hat{\alpha}_4 - 3}{4}\widehat{SR}^2\right). \end{aligned}$$

□

Here are some important comments for these two asymptotic distribution.

1. Note that for normally distributed returns, the skewness α_3 and (historical) kurtosis α_4 of the returns distribution are both zero, and so Mertens' form reduces to Jobson and Korkie's form. These are unknown in practice, and have to be estimated from the data, which results in some mis-estimation of the standard error when skew is extreme.

2. Since the population SR is unknown, Lo suggests to approximate it with the estimated Sharpe ratio $\widehat{SR}$. In practice, the asymptotically equivalent form $\widehat{SR} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{1 + \frac{1}{2}\widehat{SR}^2}{n-1}}$ has better small sample coverage for normal returns.

1.2.3.3 Asymptotic Statistical Properties of Estimated Sharpe Ratio Under IID Return (II)

Another way to obtain the asymptotic statistical properties of estimated Sharpe ratio is by asymptotic approximation to noncentral t distribution, which is the exact distribution of the estimated Sharpe ratio under IID normal return or Central Limit Theorem.

Bentkus, Jing, Shao and Zhou (2007) investigated the limiting behavior of the noncentral t statistic and gave a systematic description of its limiting distribution. They showed that by only assuming that $E\left[r_t^2\right] < \infty$, the limiting distribution of the noncentral t statistic can be nonnormal while those of the central statistic are known to be asymptotically normal. In fact, when $E\left[r_t^2\right] < \infty$, the asymptotic behavior of the noncentral t statistic critically depends on whether or not $E\left[r_t^4\right] = \infty$: if $E\left[r_t^4\right] < \infty$, the limiting distributions of $T_\nu(\delta)$ are related to the normal or a square of normal distribution; if $E\left[r_t^4\right] = \infty$, the limit is related to other stable distributions. All of them have very different convergence rates.

In particular, assuming $E\left[r_t^4\right] < \infty$ and r_t is not a specific linear function of a standardized Bernoulli random variable, Bentkus, Jing, Shao and Zhou (2007) derived the following limiting distribution of the noncentral t statistic

$$T_{n-1}(\delta) - \delta \xrightarrow{d} N\left(0, 1 - \frac{\delta}{\sqrt{n}}\alpha_3 + \frac{\delta^2(\alpha_4 - 1)}{4n}\right).$$

We can observe that Mertens' result (Mertens 2002) replicates this known asymptotic distribution. In addition, Akahira(1995) obtained another higher order approximation to noncentral t distribution.

Lemma. *Akahira (1995) derived the following higher order approximation result:*

$$\frac{\widehat{SR}b_{n-1} - SR}{\sqrt{\frac{1}{n} + \widehat{SR}^2\left(1 - b_{n-1}^2\right)}} = Z - \frac{\widehat{SR}^3\left(Z^2 - 1\right)}{24\left(\frac{1}{n} + \widehat{SR}^2\left(1 - b_{n-1}^2\right)\right)^{\frac{3}{2}}}\left\{\frac{1}{(n-1)^2} + \frac{1}{4(n-1)^3}\right\} + O\left(n^{-4}\right),$$

with confidence interval

$$\left(\widehat{SR}b_{n-1} - Z_{\frac{\alpha}{2}}\sqrt{\frac{1}{n} + \widehat{SR}^2\left(1 - b_{n-1}^2\right)} + \frac{\widehat{SR}^3\left(Z_\alpha^2 - 1\right)}{24\left(\frac{1}{n} + \widehat{SR}^2\left(1 - b_{n-1}^2\right)\right)}\left\{\frac{1}{(n-1)^2} + \frac{1}{4(n-1)^3}\right\} + O\left(n^{-4}\right),\right.$$
$$\left.\widehat{SR}b_{n-1} + Z_{\frac{\alpha}{2}}\sqrt{\frac{1}{n} + \widehat{SR}^2\left(1 - b_{n-1}^2\right)} - \frac{\widehat{SR}^3\left(Z_\alpha^2 - 1\right)}{24\left(\frac{1}{n} + \widehat{SR}^2\left(1 - b_{n-1}^2\right)\right)}\left\{\frac{1}{(n-1)^2} + \frac{1}{4(n-1)^3}\right\} + O\left(n^{-4}\right)\right).$$

Proof. From the definition of noncentral t distribution (1.2.3), letting $S = \sqrt{\frac{\chi_\nu^2}{\nu}} = \frac{\chi_\nu}{\sqrt{\nu}}$, we have $T_\nu(\delta) = \frac{Z+\delta}{S}$. Since Z and χ_ν^2 are independent, Z and S are independent too.

Given the first few moments of χ_ν, we can get the first few moments of S. [4]

$$E[S] = \sqrt{\frac{2}{\nu}}\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} = \frac{1}{d_{\nu+2}}\sqrt{\frac{\nu+1}{\nu}} = b_\nu ,$$

$$E\left[S^2\right] = \frac{E\left[\chi_\nu^2\right]}{\nu} = 1 ,$$

$$E\left[S^3\right] = \frac{E\left[\chi_\nu^3\right]}{\left(\sqrt{\nu}\right)^3} = \left(1+\frac{1}{\nu}\right) b_\nu ,$$

$$E\left[S^4\right] = \frac{E\left[\chi_\nu^4\right]}{\nu^2} = 1+\frac{2}{\nu} ,$$

$$E\left(S - E[S]\right)^2 = Var[S] = \frac{Var[\chi_\nu]}{\nu} = 1 - b_\nu^2 ,$$

$$E\left(S - E[S]\right)^3 = \frac{E\left(\chi_\nu - E[\chi_\nu]\right)^3}{\left(\sqrt{\nu}\right)^3} = b_\nu\left\{2\left(b_\nu^2 - 1\right) + \frac{1}{\nu}\right\} ,$$

$$E\left(S - E[S]\right)^4 = \frac{E\left(\chi_\nu - E[\chi_\nu]\right)^4}{\nu^2} = \frac{2}{\nu}\left(1 - 2b_\nu^2\right) + \left(1 - b_\nu^2\right)\left(1 + 3b_\nu^2\right) .$$

For any α with condition $0 < \alpha < 1$, there exists a $t_{\alpha,\nu}(\delta)$ such that $P\{T_\nu(\delta) < t_{\alpha,\nu}(\delta)\} = 1 - \alpha$. The $t_{\alpha,\nu}(\delta)$ is called the upper 100α percentile

[4] $E[\chi_\nu] = \mu_{\chi_\nu} = \sqrt{2}\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} = \frac{\sqrt{\nu+1}}{d_{\nu+2}} = b_\nu\sqrt{\nu}$
$E\left[\chi_\nu^2\right] = \nu$
$E\left[\chi_\nu^3\right] = 2\sqrt{2}\frac{\Gamma((\nu+3)/2)}{\Gamma(\nu/2)} = \sqrt{\nu}(\nu+1)b_\nu$
$E\left[\chi_\nu^4\right] = \nu(\nu+2)$
$E\left(\chi_\nu - E[\chi_\nu]\right)^2 = \sigma_{\chi_\nu}^2 = E\left[\chi_\nu^2\right] - \mu_{\chi_\nu}^2 = \nu\left(1 - b_\nu^2\right)$
$E\left(\chi_\nu - E[\chi_\nu]\right)^3 = E\left[\chi_\nu^3\right] - 3E\left[\chi_\nu^2\right]\mu_{\chi_\nu} + 2\mu_{\chi_\nu}^3 = \sqrt{\nu}b_\nu\left\{1 + 2\nu\left(b_\nu^2 - 1\right)\right\}$
$E\left(\chi_\nu - E[\chi_\nu]\right)^4 = E\left[\chi_\nu^4\right] - 4E\left[\chi_\nu^3\right]\mu_{\chi_\nu} + 6E\left[\chi_\nu^2\right]\mu_{\chi_\nu}^2 - 3\mu_{\chi_\nu}^4 = 2\nu\left(1 - 2b_\nu^2\right) + \nu^2\left(1 - b_\nu^2\right)\left(1 + 3b_\nu^2\right)$

of the noncentral t distribution. Then we have

$$\begin{aligned}
1-\alpha &= P\left\{T_\nu(\delta) < t_{\alpha,\nu}(\delta)\right\} \\
&= P\left\{\frac{Z+\delta}{S} < t_{\alpha,\nu}(\delta)\right\} = P\left\{Z - t_{\alpha,\nu}(\delta) S < -\delta\right\} \\
&= P\left\{\frac{Z - t_{\alpha,\nu}(\delta) S - E\left[Z - t_{\alpha,\nu}(\delta) S\right]}{Var\left[Z - t_{\alpha,\nu}(\delta) S\right]} < \frac{-\delta - \left[Z - t_{\alpha,\nu}(\delta) S\right]}{Var\left[Z - t_{\alpha,\nu}(\delta) S\right]}\right\} \\
&= P\left\{W = \frac{Z - t_{\alpha,\nu}(\delta)(S - b_\nu)}{\sqrt{1 + t^2_{\alpha,\nu}(\delta)(1 - b^2_\nu)}} < \frac{t_{\alpha,\nu}(\delta) b_\nu - \delta}{\sqrt{1 + t^2_{\alpha,\nu}(\delta)(1 - b^2_\nu)}}\right\}.
\end{aligned}$$

Note that the statistic W is based on a linear combination of a normal random variable Z and a chi-statistic S, with $E[W] = 0$ and $Var[W] = 1$. In order to use the Cornish-Fisher expansion for the statistic W up to the order $O\left(\nu^{-3}\right)$, we need the third and fourth cumulants of W up to the same order.

$$\begin{aligned}
&\kappa_3\left(W = \frac{Z - t_{\alpha,\nu}(\delta)(S - b_\nu)}{\sqrt{1 + t^2_{\alpha,\nu}(\delta)(1 - b^2_\nu)}}\right) \\
&= \frac{\kappa_3(Z) - (t_{\alpha,\nu}(\delta))^3 \kappa_3(S - b_\nu)}{\left(1 + t^2_{\alpha,\nu}(\delta)(1 - b^2_\nu)\right)^{\frac{3}{2}}} \\
&= \frac{0 - (t_{\alpha,\nu}(\delta))^3 E(S - E[S])^3}{\left(1 + t^2_{\alpha,\nu}(\delta)(1 - b^2_\nu)\right)^{\frac{3}{2}}} \\
&= \frac{(t_{\alpha,\nu}(\delta))^3 b_\nu}{\left(1 + t^2_{\alpha,\nu}(\delta)(1 - b^2_\nu)\right)^{\frac{3}{2}}}\left\{2\left(1 - b^2_\nu\right) - \frac{1}{\nu}\right\} \\
&\quad \text{with } b_\nu \text{expanded by footnote 1} \\
&= \frac{(t_{\alpha,\nu}(\delta))^3}{\left(1 + t^2_{\alpha,\nu}(\delta)(1 - b^2_\nu)\right)^{\frac{3}{2}}} \cdot \left(-\frac{1}{4}\right)\left\{\frac{1}{\nu^2} + \frac{1}{4\nu^3} + O\left(\nu^{-4}\right)\right\}.
\end{aligned}$$

$$
\begin{aligned}
& \kappa_4\left(W=\frac{Z-t_{\alpha,\nu}(\delta)(S-b_\nu)}{\sqrt{1+t_{\alpha,\nu}^2(\delta)(1-b_\nu^2)}}\right) \\
= & \frac{\kappa_4(Z)+(t_{\alpha,\nu}(\delta))^4\kappa_4(S-b_\nu)}{\left(1+t_{\alpha,\nu}^2(\delta)(1-b_\nu^2)\right)^2} \\
= & \frac{0+(t_{\alpha,\nu}(\delta))^4\left\{E(S-E[S])^4-3\left[E(S-E[S])^2\right]^2\right\}}{\left(1+t_{\alpha,\nu}^2(\delta)(1-b_\nu^2)\right)^2} \\
= & \frac{(t_{\alpha,\nu}(\delta))^4\left\{\frac{2}{\nu}\left(1-2b_\nu^2\right)+\left(1-b_\nu^2\right)\left(1+3b_\nu^2\right)-3\left[1-b_\nu^2\right]^2\right\}}{\left(1+t_{\alpha,\nu}^2(\delta)(1-b_\nu^2)\right)^2} \\
= & \frac{2(t_{\alpha,\nu}(\delta))^4}{\left(1+t_{\alpha,\nu}^2(\delta)(1-b_\nu^2)\right)^2}\left\{\left(1-b_\nu^2\right)\left(3b_\nu^2-1\right)+\frac{1}{\nu}\left(1-2b_\nu^2\right)\right\} \\
& \text{with } b_\nu \text{expanded by footnote 1} \\
= & O\left(\nu^{-4}\right) .
\end{aligned}
$$

The fourth cumulant is usually of order ν^{-3}, but in this case the term of the order vanishes (Kendall and Stuart 1969 p372).

By the Cornish-Fisher expansion, we can obtain a higher order approximation formula of a percentage point of the noncentral t distribution, $t_{\alpha,\nu}(\delta)$.

$$
\begin{aligned}
& \frac{t_{\alpha,\nu}(\delta)b_\nu-\delta}{\sqrt{1+t_{\alpha,\nu}^2(\delta)(1-b_\nu^2)}} \\
= & z_\alpha+\frac{1}{6}\kappa_3(W)\left(z_\alpha^2-1\right)+\frac{1}{24}\kappa_4(W)\left(z_\alpha^4-3z_\alpha\right)+O\left(\nu^{-4}\right) \\
= & z_\alpha-\frac{(t_{\alpha,\nu}(\delta))^3\left(z_\alpha^2-1\right)}{24\left(1+t_{\alpha,\nu}^2(\delta)(1-b_\nu^2)\right)^{\frac{3}{2}}}\left\{\frac{1}{\nu^2}+\frac{1}{4\nu^3}\right\}+O\left(\nu^{-4}\right) ,
\end{aligned}
$$

where z_α is the upper 100α percentile of the standard normal distribution. Regarding it as an equation of $t_{\alpha,\nu}(\delta)$, the existence and uniqueness of a solution of the equation is guaranteed when $0.1 \leq \alpha \leq 0.15$ for $\nu = 1$, $0.03 \leq \alpha \leq 0.15$ for $\nu = 2$, $0.006 \leq \alpha \leq 0.15$ for $\nu = 3$ and $0.003 \leq \alpha \leq 0.15$ for $\nu \geq 4$ (Akahira, Sato and Torigoe 1995). In addition, from this result, we can obtain the confidence interval for the noncentrality parameter δ of level $1-\alpha$.

$$\left(T_\nu(\delta) b_\nu - Z_{\frac{\alpha}{2}} \sqrt{1 + T_\nu^2(\delta)(1 - b_\nu^2)} + \right.$$
$$\left. \frac{(T_\nu(\delta))^3 \left(Z_{\frac{\alpha}{2}}^2 - 1 \right)}{24 (1 + T_\nu^2(\delta)(1 - b_\nu^2))} \left\{ \frac{1}{\nu^2} + \frac{1}{4\nu^3} \right\} + O\left(\nu^{-4}\right), \right.$$

$$T_\nu(\delta) b_\nu + Z_{\frac{\alpha}{2}} \sqrt{1 + T_\nu^2(\delta)(1 - b_\nu^2)} -$$
$$\left. \frac{(T_\nu(\delta))^3 \left(Z_{\frac{\alpha}{2}}^2 - 1 \right)}{24 (1 + T_\nu^2(\delta)(1 - b_\nu^2))} \left\{ \frac{1}{\nu^2} + \frac{1}{4\nu^3} \right\} + O\left(\nu^{-4}\right) \right).$$

□

If we only consider the above approximation up to the order $O\left(n^{-1}\right)$, a very neat normality result can be obtained (Abramowitz and Stegun 1964 p949 and Walck 2007 p118).

$$\widehat{SR}\left(1 - \frac{1}{4(n-1)}\right) \xrightarrow{d} N\left(SR, \frac{1}{n} + \frac{\widehat{SR}^2}{2(n-1)}\right). \tag{1.2.6}$$

1.2.4 Statistical Properties of Estimated Sharpe Ratio Under Autoregressive Return

The simplest relaxation of the IID assumption of the return is to assume the time series of returns has a autocorrelation. Lo (2002) proposed that under non-IID returns people can obtain the distribution of estimated Sharpe ratio by using MLE plus Delta method. Specifically,

$$\widehat{SR} = \psi\left(\hat{\boldsymbol{\theta}}\right) \xrightarrow{d} N\left(\psi(\boldsymbol{\theta}), \left(\left.\frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}}\right)' \mathbf{I}^{-1}(\boldsymbol{\theta}) \left(\left.\frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}}\right)\right),$$

Or

$$\widehat{SR} = \psi\left(\hat{\boldsymbol{\theta}}\right) \xrightarrow{d} N\left(\psi(\boldsymbol{\theta}), \left(\left.\frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}}\right)' \mathbf{j}^{-1}\left(\hat{\boldsymbol{\theta}}\right) \left(\left.\frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}}\right)\right).$$

In addition, Van Belle (2002) noted a special rule of thumb, under formulation of AR(1) with ρ being the autocorrelation of the series of returns and $\mu = r_f$, the noncentral t statistic becomes a central t statistic; and we

have

$$\sqrt{n}\widehat{SR} = t_{n-1} \xrightarrow{d} N\left(0, \frac{1+\rho}{1-\rho}\right) .$$

In this Chapter, we provide a general review on statistical properties of estimated Sharpe ratio. Note that all the asymptotic distributions for estimated Sharpe ratio introduced here are first order or second order results, meaning that they will need a relatively large sample to make the inference results accurate. In Chapter 3 and 4, we will set these results of Chapter 1 as a reference group to compare with our proposed third order methodology. But before that, we will detail the mechanics of the likelihood-based third-order methods in the next Chapter.

Chapter 2

Likelihood-based Statistical Inference Methods

2.1 Introduction

The following is the notation used throughout this thesis:

- Upper case letters, for example, X and Y are scalar random variables.
- Lower case letters, for example, x and y are scalar realizations.
- Bold letters or symbols, for example, $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{y}$ or $\boldsymbol{\theta}$, are matrices or vectors. In addition, all of the vectors are column vectors.

In particular, we let $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ be a n-dimensional vector of random variables with probability density function or pdf $f(\cdot; \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$ is a p-dimensional vector of parameters. Nuisance parameters $\boldsymbol{\lambda}(\boldsymbol{\theta})$ arise in a variety of settings, and typically they are included to make the model more realistic for the application of interest. The goal of statistical inference is to draw conclusions about the parameter of interest, $\psi(\boldsymbol{\theta})$, based on an observed sample vector $\mathbf{y} = (y_1, \ldots, y_n)'$. Furthermore, throughout this dissertation, we assume that:

1. $\dim(\psi(\boldsymbol{\theta})) = 1$,
2. $p < n$.

2.2 Dimension Reduction-Sufficiency and Ancillarity

The process of reducing the dimension of data without loss of information is referred to as dimension reduction. In this section, we review some results on dimension reduction by means of sufficiency and ancillarity. Sufficiency and ancillarity reduction are very useful in constructing marginal and conditional distributions which will depend only on the parameter of interest from the original model, and then these distributions can be used to draw inference about parameters of interest.

2.2.1 Sufficiency

2.2.1.1 Definition

Fisher (1922) introduced sufficiency as a method to reduce the dimension of a statistical model. It was further developed by Kalbfleisch (1975), Huzurbazar (1976), Cox and Hinkley (1979), Fraser (1979) and others.

Statistic, Sufficient Statistic and Minimal Sufficient Statistic are defined as follows. A **statistic** is a function of sample data that does not depend on any unknown parameters and the probability distribution of the statistic is called the sampling distribution of the statistic. In addition, let $\mathbf{Y}$ be a random vector whose distribution depends on the parameter $\boldsymbol{\theta}$. A statistic vector $\mathbf{S} = \mathbf{S}(\mathbf{Y})$ ($\dim(\mathbf{S}) = \dim(\boldsymbol{\theta}) = p$) is said to be **sufficient** for $\boldsymbol{\theta}$ if, for each s, the conditional distribution of $\mathbf{Y}$ given $\mathbf{S}(\mathbf{Y}) = \mathrm{s}$ does not depend on $\boldsymbol{\theta}$. Also, it is easy to see that if $\mathrm{g}(\cdot)$ is a one to one function ($\mathbb{R}^p \to \mathbb{R}^p$) and $\mathbf{S}$ is a sufficient statistic, then $\mathrm{g}(\mathbf{S})$ is also a sufficient statistic. Finally, a sufficient statistic $\mathbf{S}(\mathbf{Y})$ is a **minimal sufficient statistic** if $\mathbf{S}(\mathbf{Y})$ is a function of $\mathbf{S}^*(\mathbf{Y})$ for any other sufficient statistic $\mathbf{S}^*(\mathbf{Y})$.

To motivate the mathematical definition, we consider the following example. There are two people A and B. A knows the entire random sample y while B only knows the value of sufficient statistic $\mathbf{S}(\mathrm{y}) = \mathrm{s}$. Since the conditional distribution of $\mathbf{Y}$ given $\mathbf{S}$ does not depend on $\boldsymbol{\theta}$, people who know the value of $\mathbf{S}$ would also know this conditional distribution. And thus B can use his computer to generate a new random sample y^* from this conditional distribution and his new random sample has the same distribution as a random sample drawn from the population with unknown value of $\boldsymbol{\theta}$. Finally B can use his random sample y^* to compute whatever A computes using his random sample y and B can do, on average, just as good a job of estimating the unknown parameter $\boldsymbol{\theta}$ as A. Or intuitively, all of the information needed for inference from the data about ψ is contained in the statistic $\mathbf{S}$ in this reduction of data.

Hogg, Tanis and Rao (1977) suggest the application of **Factorization Theorem** to check if a statistic is sufficient; Specifically, a statistic vector $\mathbf{S}(\mathbf{Y})$ is sufficient if and only if the density function can be factored as

follows:

$$f(\mathbf{y}|\boldsymbol{\theta}) = u(\mathbf{y})v(\mathbf{S}(\mathbf{y}), \boldsymbol{\theta}) \ , \tag{2.2.1}$$

where u and v are non-negative functions.

Example. We consider a Normal distribution. Let $Y_1, \ldots, Y_n$ be a random sample independent identically distributed or IID as $N(\mu, \sigma^2)$. In this case $\boldsymbol{\theta} = \left(\mu, \sigma^2\right)$. The joint density of the sample is

$$\begin{aligned}
& f(y_1, \ldots, y_n|\mu, \sigma^2) \\
= \ & (2\pi)^{-\frac{n}{2}} \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n} (y_i - \mu)^2}{2\sigma^2} \right) \qquad (2.2.2) \\
= \ & (2\pi)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2} + \frac{\mu \sum_{i=1}^{n} y_i}{\sigma^2} - \frac{n\mu^2}{2\sigma^2} - \frac{n \ln \sigma^2}{2} \right) \\
= \ & (2\pi)^{-\frac{n}{2}} \exp\left(\begin{pmatrix} \frac{\mu}{\sigma^2} & -\frac{1}{2\sigma^2} \end{pmatrix} \cdot \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} y_i^2 \end{pmatrix} - \frac{n\mu^2}{2\sigma^2} - \frac{n \ln \sigma^2}{2} \right) \\
= \ & (2\pi)^{-\frac{n}{2}} \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})' \mathbf{S}(\mathbf{y}) - A(\boldsymbol{\theta}) \right) \ . \qquad (2.2.3)
\end{aligned}$$

From (2.2.3) and Factorization Theorem (2.2.1), we can find that

$$u(\mathbf{y}) = (2\pi)^{-\frac{n}{2}} \ ,$$

$$v(\mathbf{S}(\mathbf{y}), \boldsymbol{\theta}) = \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})' \mathbf{S}(\mathbf{y}) - A(\boldsymbol{\theta}) \right) \ .$$

The sufficient statistic is

$$\mathbf{S}(\mathbf{Y}) = \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} Y_i^2 \end{pmatrix} \ . \tag{2.2.4}$$

The natural parameter or the canonical parameter is

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix} \ . \tag{2.2.5}$$

Log-partition function is $A(\boldsymbol{\theta}) = \frac{n\mu^2}{2\sigma^2} + \frac{n \ln \sigma^2}{2}$ or $A(\boldsymbol{\eta}) =$n$(-\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-2\eta_2))$.

An alternative common sufficient statistic can be achieved by one-to-one transformation from (2.2.4):

$$\mathbf{S}^*(\mathbf{Y}) = \begin{pmatrix} \frac{\sum_{i=1}^{n} Y_i}{n} \\ \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 \end{pmatrix}.$$

According to Fraser (1976), the first sufficient statistic in random variable form $\bar{Y} = \left(\sum_{i=1}^{n} Y_i\right)/n$ is normally distributed with mean μ and variance σ^2/n, that is $N(\mu, \sigma^2/n)$. Then $\left(\bar{Y} - \mu\right) / \left(SE/\sqrt{n}\right)$ is distributed as the Student t distribution with $(n-1)$ degrees of freedom, where $SE^2 = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 / (n-1)$. Since the distribution of this statistic depends only on one of the parameters of interest $\psi_1(\boldsymbol{\theta}) = \mu$, it then can be used for inference concerning μ. For example, the classical 95% confidence interval for μ is $\bar{y} \pm t^* \frac{se}{\sqrt{n}}$ with $P(t_{n-1} > t^*) = 0.025$. On the other hand, the second sufficient statistic $\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$ is distributed as $\sigma^2 \chi^2_{n-1}$, where χ^2_{n-1} stands for a Chi-squared distribution with degree of freedom $(n-1)$. It depends only on the other parameter of interest $\psi_2(\boldsymbol{\theta}) = \sigma^2$ and thus can be used for inference about σ^2.

Hence we successfully reduce the dimension of data from n to 2. In other words, rather than working with the original $Y_1, \ldots, Y_n$, it is sufficient to work with $\mathbf{S}(\mathbf{Y})$ or $\mathbf{S}^*(\mathbf{Y})$ with no loss of information about $\boldsymbol{\theta}$. □

2.2.1.2 Sufficient Statistic and Conditioning

Cox (1988) identifies at least four interrelated roles for conditioning in inference. First, to make probability calculations relevant to the data under study; second, to recover information lost in reducing the dimension of the problem; third, to eliminate nuisance parameters; and finally, to enable computation of accurate approximations to densities. Making probability calculations relevant and recovering information are usually associated with conditioning on ancillary or approximately ancillary statistics; eliminating nuisance parameters is usually associated with conditioning on sufficient or approximately sufficient statistics. These two types of conditioning are most transparent in transformation family models and exponential family modes, respectively.

We start with an example to discuss sufficient statistic and conditioning.

Example. Let Y_1 and Y_2 be IID as a normal distribution with mean θ and variance unity, $N(\theta, 1)$. Also let $S = Y_1 + Y_2$ and $T = Y_1 - Y_2$ be a one-to-one transformation from (Y_1, Y_2) to (S, T). Thus, $S \sim N(2\theta, 2)$ and $T \sim N(0, 2)$. Note that the distribution of T is free of θ and S and T are independent. Then the conditional distribution of T given S is free of θ:

$$f_T(t) = f_{T|S}(t|s) = \frac{f_{S,T}(s, t; \theta)}{f_S(s; \theta)} = \frac{f_{\mathbf{Y}}(\mathbf{y}; \theta)}{f_S(s; \theta)}.$$

By Factorization Theorem, we know statistic S is sufficient and the dimen-

sion of the variable is reduced from 2 to 1 by the marginalization in going from $\mathbf{Y}$ to S. Thus the model for the original sample Y_1 and Y_2 is replaced by the marginal model for the new variable S and inference about θ can be obtained from the marginal density of S. □

To extend the preceding example, we generalize a type of sufficiency reduction which is useful in reducing the dimension of the initial variable $\mathbf{Y}$ to p, the dimension of the sufficient statistic $\mathbf{S}$. Suppose there exists a one-to-one transformation from $\mathbf{Y}$ to $(\mathbf{S}, \mathbf{T})$ with the Jacobian of the transformation $\mathbf{J}$ such that

$$f(\mathbf{y};\boldsymbol{\theta}) = f(\mathbf{s};\boldsymbol{\theta}) f(\mathbf{t}|\mathbf{s}) \left|\mathbf{J}\right| \ , \tag{2.2.6}$$

then $\mathbf{S} = \mathbf{S}(\mathbf{Y})$ is sufficient for $\boldsymbol{\theta}$ and the marginal density of $\mathbf{S}$, $f(\mathbf{s};\boldsymbol{\theta})$, is an appropriate basis for inference about $\boldsymbol{\theta}$. However, the conditional density $f(\mathbf{t}|\mathbf{s})$ does have a role to play in inference, but not in inference for $\boldsymbol{\theta}$. As is suggested in Cox and Hinkley (1979), the conditional density is useful for model checking; in particular because it does not depend on which value of $\boldsymbol{\theta}$ generated the data $\mathbf{Y}$. From the model-checking point of view, $\boldsymbol{\theta}$ is a nuisance parameter which is eliminated in the conditional density.

The most well-known class of models which allows a sufficiency reduction of the type (2.2.6) is the **family of linear exponential models**.

$$f(\mathbf{y};\boldsymbol{\theta}) = \exp\left(\boldsymbol{\eta}'(\boldsymbol{\theta})\mathbf{s}(\mathbf{y}) - A(\boldsymbol{\theta})\right) u(\mathbf{y}) \ , \tag{2.2.7}$$

where $\mathbf{S} = \mathbf{S}(\mathbf{Y})$ is the minimal sufficient statistic with same dimension p as canonical parameter $\boldsymbol{\eta}(\boldsymbol{\theta})$. According to the Pitman–Koopman–Darmois Theorem (Pitman, 1936 and etc), among families of probability distributions whose domain does not vary with the parameter being estimated, only in exponential families is there a sufficient statistic whose dimension remains bounded as sample size increases. Note that the form is non-unique, since $\boldsymbol{\eta}(\boldsymbol{\theta})$ can be multiplied by any nonzero constant, provided that $\mathbf{S}(\mathbf{Y})$ is multiplied by that constant's reciprocal, or a constant c can be added to $\boldsymbol{\eta}(\boldsymbol{\theta})$ and $u(\mathbf{y})$ multiplied by $e^{-c\mathbf{s}(\mathbf{y})}$ to offset it.

Note that $A(\boldsymbol{\theta})$ can always be written as functions of $\boldsymbol{\eta}$, regardless of the form of the transformation that generates $\boldsymbol{\eta}$ from $\boldsymbol{\theta}$. It is so even when $\boldsymbol{\eta}(\boldsymbol{\theta})$ is not a one-to-one function and cannot be inverted, i.e. two or more different values of $\boldsymbol{\theta}$ map to the same value of $\boldsymbol{\eta}(\boldsymbol{\theta})$. Thus, by defining a transformed natural parameter $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\theta})$ we can always write (2.2.7) as $f(\mathbf{y};\boldsymbol{\eta}) = \exp\left(\boldsymbol{\eta}'\mathbf{s}(\mathbf{y}) - A(\boldsymbol{\eta})\right) u(\mathbf{y})$. Or, more generally, let $\boldsymbol{\theta}$ represent $\boldsymbol{\eta}$, written as $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$, and in this case the exponential family is said to be in canonical form or in natural form:

$$f(\mathbf{y};\boldsymbol{\theta}) = \exp\left(\boldsymbol{\theta}'\mathbf{s}(\mathbf{y}) - A(\boldsymbol{\theta})\right) u(\mathbf{y}) \ . \tag{2.2.8}$$

Following the basic setup at the beginning of this chapter, ψ can be an interest canonical parameter in (2.2.7) or (2.2.8) and then the corresponding

exponential model is sometimes expressed as

$$f(\mathbf{y};\boldsymbol{\theta}) = \exp\left(\psi s_1 + \boldsymbol{\lambda}'\mathbf{s}_2 - A(\boldsymbol{\theta})\right) u(\mathbf{y}) \,. \tag{2.2.9}$$

The distribution of $S_1|\mathbf{S}_2 = \mathbf{s}_2$ depends only on the interest parameter ψ and accordingly the conditional distribution of of S_1 given $\mathbf{S}_2 = \mathbf{s}_2$ is an appropriate distribution for inference concerning ψ free of the nuisance parameter $\boldsymbol{\lambda}$. The extension of this point will be presented in the example at subsection (2.2.3.2).

2.2.2 Ancillarity

2.2.2.1 Definition.

We proceed in a different direction and consider statistics whose distribution is free of the parameter. Formally, a statistic $\mathbf{T}(\mathbf{Y})$ whose distribution does not depend on the parameter $\boldsymbol{\theta}$ is called an **ancillary** statistic and a statistic $\mathbf{T}(\mathbf{Y})$ is **maximal ancillary** if every other ancillary statistic is a function of $\mathbf{T}(\mathbf{Y})$. In addition, **approximate ancillary** means that the distribution of $\mathbf{T}(\mathbf{Y})$ depends on $\boldsymbol{\theta}$ only in terms of $O\left(n^{-1}\right)$ or higher, for $\boldsymbol{\theta}$ within $O\left(n^{-\frac{1}{2}}\right)$ of its true value.

We can see from the definition that ancillary statistic is a pivotal quantity as well as a statistic and ancillary statistics can be used to construct prediction intervals. However, an ancillary statistic may not exist and a general method for constructing an ancillary statistic does not exist either. If one exists at some special cases, an ancillary statistic may not be unique.

Example. Let $X_1, \ldots, X_n$ be IID $N(\mu, 1)$. In this case, let $\psi = \mu$. The following statistical measures of dispersion of the sample are all ancillary statistics, because their sampling distributions does not change as location μ changes:

(1) Range: $\max\{X_1, \ldots, X_n\} - \min\{X_1, \ldots, X_n\}$

(2) Interquartilde Range: $Q_3 - Q_1$

(3) Sample Variance: $\hat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/n$

Conversely, let $X_1, \ldots, X_n$ be IID $N(1, \sigma^2)$, and here $\psi = \sigma^2$. In this case the sample mean $\bar{X}$ is, however, not an ancillary statistic of the variance, as the sampling distribution of $\bar{X}$ is $N(1, \frac{\sigma^2}{n})$, which does depend on σ^2, and this measure of location depends on dispersion. □

2.2.2.2 Ancillary Statistic and Conditioning

Ideally, one would like to have the dimension of the minimal sufficient statistic to be equal to p which is the dimension of the parameter $\boldsymbol{\theta}$. However, Cox and Hinkley (1979) pointed out that it is common to have the

dimension of minimal sufficient statistic larger than p and thus a reduction method is needed to obtain inference on $\boldsymbol{\theta}$. In particular, one can make initial reduction by sufficiency (2.2.6) or (2.2.1), then apply ancillary methods partitioning the minimal sufficient statistic $\mathbf{S}$ into $\begin{pmatrix} \mathbf{M} \\ \mathbf{A} \end{pmatrix}$ such that the marginal distribution of $\mathbf{A}$ does not depend on $\boldsymbol{\theta}$. Hence, we can factorize $f(\mathbf{s};\boldsymbol{\theta})$, the density of the minimal sufficient statistic $\mathbf{S}$, to $f(\mathbf{s};\boldsymbol{\theta}) \propto f(\mathbf{m}|\mathbf{a};\boldsymbol{\theta})f(\mathbf{a})$. Substitute this result to (2.2.6) and we can obtain

$$f(\mathbf{y};\boldsymbol{\theta}) = f(\mathbf{s};\boldsymbol{\theta})f(\mathbf{t}|\mathbf{s})\,|\mathbf{J}| \propto f(\mathbf{s};\boldsymbol{\theta}) \propto f(\mathbf{m}|\mathbf{a};\boldsymbol{\theta})f(\mathbf{a})\,. \tag{2.2.10}$$

In (2.2.10) $\mathbf{A} = \mathbf{A}(\mathbf{Y})$ is said to be ancillary for $\boldsymbol{\theta}$ and the conditional density of $\mathbf{M}$ given $\mathbf{A}$ entails no information loss about $\boldsymbol{\theta}$ and thus can be used for inference concerning $\boldsymbol{\theta}$. In particular, if one takes $\mathbf{M}$ to be maximum likelihood estimator $\hat{\boldsymbol{\theta}}$, which in general will not be sufficient, then one can ask for an ancillary complement. Since the requirement that $\mathbf{A}$ be ancillary ensures that there is little information about $\boldsymbol{\theta}$ in the marginal density for $\mathbf{A}$, the conditional distribution of $\hat{\boldsymbol{\theta}}$ given $\mathbf{A}$ is expected to provide a good inference for $\boldsymbol{\theta}$. Intuitively, an ancillary complement add the missing information of $\mathbf{M}$ without duplicating any. We will see this point again when introducing p^* formula at section (2.5).

Fisher (1934) introduced inference conditional on ancillaries as a method to reduce the size of the sample space and yet retain all the relevant information in the original sample; he claimed that the argument often advanced for using this method for inference about $\boldsymbol{\theta}$ is either that the conditional density gives more precise inference for $\boldsymbol{\theta}$ or that the subset of the sample space defined by fixing the value of $\mathbf{A}$ is the relevant subset for inference about $\boldsymbol{\theta}$. For further discussion see Kass(1989) and Dawid(1991) .

However, the idea of using the conditional distribution in (2.2.10) for inference has not been as widely accepted as using the marginal density in (2.2.6). According to Reid (1995), this may be because the partition in (2.2.10) is not typically unique, whereas that in (2.2.6) is essentially unique.

The main class of models which allows an ancillary reduction of type (2.2.10) is the **transformation family models**, that is, models generated by a group of transformations acting on the sample space. For the transformation model exact ancillary statistics exist and are sometimes referred to as the configuration of the sample.

Example. (1) Denoting the base model by $f_0(\cdot)$, the transformation model is $f(\mathbf{y};\boldsymbol{\theta}) = f_0(g_\theta \mathbf{y})$, where g_θ is a member of a group of transformations indexed by θ that leaves the original sample space unchanged. For example, in a location-scale model, $g_\theta Y = \frac{Y-\theta_1}{\theta_2}$. In a sample of size n from a transformation model a partition of the form (2.2.10) obtains with $\mathbf{M} = \mathbf{M}(\mathbf{Y}) = \hat{\boldsymbol{\theta}}$, the maximum likelihood estimate, and $\mathbf{A} = \mathbf{A}(\mathbf{Y}) = g_{\hat{\theta}}\mathbf{Y}$, the maximal invariant for the group. Again $\mathbf{M}$ has the same dimension as the parameter

$\boldsymbol{\theta}$. In a sample of size n from a location model, $\mathbf{A}(\mathbf{Y})$=$\Big(Y_1 - \hat{\theta}, \ldots, Y_n - \hat{\theta}\Big)$; from a location-scale model, $\mathbf{A}(\mathbf{Y}) = \Big(\frac{Y_1-\hat{\theta}_1}{\hat{\theta}_2}, \ldots, \frac{Y_n-\hat{\theta}_1}{\hat{\theta}_2}\Big)$.

(2) Let $(X_1, \ldots, X_n)$ be a random sample from a statistical model $\frac{1}{\sigma} f(\frac{x-\mu}{\sigma})$, $\boldsymbol{\theta} = \big(\mu, \sigma^2\big)$, $S_1 = \bar{X}$ and $S_2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$ then (S_1, S_2) is a sufficient statistic given the configuration statistic (Fisher, 1934). Then $\mathbf{A} = \Big(\frac{X_1-\hat{\mu}}{\hat{\sigma}}, \ldots, \frac{X_n-\hat{\mu}}{\hat{\sigma}}\Big)$ which is in turn ancillary; the ancillary defines the orbit on which the observations reside. Position on an orbit is determined by the MLE $\big(\hat{\mu}, \hat{\sigma}^2\big)$. Accordingly any sample point can be represented by the pair, the MLE and the configuration. For detailed discussion see Fraser (1968). □

For a continuous statistical model $f(\mathbf{y}; \boldsymbol{\theta})$ with asymptotic properties, the current methods of constructing ancillaries for third order analysis include the tangent location model of Fraser (1964) and the cumulant method of McCullagh (1987). The cumulant method uses the cumulant of the log-likelihood derivative to obtain third order ancillaries(McCullagh, 1987) and the tangent location method constructs tangent directions for first derivative ancillaries at the data point $\mathbf{Y}^0$. The analysis in Fraser and Reid (1995) shows that first order derivative ancillaries can be adjusted to give second order ancillaries and no further upgrading is required for third order inference: these second order ancillaries can however be upgraded to third order ancillaries (Skovgaard, 1986).

2.2.3 Nuisance Parameters

We assume if the model $f(\mathbf{y}; \boldsymbol{\theta})$ satisfies (2.2.6) or (2.2.10), then the marginal density of sufficient statistic or conditional density given ancillary statistic will be used as the basis for inference and the dimension of the initial problem is reduced to that of S or that of M. In exponential or transformation models, the dimensions of S or M and $\boldsymbol{\theta}$ are the same.

For models with nuisance parameters, various generalizations of factorizations (2.2.6) and (2.2.10) are available according to Reid (1995). We summarize them into three types.

2.2.3.1 Extension Type One

The simplest extension is

$$f\left(\mathbf{s}; \psi, \boldsymbol{\lambda}\right) = f(s_1|\mathbf{s}_2; \psi) f(\mathbf{s}_2; \boldsymbol{\lambda}) \,. \tag{2.2.11}$$

In this setting, $\mathbf{S}_2$ is sufficient for $\boldsymbol{\lambda}$ in a general sense as in (2.2.6) and the marginal density for $\mathbf{S}_2$ can then be used for inference concerning the interest parameter $\boldsymbol{\lambda}$ with the nuisance parameter ψ eliminated by marginalization. On the other hand if interest lies in obtaining inference about the interest parameter ψ, since $\mathbf{S}_2$ is ancillary for ψ as in (2.2.10), the conditional

density of $S_1|\mathbf{S}_2$ would be an appropriate choice. For this case, conditioning eliminates the nuisance parameter $\boldsymbol{\lambda}$.

Many definitions of sufficiency and ancillarity in the presence of nuisance parameters require the parameters of the model to split into component densities in the manner of (2.2.11) (Fraser, 1956; Basu, 1977; Cox and Hinkley, 1979). Cox and Hinkley (1979) refer to S_1 as "conditionally sufficient for ψ". Barndorff-Nielsen (2014) uses the terminology "S_1 is S-sufficient for ψ".

2.2.3.2 Extension Type Two

Very few models with nuisance parameters admit a factorization of the form (2.2.11). One generalization of it is

$$f(\mathbf{s};\boldsymbol{\theta}) = f(s_1|\mathbf{s}_2;\psi)f(\mathbf{s}_2;\psi,\boldsymbol{\lambda}) \ . \tag{2.2.12}$$

In this case, $\mathbf{S}_2$ is no longer ancillary for ψ, but it is sufficient for $\boldsymbol{\lambda}$ in the sense that the nuisance parameter $\boldsymbol{\lambda}$ has been eliminated in the conditional distribution. $f(\mathbf{s}_2;\psi,\boldsymbol{\lambda})$ or $f(\mathbf{s};\boldsymbol{\theta})$ can be used for testing $\boldsymbol{\lambda}$ if plausible values of ψ can be constructed from the conditional density of S_1 given $\mathbf{S}_2$.

One motivation for using $f(s_1|\mathbf{s}_2;\psi)$ for inference about ψ is merely pragmatic: we can do this without specifying a value for the unknown parameter $\boldsymbol{\lambda}$. A more theoretical motivation is that, in testing a hypothesis about ψ with $\boldsymbol{\lambda}$ unspecified, any test having a type I error that does not depend on $\boldsymbol{\lambda}$ must be constructed from the conditional distribution, at least if $\mathbf{S}_2$ is complete (Lehmann, 1986). However, there is potentially information about ψ in the marginal density of $\mathbf{S}_2$ as is indicated explicitly in the notation.

A systematic investigation into ways of quantifying the information in such marginal or conditional densities is given in Jorgensen (1994) and Barndorff-Nielsen (2014) .

Example. .

(1) The most common models admitting a factorization of the form (2.2.12) are exponential family linear models, where $f(\mathbf{s};\boldsymbol{\theta}) = \exp(\psi s_1 + \boldsymbol{\lambda}'\mathbf{s}_2 - A(\psi,\boldsymbol{\lambda}) - d(s_1,\mathbf{s}_2))$. It is easy to show that

$$f(s_1|\mathbf{s}_2;\psi) = \exp(\psi s_1 - A_2(\psi) - d_2(s_1)) \ , \tag{2.2.13}$$

and that the marginal distribution of $\mathbf{S}_2$ depends on $(\psi,\boldsymbol{\lambda})$ (Lehmann, 1986, Ch.2). In (2.2.13) the functions $A_2(\psi)$ and $d_2(S_1)$ depend on $\mathbf{S}_2$ and are usually difficult to calculate. However, tests of hypotheses about ψ based on (2.2.13) have an unconditional optimality property: they are uniformly most powerful among the class of unbiased tests. Conditional inference based on (2.2.13) is discussed in detail in Lehmann (1986)) from this point of view. A third order approximation to the likelihood function from the conditional distribution (2.2.13) is introduced at later section in (2.5.6).

(2) The shape parameter of a gamma distribution is a component of the canonical parameter. Suppose we have a sample of size n from the density $f(y;\psi,\lambda) = \frac{1}{\Gamma(\psi)}\lambda^{\psi}y^{\psi-1}e^{-\lambda y}$. The minimal sufficient statistic is $(S_1,S_2) = (\sum \log Y_i, \sum Y_i)$ and the conditional density of S_1 given S_2 is given by (2.2.13), where versions of $A_2(\psi)$ and $d_2(S_1)$ are $\exp(A_2(\psi)) = \int \exp\left((\psi-1)s_1\right) h\left(s_1,s_2\right) ds_1$, $\exp\left(-d_2(s_1)\right) = \exp\left(-s_1\right) h\left(s_1,s_2\right)$, $h\left(s_1,s_2\right) = \exp\left(-d(s_1,s_2)\right) = \int_S dy_1 \cdots dy_n$ with $S = \{(y_1 \ldots y_n) : \sum y_i = s_2, \sum \log y_i = s_1\}$. □

2.2.3.3 Extension Type Three

A different generalization of (2.2.11) is the case where $\mathbf{S}_2$ is no longer sufficient for $\boldsymbol{\lambda}$, but is ancillary for ψ:

$$f\left(\mathbf{s};\psi,\boldsymbol{\lambda}\right) = f\left(s_1|\mathbf{s}_2;\psi,\boldsymbol{\lambda}\right) f\left(\mathbf{s}_2;\boldsymbol{\lambda}\right) \ . \tag{2.2.14}$$

By analogy with the situation in (2.2.12), one motivation for using marginal density of $\mathbf{S}_2$ is pragmatism; we can construct inference for $\boldsymbol{\lambda}$ without specifying any value for the nuisance parameter ψ.

If our interest is in the parameter ψ, then we will either use $f(\mathbf{s};\psi,\boldsymbol{\lambda})$ or $f\left(s_1|\mathbf{s}_2;\psi,\boldsymbol{\lambda}\right)$ for inference about ψ. If we use the conditional density, plausible values of $\boldsymbol{\lambda}$ might be obtained from the marginal density for $\mathbf{S}_2$.

This is a more direct generalization of the ancillarity definition (2.2.10): the distribution of the ancillary statistic depends on an additional parameter rather than being completely known as in (2.2.10).

Example. Suppose we have a sample of size n from $N\left(\mu;\sigma^2\right)$. $\bar{Y}$ and S^2 stand for the sample mean and variance, we have

$$f\left(\bar{y},s^2;\mu,\sigma^2\right) = f_1\left(\bar{y};\mu,\sigma^2\right) f_2\left(s^2;\sigma^2\right) \ ,$$

where f_1 is the $N\left(\mu,\sigma^2/n\right)$ density and f_2 is the $\sigma^2\chi^2_{n-1}$ density. Because $\bar{Y}$ and S^2 are independent, this is actually an application of both (2.2.12) and (2.2.14).

First of all, inference about σ^2 can be based on the marginal distribution of S^2 as in (2.2.14) S^2 is ancillary for μ or the conditional distribution of S^2 given $\bar{Y}$ as in (2.2.12) $\bar{Y}$ is sufficient for μ. In fact this example highlights the difficulty with the extended definitions of sufficiency and ancillarity. Although $\bar{Y}$ is sufficient for μ, in the sense of definition (2.2.12), it is not sufficient in our usual understanding of the definition based on (2.2.6); that is, inference for μ, cannot be constructed from $\bar{Y}$ alone. This is also the reason why in (2.2.12) $f(\mathbf{s}_2;\psi,\boldsymbol{\lambda})$ or $f(\mathbf{s};\boldsymbol{\theta})$ can be used for testing $\boldsymbol{\lambda}$ if plausible values of ψ can be constructed from the conditional density of S_1 given $\mathbf{S}_2$, as well as in (2.2.14) we will either use $f(\mathbf{s};\psi,\boldsymbol{\lambda})$ or $f\left(s_1|\mathbf{s}_2;\psi,\boldsymbol{\lambda}\right)$ for inference about ψ if plausible values of $\boldsymbol{\lambda}$ might be obtained from the marginal density for $\mathbf{S}_2$.

On the other hand, inference for μ is constructed from the t-statistic $\sqrt{n}\left(\bar{Y}-\mu\right)/S$. The t-test can be derived by formal considerations related to (2.2.12) and (2.2.14). One derivation is via the construction of similar tests for the ratio of canonical parameters in the exponential family (Lehmann, 1986) and the other is via construction of an invariant test (Lehmann, 1986). Need to mention that t-statistic is the most well-known example of a pivotal statistic, that is, a function of the data and parameters whose distribution is known exactly. The development of inference based on pivotal statistics proceeded somewhat separately from that of conditional inference, although recent work emphasizes the connections between them.

Finally, the normal distribution is both an exponential family and a transformation family, which is why arguments based on sufficiency or ancillarity lead to the same result.

Corresponding to the example (1) in (2.2.3.2), in transformation models tests based on the marginal distribution of the ancillary statistic (the maximal invariant) also have an unconditional optimality property: they are uniformly most powerful among the class of invariant tests. This is the point of view from which the t-test is derived in Lehmann (1986). □

For further consideration of various definitions of sufficiency and ancillarity in the presence of nuisance parameters see (Fraser, 1956; Basu, 1977; Cox and Hinkley, 1979; Reid, 1996; Lindsey, 1996; Barndorff-Nielsen, 2014).

2.3 First Order Approximation

In statistical inference, people frequently encounter distributional problems that either have no exact solutions or have solutions so complicated that they cannot be used directly. Such situations are frequently addressed by asymptotic statistical theory. In this section, we will review some standard first-order likelihood-based methods. In the later sections, we will improve these methods to achieve a higher-order of accuracy.

The likelihood function of $\boldsymbol{\theta}$ from the observed response value $\mathbf{y}$ is defined as proportional to the sampling density, $\mathcal{L}\left(\boldsymbol{\theta};\mathbf{y}\right)=c\cdot\prod_{i=1}^{n}f\left(y_i;\boldsymbol{\theta}\right)$ where $c=c\left(\mathbf{y}\right)\in(0,+\infty)$ is an arbitrary constant; and the **log-likelihood function** is defined as:

$$\begin{aligned}\ell\left(\boldsymbol{\theta}\right) &= \ell\left(\boldsymbol{\theta};\mathbf{y}\right)=\ell\left(\psi,\boldsymbol{\lambda};\mathbf{y}\right)=\log\left(\mathcal{L}\left(\boldsymbol{\theta};\mathbf{y}\right)\right)\\ &= a+\log\left(\prod_{i=1}^{n}f\left(y_i;\boldsymbol{\theta}\right)\right)=a+\sum_{i=1}^{n}\log f\left(y_i;\boldsymbol{\theta}\right)=\sum_{i=1}^{n}\ell\left(\boldsymbol{\theta};y_i\right)\end{aligned}\tag{2.3.1}$$

where $a\in(-\infty,+\infty)$ is an arbitrary constant. We can see that the likelihood is a function of the parameter as determined by the data. It can be

viewed as summarizing all the information in the data about the parameter.

The followings represent the notation that will be used throughout this dissertation.

- $f_{\boldsymbol{\theta}}(\mathbf{y};\boldsymbol{\theta}) = \frac{\partial f(\mathbf{y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ and $f_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\mathbf{y};\boldsymbol{\theta}) = \frac{\partial^2 f(\mathbf{y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}$;
- $\mathbf{s}(\boldsymbol{\theta}) = \ell_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is **the score function** and $\mathbf{s}\left(\hat{\theta}\right) = \mathbf{0}$ is **the estimating equation**;
- $\ell_{;\mathbf{y}'} = \frac{\partial \ell(\boldsymbol{\theta};\mathbf{y})}{\partial \mathbf{y}'} = \left(\frac{\partial \ell(\boldsymbol{\theta};\mathbf{y})}{\partial y_1}, \ldots, \frac{\partial \ell(\boldsymbol{\theta};\mathbf{y})}{\partial y_n}\right)$ is the **log likelihood gradient with respect to the variable** $\mathbf{y}$ and a second order derivative $\ell_{\boldsymbol{\theta};\mathbf{y}'} = \frac{\partial^2 \ell(\boldsymbol{\theta};\mathbf{y})}{\partial \mathbf{y}'\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial^2 \ell(\boldsymbol{\theta};\mathbf{y})}{\partial y_1 \partial \theta_1} & \cdots & \frac{\partial^2 \ell(\boldsymbol{\theta};\mathbf{y})}{\partial y_n \partial \theta_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \ell(\boldsymbol{\theta};\mathbf{y})}{\partial y_1 \partial \theta_p} & \cdots & \frac{\partial^2 \ell(\boldsymbol{\theta};\mathbf{y})}{\partial y_n \partial \theta_p} \end{pmatrix}$.
- If $\mathbf{V} = (\mathbf{V}_1, \ldots, \mathbf{V}_p)$ is a set of p linearly independent vectors and $\mathbf{V}_i \in \mathbb{R}^n$ then the **log likelihood gradient in the direction of** $\mathbf{V}$ is $1 \times p$ row vector $\ell_{;\mathbf{V}} = \frac{\partial \ell(\boldsymbol{\theta};\mathbf{y})}{\partial \mathbf{V}} = \ell_{;\mathbf{y}'} \cdot \mathbf{V} = \left(\sum_{k=1}^{n} \frac{\partial \ell(\boldsymbol{\theta};\mathbf{y})}{\partial y_k} v_{k1}, \ldots, \sum_{k=1}^{n} \frac{\partial \ell(\boldsymbol{\theta};\mathbf{y})}{\partial y_k} v_{kp}\right)$.
- $\mathbf{j}(\boldsymbol{\theta}) = \mathbf{j}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\boldsymbol{\theta}) = -\mathbf{s}_{\boldsymbol{\theta}'}(\boldsymbol{\theta}) = -\ell_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\boldsymbol{\theta}) = -\frac{\partial^2 \ell(\boldsymbol{\theta};\mathbf{y})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} = \begin{pmatrix} j_{\psi\psi}(\boldsymbol{\theta}) & \mathbf{j}_{\psi\boldsymbol{\lambda}}(\boldsymbol{\theta}) \\ \mathbf{j}_{\boldsymbol{\lambda}\psi}(\boldsymbol{\theta}) & \mathbf{j}_{\boldsymbol{\lambda}\boldsymbol{\lambda}'}(\boldsymbol{\theta}) \end{pmatrix}$ is the **observed Fisher full information matrix** and $\mathbf{I}(\boldsymbol{\theta}) = var[\mathbf{s}(\boldsymbol{\theta})]$ is the **expected Fisher full information matrix**. Under certain regularity conditions (Lehmann and Casella 1998) an alternative expression for the expected information is given by $\mathbf{I}(\boldsymbol{\theta}) = var[\mathbf{s}(\boldsymbol{\theta})] = E[\mathbf{s}(\boldsymbol{\theta})\mathbf{s}'(\boldsymbol{\theta})] = E[\mathbf{j}(\boldsymbol{\theta})] = \begin{pmatrix} I_{\psi\psi}(\boldsymbol{\theta}) & \mathbf{I}_{\psi\boldsymbol{\lambda}}(\boldsymbol{\theta}) \\ \mathbf{I}_{\boldsymbol{\lambda}\psi}(\boldsymbol{\theta}) & \mathbf{I}_{\boldsymbol{\lambda}\boldsymbol{\lambda}'}(\boldsymbol{\theta}) \end{pmatrix}$.
- $\mathbf{j}_{\boldsymbol{\lambda}\boldsymbol{\lambda}'}(\boldsymbol{\theta}) = -\ell_{\boldsymbol{\lambda}\boldsymbol{\lambda}'}(\boldsymbol{\theta}) = -\frac{\partial^2 \ell(\boldsymbol{\theta};\mathbf{y})}{\partial \boldsymbol{\lambda}\partial \boldsymbol{\lambda}'}$ is the observed information concerning the nuisance parameter $\boldsymbol{\lambda}$ for given ψ, and it is also called **the observed nuisance information matrix**;
- $\mathbf{I}^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} I^{\psi\psi}(\boldsymbol{\theta}) & \mathbf{I}^{\psi\boldsymbol{\lambda}}(\boldsymbol{\theta}) \\ \mathbf{I}^{\boldsymbol{\lambda}\psi}(\boldsymbol{\theta}) & \mathbf{I}^{\boldsymbol{\lambda}\boldsymbol{\lambda}'}(\boldsymbol{\theta}) \end{pmatrix}$ and $\mathbf{j}^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} j^{\psi\psi}(\boldsymbol{\theta}) & \mathbf{j}^{\psi\boldsymbol{\lambda}}(\boldsymbol{\theta}) \\ \mathbf{j}^{\boldsymbol{\lambda}\psi}(\boldsymbol{\theta}) & \mathbf{j}^{\boldsymbol{\lambda}\boldsymbol{\lambda}'}(\boldsymbol{\theta}) \end{pmatrix}$ are the inverse of the original block information; [1]

---

[1] From $\det\begin{pmatrix} \mathbf{A}_{n\times n} & \mathbf{B}_{n\times m} \\ \mathbf{C}_{m\times n} & \mathbf{D}_{m\times m} \end{pmatrix} = \det(\mathbf{D})\det\left(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)$ and $\begin{pmatrix} \mathbf{A}_{n\times n} & \mathbf{B}_{n\times m} \\ \mathbf{C}_{m\times n} & \mathbf{D}_{m\times m} \end{pmatrix}^{-1} = \begin{pmatrix} \left(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1} & -\mathbf{A}^{-1}\mathbf{B}\left(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\left(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1} & \left(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)^{-1} \end{pmatrix}$, we can obtain a result on the inverse of information matrix $\mathbf{I}^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} I^{\psi\psi} & \mathbf{I}^{\psi\boldsymbol{\lambda}} \\ \mathbf{I}^{\boldsymbol{\lambda}\psi} & \mathbf{I}^{\boldsymbol{\lambda}\boldsymbol{\lambda}'} \end{pmatrix} = \begin{pmatrix} \left(I_{\psi\psi} - \mathbf{I}_{\psi\boldsymbol{\lambda}}\mathbf{I}^{-1}_{\boldsymbol{\lambda}\boldsymbol{\lambda}'}\mathbf{I}_{\boldsymbol{\lambda}\psi}\right)^{-1} = \left(\frac{|\mathbf{I}(\boldsymbol{\theta})|}{|\mathbf{I}_{\boldsymbol{\lambda}\boldsymbol{\lambda}'}|}\right)^{-1} & -I^{-1}_{\psi\psi}\mathbf{I}_{\psi\boldsymbol{\lambda}}\left(\mathbf{I}_{\boldsymbol{\lambda}\boldsymbol{\lambda}'} - \mathbf{I}_{\boldsymbol{\lambda}\psi}I^{-1}_{\psi\psi}\mathbf{I}_{\psi\boldsymbol{\lambda}}\right)^{-1} \\ -\mathbf{I}^{-1}_{\boldsymbol{\lambda}\boldsymbol{\lambda}'}\mathbf{I}_{\boldsymbol{\lambda}\psi}\left(I_{\psi\psi} - \mathbf{I}_{\psi\boldsymbol{\lambda}}\mathbf{I}^{-1}_{\boldsymbol{\lambda}\boldsymbol{\lambda}'}\mathbf{I}_{\boldsymbol{\lambda}\psi}\right)^{-1} & \left(\mathbf{I}_{\boldsymbol{\lambda}\boldsymbol{\lambda}'} - \mathbf{I}_{\boldsymbol{\lambda}\psi}I^{-1}_{\psi\psi}\mathbf{I}_{\psi\boldsymbol{\lambda}}\right)^{-1} \end{pmatrix}$

- $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative distribution function and probability density function for the standard normal distribution, respectively.

Moreover, throughout this dissertation with $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the following regularity conditions are assumed to hold:

- $f(\mathbf{y}; \boldsymbol{\theta}) > 0$ is twice continuously differentiable in $\boldsymbol{\theta}$ in a neighborhood $\mathbb{N}$ of $\boldsymbol{\theta}$;
- $\int \sup_{\boldsymbol{\theta} \in \mathbb{N}} |f_{\boldsymbol{\theta}}(\mathbf{y}; \boldsymbol{\theta})| d\mathbf{y} < \infty$ and $\int \sup_{\boldsymbol{\theta} \in \mathbb{N}} |f_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\mathbf{y}; \boldsymbol{\theta})| d\mathbf{y} < \infty$;
- $E\left[\ell_{\boldsymbol{\theta}}(\mathbf{y}; \boldsymbol{\theta}) \ell'_{\boldsymbol{\theta}}(\mathbf{y}; \boldsymbol{\theta})\right]$ exists and is nonsingular;
- $\int \sup_{\boldsymbol{\theta} \in \mathbb{N}} |\ell_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\mathbf{y}; \boldsymbol{\theta})| d\mathbf{y} < \infty$.

Likelihood analysis typically involves two types of maximum likelihood estimation: unconstrained maximum likelihood estimation and constrained maximum likelihood estimation.

2.3.1 Unconstrained Maximum Likelihood Estimation

The (unconstrained or overall) maximum likelihood estimation aims to solve the (unconstrained or overall) **maximum likelihood estimator** or MLE $\hat{\boldsymbol{\theta}} = \left(\hat{\psi}, \hat{\boldsymbol{\lambda}}'\right)' = \hat{\boldsymbol{\theta}}(\mathbf{y}) = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{y})$. $\hat{\boldsymbol{\theta}}$ is usually obtained from solving the first order condition, or equivalently the estimating equation: $\mathbf{s}\left(\hat{\boldsymbol{\theta}}\right) = \ell_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}\right) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \mathbf{0}$ and $\hat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$.

The study of parametric statistics based on likelihood function was initiated by Fisher (1922, 1925). In regular parametric models when the amount of information is large[2], first order asymptotic theory is available. For these models the Central Limit Theorem and the Law of Large Numbers provide access to a range of statistical procedures. In particular, for models with independent random variables, the score function, being a sum of independent components is asymptotically normal. Local linearization then relates the maximum likelihood estimate and likelihood ratio statistic to the score function. These three, score, maximum likelihood and likelihood ratio, provide important and powerful methods referred to as first order asymptotic theory. Specifically, under the regularity conditions stated above, first order approximations for testing H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ are obtained by Cox and Hinkley (1979):

1.

$$\mathbf{s}(\boldsymbol{\theta}) \xrightarrow{d} N_p(0, \mathbf{I}(\boldsymbol{\theta})) \ , \tag{2.3.2}$$

or equivalently the Rao Statistic or Lagrange multiplier statistic

[2]In more general contexts Central Limit Theorem type results hold as long as the quantity of information supplied by the sample tends to infinity. Increasing the sample size is just one simple way of increasing the quantity of information.

$$\mathbf{S}(\boldsymbol{\theta}) = \mathbf{s}'(\boldsymbol{\theta})\,\mathbf{I}^{-1}(\boldsymbol{\theta})\,\mathbf{s}(\boldsymbol{\theta}) \xrightarrow{d} \chi^2_p\,; \tag{2.3.3}$$

2.

$$\hat{\boldsymbol{\theta}} \xrightarrow{d} N_p\left(\boldsymbol{\theta}, \mathbf{I}^{-1}(\boldsymbol{\theta})\right)\,, \tag{2.3.4}$$

or equivalently the Wald statistic (see Wald 1941)

$$q^2(\boldsymbol{\theta}) = \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' \mathbf{I}(\boldsymbol{\theta}) \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} \chi^2_p\,; \tag{2.3.5}$$

3. The log likelihood ratio statistic or the Wilks statistic (see Wilks 1938)

$$R^2(\boldsymbol{\theta}) = 2\left[\ell\left(\hat{\boldsymbol{\theta}}\right) - \ell(\boldsymbol{\theta})\right] \xrightarrow{d} \chi^2_p\,; \tag{2.3.6}$$

4. By applying Taylor series expansion to $\mathbf{I}(\boldsymbol{\theta})$ at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, we have the asymptotic equivalence of the expected Fisher information and the observed information evaluated at $\hat{\boldsymbol{\theta}}$, i.e. $\mathbf{I}(\boldsymbol{\theta}) \approx \mathbf{j}\left(\hat{\boldsymbol{\theta}}\right)$. Therefore, when $\mathbf{I}(\boldsymbol{\theta})$ is difficult to obtain, then it can be approximated and replaced by $\mathbf{j}\left(\hat{\boldsymbol{\theta}}\right)$ and the resulting statistics still converge in distribution to the Central Chi-square distribution with p degrees of freedom.

When $\theta = \psi$ is a scalar parameter of interest, then the three test statistics can be applied in square root version: (1) standardized score statistic $S = S(\theta) = \frac{s(\theta)}{\sqrt{I(\theta)}} \xrightarrow{d} N(0,1)$; (2) standardized maximum likelihood departure statistic $q = q(\theta) = (\hat{\theta} - \theta)\sqrt{I(\theta)} \xrightarrow{d} N(0,1)$; and (3) signed log-likelihood ratio statistic or deviance statistic[3] $\operatorname{sgn}\left(\hat{\theta} - \theta\right)\sqrt{2\left[\ell\left(\hat{\theta}\right) - \ell(\theta)\right]} \xrightarrow{d} N(0,1)$. Again, (4) $I(\theta)$ can be replaced by $j\left(\hat{\theta}\right)$ and the resulting statistics still converge in distribution to the standard normal distribution. (Cox and Hinkley 1979)

2.3.2 Constrained Maximum Likelihood Estimation

Constrained maximum likelihood estimation is to get the **constrained maximum likelihood estimator** or constrained MLE $\hat{\boldsymbol{\theta}}_\psi = \left(\psi, \hat{\boldsymbol{\lambda}}'_\psi\right)' = \arg\max_{\boldsymbol{\lambda}} \ell(\boldsymbol{\theta}, \mathbf{y})$ where the maximum is taken given fixed value of ψ. $\hat{\boldsymbol{\theta}}_\psi$ is generally obtained from solving $l_{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\theta}}_\psi\right) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\lambda}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_\psi} = \mathbf{0}$ when $\boldsymbol{\lambda}$ is explicitly known. However, when $\boldsymbol{\lambda}$ is not explicitly available, $\hat{\boldsymbol{\theta}}_\psi$ is often obtained by applying the Lagrange multiplier method. In particular, the Lagrangian

[3] $2\left[\ell\left(\hat{\theta}\right) - \ell(\theta)\right]$ is called the deviance and R is called the directed deviance by Lindsey (1996)

function is

$$H(\boldsymbol{\theta}, \alpha) = \ell(\boldsymbol{\theta}; \mathbf{y}) + \alpha[\psi(\boldsymbol{\theta}) - \psi] \ . \tag{2.3.7}$$

In addition, for any given value of ψ, the constrained MLE $\hat{\boldsymbol{\theta}}_\psi$ and the Lagrange multiplier estimator $\hat{\alpha}$ satisfies the following first order condition:

$$\left.\frac{\partial H(\boldsymbol{\theta}, \alpha)}{\partial \boldsymbol{\theta}}\right|_{\left(\hat{\boldsymbol{\theta}}_\psi, \hat{\alpha}\right)} = \mathbf{0} \ , \tag{2.3.8}$$

$$\left.\frac{\partial H(\boldsymbol{\theta}, \alpha)}{\partial \alpha}\right|_{\left(\hat{\boldsymbol{\theta}}_\psi, \hat{\alpha}\right)} = 0 \ . \tag{2.3.9}$$

Note that since the closed form of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_\psi$ are not always available, numerical methods are often required to calculate them.

The tilted log-likelihood function which will be used later is defined as

$$\tilde{\ell}(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + \hat{\alpha}(\psi(\boldsymbol{\theta}) - \psi) \ . \tag{2.3.10}$$

It is easy to show that $\tilde{\ell}\left(\hat{\boldsymbol{\theta}}_\psi\right) = \ell\left(\hat{\boldsymbol{\theta}}_\psi\right)$. Also we define $\tilde{\mathbf{j}}(\boldsymbol{\theta}) = \tilde{\mathbf{j}}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\boldsymbol{\theta}) = -\tilde{\ell}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\boldsymbol{\theta}) = -\frac{\partial^2 \tilde{\ell}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$ to be the **observed information matrix for the tilted log-likelihood function.**

To test a parameter of interest ψ in the presence of a vector nuisance parameter $\boldsymbol{\lambda}$, the analogs of (2.3.2) to (2.3.6) are indicated by the asymptotic distribution of $\hat{\psi}$ and $\ell_\psi(\boldsymbol{\theta})$ and (2.3.6):

$$\hat{\psi} \xrightarrow{d} N\left(\psi, I^{\psi\psi}(\boldsymbol{\theta})\right) = N\left(\psi, \left(\frac{|\mathbf{I}(\boldsymbol{\theta})|}{|\mathbf{I}_{\boldsymbol{\lambda}\boldsymbol{\lambda}'}(\boldsymbol{\theta})|}\right)^{-1}\right) \ , \tag{2.3.11}$$

$$\ell_\psi(\boldsymbol{\theta}) \xrightarrow{d} N\left(0, \left(I^{\psi\psi}(\boldsymbol{\theta})\right)^{-1}\right) = N\left(0, \left(\frac{|\mathbf{I}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\boldsymbol{\theta})|}{|\mathbf{I}(\boldsymbol{\theta})|}\right)^{-1}\right) \ . \tag{2.3.12}$$

Hence, for testing the hypothesis, say H_0: $\psi = \psi_0$, we have:

$$q = q(\psi) = \left(\hat{\psi} - \psi\right)\sqrt{\frac{1}{I^{\psi\psi}\left(\hat{\boldsymbol{\theta}}_\psi\right)}} \approx \left(\hat{\psi} - \psi\right)\sqrt{\frac{|\mathbf{j}\left(\hat{\boldsymbol{\theta}}\right)|}{|\mathbf{j}_{\boldsymbol{\lambda}\boldsymbol{\lambda}'}\left(\hat{\boldsymbol{\theta}}_\psi\right)|}} \xrightarrow{d} N(0,1) \ , \tag{2.3.13}$$

$$S = S(\psi) = \ell_\psi\left(\hat{\boldsymbol{\theta}}_\psi\right)\sqrt{I^{\psi\psi}\left(\hat{\boldsymbol{\theta}}_\psi\right)} \approx \ell_\psi\left(\hat{\boldsymbol{\theta}}_\psi\right)\sqrt{\frac{|\mathbf{j}_{\boldsymbol{\lambda}\boldsymbol{\lambda}'}\left(\hat{\boldsymbol{\theta}}_\psi\right)|}{|\mathbf{j}\left(\hat{\boldsymbol{\theta}}\right)|}} \xrightarrow{d} N(0,1) \ , \tag{2.3.14}$$

$$R = R(\psi) = \text{sgn}\left(\hat{\psi} - \psi\right)\sqrt{2\left[\ell_p\left(\hat{\psi}\right) - \ell_p(\psi)\right]}$$
$$= \text{sgn}\left(\hat{\psi} - \psi\right)\sqrt{2\left[\ell\left(\hat{\boldsymbol{\theta}}\right) - \ell\left(\hat{\boldsymbol{\theta}}_\psi\right)\right]} \xrightarrow{d} N(0,1) \ , \qquad (2.3.15)$$

where $\ell_p(\psi)$ is the profile log-likelihood function which can be transformed to ordinary log-likelihood function like $\ell_p(\psi) = \ell\left(\hat{\boldsymbol{\theta}}_\psi\right) = \ell\left(\psi, \hat{\boldsymbol{\lambda}}_\psi\right)$ and $\ell_p\left(\hat{\psi}\right) = \ell\left(\hat{\boldsymbol{\theta}}\right) = \ell\left(\hat{\psi}, \hat{\boldsymbol{\lambda}}\right)$. Also $\mathbf{j}_{\boldsymbol{\lambda}\boldsymbol{\lambda}'}\left(\hat{\boldsymbol{\theta}}_\psi\right)$ is the observed nuisance information concerning the nuisance parameter $\boldsymbol{\lambda}$ for given ψ. For detailed descriptions, see Barndorff-Nielsen and Cox (1994 p.91).

Before we end this subsection, we need to mention that besides (2.3.11) and (2.3.12) which come from simple multivariate normal theory, we can also apply Delta method for approximation:

$$\hat{\psi} = \psi\left(\hat{\boldsymbol{\theta}}\right) \xrightarrow{d} N\left(\psi(\boldsymbol{\theta}), \psi'_{\boldsymbol{\theta}}(\boldsymbol{\theta})\mathbf{I}^{-1}(\boldsymbol{\theta})\psi_{\boldsymbol{\theta}}(\boldsymbol{\theta})\right) \ . \qquad (2.3.16)$$
$$\text{or} \xrightarrow{d} N\left(\psi(\boldsymbol{\theta}), \psi'_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}\right)\mathbf{j}^{-1}\left(\hat{\boldsymbol{\theta}}\right)\psi_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}\right)\right) \qquad (2.3.17)$$

It corresponds to (2.3.11) and the unknown $I^{\psi\psi}(\boldsymbol{\theta})$ can be approximated either by $\psi'_{\boldsymbol{\theta}}(\boldsymbol{\theta})\mathbf{I}^{-1}(\boldsymbol{\theta})\psi_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ or by $\psi'_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}\right)\mathbf{j}^{-1}\left(\hat{\boldsymbol{\theta}}\right)\psi_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}\right)$.[4]

2.3.3 Analysis Over the Three Types of Test Statistics

First, these statistics are asymptotically normal or chi-squared distribution, and the relative errors are of order $O\left(n^{-\frac{1}{2}}\right)$. In other words, these three statistics have first order accuracy. Engle (1984) showed that the three tests are asymptotically equivalent. **First order asymptotic** theory is very useful and has been widely used in practice. However, it has many inadequacies that provide the stimulus to further develop asymptotic theory.

1. First-order methods depend heavily on the model being approximately normally distributed and the sample size or some information measure being large. In addition when the number of nuisance parameters is large, first order theory may fail to give reasonable and accurate approximations. Therefore there is a great need for more refined distributional approximations.

2. In addition, first order theory yields different measures of departure

[4]Note that Delta method itself is a first order approximation plus the first order approximation of (2.3.4), it is thus an approximation's approximation and theoretically less accurate than (2.3.11) or (2.3.12).

and then different test properties. They may, however, in some situations produce strikingly different results.

3. Finally, in some problems, when first order theory provides surprisingly good approximations, the application of the higher order asymptotic can serve more to verify this than to provide a substantial improvement (Pierce and Peters. 1992).

Second, **the significance function** or **p-value function** or **confidence distribution function** of θ at $\hat{\theta}$ is p: $\Theta \to [0,1]$ and defined to be $p(\theta) = F\left(\hat{\theta};\theta\right) = P\left(\hat{\Theta} \leq \hat{\theta};\theta\right)$. It measures the probability to the left or right of the data point given the distribution under θ. This definition is discussed in detail in Fraser (1991). In particular, the corresponding significance functions approximated by the three methods for testing ψ are defined as $p_1(\psi) = \Phi(S)$, $p_2(\psi) = \Phi(q)$ and $p_3(\psi) = \Phi(R)$ and we have $p_i(\psi) = p(\psi) + O\left(n^{-\frac{1}{2}}\right)$ with $p(\psi)$ representing the exact probability for each case. In other words, these p-values have order of convergence $O\left(n^{-\frac{1}{2}}\right)$ and are generally referred to as first-order methods. In addition, all possible confidence intervals can be obtained by inverting $p(\theta)$. A centered $(1-\alpha)\times 100\%$ confidence interval for ψ can either be obtained by solving $P\left(|S| < z_{\alpha/2}\right) = 1-\alpha$, $P\left(|q| < z_{\alpha/2}\right) = 1-\alpha$, $P\left(|R| < z_{\alpha/2}\right) = 1-\alpha$ where z_α is the α quartile of the standard normal distribution, or be obtained by calculating

$$\left(\min\left\{p^{-1}\left(\frac{\alpha}{2}\right), p^{-1}\left(1-\frac{\alpha}{2}\right)\right\}, \max\left\{p^{-1}\left(\frac{\alpha}{2}\right), p^{-1}\left(1-\frac{\alpha}{2}\right)\right\}\right). \quad (2.3.18)$$

Example. A sample $(y_1,\ldots,y_n)$ is drawn from $N\left(\mu,\sigma_0^2\right)$, with unknown population expectation μ and known population variance σ_0^2, to test the null hypothesis $H_0 : \mu \geq \mu^*$.

We first derive the exact p-value function.

$$\begin{aligned} p(\mu^*) &= P\left(\bar{Y} < \bar{y};\mu = \mu^*\right) \\ &= P\left(\sum_{i=1}^n Y_i < \sum_{i=1}^n y_i;\mu = \mu^*\right). \end{aligned}$$

Here, we need an exact theoretical result rather than an approximation, and this result is if Y_i is IID $N\left(\mu,\sigma_0^2\right)$ then $\sum_{i=1}^n Y_i \sim N\left(n\mu, n\sigma_0^2\right)$. Replacing μ^* with any hypothesized value μ, **the exact p-value function of μ at sample mean** $\bar{y}$ is just the cumulative distribution function of $N\left(n\mu, n\sigma_0^2\right)$ at point $\sum_{i=1}^n y_i$.

Then, we derive its approximation. As is known that if Y_i is IID $N\left(\mu,\sigma^2\right)$ then $\bar{Y}\sim N\left(\mu,\frac{\sigma^2}{n}\right)$. This result comes from Central Limit Theorem with the proof listed later in the subsection (2.4.2.2) and we will see that it holds with first order accuracy $O\left(n^{-\frac{1}{2}}\right)$.

$$\begin{aligned} p\left(\mu^*\right) &= P\left(\bar{Y}<\bar{y};\mu=\mu^*\right) \\ &= P\left(\frac{\bar{Y}-\mu^*}{\sqrt{\frac{\sigma_0^2}{n}}}<\frac{\bar{y}-\mu^*}{\sqrt{\frac{\sigma_0^2}{n}}}\right)=\Phi\left(\frac{\bar{y}-\mu^*}{\sqrt{\frac{\sigma_0^2}{n}}}\right) . \end{aligned}$$

Replacing μ^* with any hypothesized value μ, we obtain the approximated p-value function of μ at the sample mean $\bar{y}$.

$$p\left(\mu\right)=\Phi\left(\frac{\bar{y}-\mu}{\sqrt{\frac{\sigma_0^2}{n}}}\right)+O\left(n^{-\frac{1}{2}}\right) .$$

On the other hand, we can also derive the approximation of p-value function from the likelihood-based statistical method. The basic likelihood results are: $\ell\left(\mu\right)=a-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}\left(y_i-\mu\right)^2$, $s\left(\mu\right)=\ell_\mu\left(\mu\right)=\frac{1}{\sigma_0^2}\sum_{i=1}^{n}\left(y_i-\mu\right)=\frac{n}{\sigma_0^2}\left(\bar{y}-\mu\right)$ with $\hat{\mu}=\bar{y}$, $\ell_{\mu\mu}\left(\mu\right)=-\frac{n}{\sigma_0^2}$ and $I\left(\theta\right)=\frac{n}{\sigma_0^2}$. From these results, we can obtain the following equivalent approximations of $p\left(\mu\right)$:

(1). Standardized score statistic is $S=S\left(\mu\right)=\frac{s(\mu)}{\sqrt{I(\mu)}}=\frac{n(\bar{y}-\mu)}{\sigma_0^2\cdot\sqrt{\frac{n}{\sigma_0^2}}}=\frac{\bar{y}-\mu}{\sqrt{\frac{\sigma_0^2}{n}}}$ and $p\left(\mu\right)=\Phi\left(S\right)+O\left(n^{-\frac{1}{2}}\right)$;

(2). Standardized maximum likelihood departure statistic is $q=q\left(\mu\right)=\left(\hat{\mu}-\mu\right)\sqrt{I\left(\mu\right)}=\left(\bar{y}-\mu\right)\sqrt{\frac{n}{\sigma_0^2}}=\frac{\bar{y}-\mu}{\sqrt{\frac{\sigma_0^2}{n}}}$ and $p\left(\mu\right)=\Phi\left(q\right)+O\left(n^{-\frac{1}{2}}\right)$;

(3). Signed log-likelihood ratio statistic is $R=R\left(\mu\right)$ and its expression will be $\operatorname{sgn}\left(\hat{\mu}-\mu\right)\sqrt{2\left[\ell\left(\hat{\mu}\right)-\ell\left(\mu\right)\right]}=\frac{\bar{y}-\mu}{\sqrt{\frac{\sigma_0^2}{n}}}$ and $p\left(\mu\right)=\Phi\left(R\right)+O\left(n^{-\frac{1}{2}}\right)$.

□

Finally, it is important to note that q is not invariant under a one-to-one transformations of parameter or the so-called reparameterization, while R is. In addition, for finite sample, Doganaksoy and Schmee (1993) found that generally R is more accurate than q. In addition, Fraser (1991) examined the accuracy of S, q and R using four different models; normal, extreme value, gamma(3) and the Cauchy distribution and his results show that quite generally R is superior to both S and q. Neyman and Pearson (1992) also proved that R gives the most powerful test. However, they also pointed out that q is more popular in applied analysis than R because of its sim-

plicity in application. In the present context the two quantities q and R are more popular in terms of application and in this dissertation only these two methods will be considered.

2.4 Edgeworth and Saddlepoint Expansion

An active area of development in theoretical statistics is that of so-called higher-order asymptotics, in which an asymptotic expansion for the density or distribution function of a statistic of interest is obtained. The purpose in obtaining such an expansion is to use the first two or three terms to provide an approximation to the density or distribution function. The inclusion of the first two or more terms would typically improve finite-sample approximation to the true distribution function than that based on just the first term. Two basic expansion methods are most frequently involved: Edgeworth expansion and saddlepoint expansion.

2.4.1 Moment Generating Function, Characteristic Function and Cumulant Generating Function

This part contains background material on three important transforms in probability theory. In the next two subsection, they will be of great use to derive Edgeworth expansion and saddlepoint expansion.

2.4.1.1 Moment Generating Function

A moment is a specific quantitative measure of the shape of a set of points. If the points represent probability density, then we have the following definitions: The nth moment of a probability density function $f(x)$ of a real-valued random variable X about a value c is $\int_{-\infty}^{+\infty}(x-c)^n f(x)\,dx$. However, the n th moment of a function, without further explanation, usually refers to the nth moment about zero, the so-called **raw moment** or crude moment: $m'_n = \int_{-\infty}^{+\infty} x^n f(x)\,dx = E[X^n]$. In addition, the moments about its mean $\mu = E[X]$ are called **central moments**, $m_n = \int (x-\mu)^n f(x)\,dx = E[(X-\mu)^n] = \sum_{j=0}^{n} \left(\begin{array}{c} n \\ j \end{array} \right) (-\mu)^{n-j} m'_j$.[5] For the second and higher moments, central moments are used in preference to raw moments, because higher-order central moments provide clearer information about the spread and shape of the distribution, independently of its location. Finally, n th **standardized moment** of a probability distribution is $\alpha_n = \int_{-\infty}^{+\infty} \left(\frac{x-\mu}{\sigma}\right)^n f(x)\,dx = E\left[\left(\frac{X-\mu}{\sigma}\right)^n\right] = \frac{m_n}{\sigma^n}$. It is the normalization of the n th moment with respect to standard deviation. Since $m_k(\lambda X) = \lambda^k m_k$ and $m'_k(\lambda X) = \lambda^k m'_k$,

[5] For random variables that have no mean, such as the Cauchy distribution, central moments are not defined.

m_k and m'_k are homogeneous polynomials of degree k, and furthermore, the standardized moment is scale invariant. This can be understood in a way that moments have dimension while the dimension cancels in $\frac{m_n}{\sigma^n}$, leaving dimensionless numbers, and that the distribution is independent of any linear change of scale.

When $E\left[|X^n|\right] = \infty$, then the moment is said not to exist. However, if the n th moment about any point exists, so does the $(n–1)$ th moment, and thus all lower-order moments, about every point. For a bounded distribution of mass or probability, the collection of all the moments of all orders, from 0 to ∞, uniquely determines the distribution.

For all k, the k th raw moment of a population can be estimated using the k th raw sample moment $\frac{\sum_{j=1}^{n} X_j^k}{n}$ applied to a sample $X_1, \ldots, X_n$ drawn from the population. It can be shown that if m'_k exists, then $m'_k = E\left[\frac{\sum_{j=1}^{n} X_j^k}{n}\right]$. It is thus an unbiased estimator. This contrasts with the situation for central moments, whose computation uses up a degree of freedom by using the sample mean. So for example an unbiased estimate of the second central moment, the population variance, is given by $\frac{\sum_{j=1}^{n}\left(X_j - \bar{X}\right)^2}{n-1}$ in which the previous denominator n has been replaced by the degrees of freedom $n–1$. This estimate of the population moment is greater than the unadjusted observed sample moment $\frac{\sum_{j=1}^{n}\left(X_j - \bar{X}\right)^2}{n}$ by a factor of $\frac{n}{n-1}$, and it is referred to as the "adjusted sample variance" or sometimes simply the "sample variance".

The **moment generating function** of a random variable X is the expectation of a function of X: $M_X(t) = E\left[e^{tX}\right] = e^{K_X(t)}$, $t \in \mathbb{R}$, wherever this expectation exists. In addition, the central moment generating function is: $C_X(t) = E\left[e^{t(X-\mu)}\right] = e^{-\mu t} M_X(t) = e^{K_X(t) - \mu t}$, $t \in \mathbb{R}$, wherever this expectation exists. The reason for defining these functions is that they can be used to find all the moments of the distribution (Bulmer 2012). Specifically

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{+\infty} e^{tx} f(x)\, dx = E\left(e^{tX}\right) = E\left[1 + tx + \frac{t^2 x^2}{2!} + \cdots\right] \\ &= 1 + t m'_1 + \frac{t^2 m'_2}{2!} + \cdots = \sum_{j=0}^{\infty} \frac{t^j m'_j}{j!}\,, \end{aligned} \tag{2.4.1}$$

$$\begin{aligned} C_X(t) &= E\left[e^{t(X-\mu)}\right] = E\left[1 + t(x-\mu) + \frac{t^2 (x-\mu)^2}{2!} + \cdots\right] \\ &= 1 + t m_1 + \frac{t^2 m_2}{2!} + \cdots = \sum_{j=0}^{\infty} \frac{t^j m_j}{j!}\,. \end{aligned}$$

Hence, both moment generating functions are the exponential generating function of the moments of the probability distribution. For a nonnegative integer n,

$$m'_n = M_X^{(n)}(0) = \left.\frac{\partial^n M_X(t)}{\partial t^n}\right|_{t=0} = \left.\frac{\partial^n e^{K_X(t)}}{\partial t^n}\right|_{t=0},$$

$$m_n = C_X^{(n)}(0) = \left.\frac{\partial^n e^{K_X(t)-\mu t}}{\partial t^n}\right|_{t=0}.$$

By Faà di Bruno's formula, the above expressions link the moments in terms of cumulants, and thus we can obtain $m'_n = \sum_{k=1}^{n} B_{n,k}(\kappa_1, \ldots, \kappa_{n-k+1})$ and $m_n = \sum_{k=1}^{n} B_{n,k}(0, \kappa_2, \ldots, \kappa_{n-k+1})$, where $B_{n,k}$ are incomplete Bell polynomials. Thus, n th moment is a n th degree polynomial in the first n cumulants, and to express the central moments as functions of the cumulants, just drop from these polynomials all terms in which κ_1 appears as a factor.

In addition, there are particularly simple results for the moment generating functions of distributions defined by the weighted sums of independent random variables:

$$M_{S_n}(t) = M_{\sum_{j=1}^{n} a_j X_j}(t) = M_{X_1}(a_1 t) M_{X_2}(a_2 t) \cdots M_{X_n}(a_n t).$$

An important property of the moment generating function is that if two distributions have the same moment generating function, then they are identical at almost all points. This statement is not equivalent to the statement "if two distributions have the same moments, then they are identical at all points." This is because in some cases, the moments exist and yet the moment-generating function does not, due to the fact that the limit $\sum_{j=0}^{\infty} \frac{t^j m'_j}{j!}$ may not exist. The lognormal distribution is an example of when this occurs.

The moment generating function provides the basis of an alternative route to analytical results compared with working directly with probability density functions or cumulative distribution functions or the characteristic functions. However, a key problem with moment generating functions is that it may not exist as the integrals defining expectation need not converge absolutely. By contrast, the characteristic function of a real-valued random variable and argument always exists for all probability distributions because it is the integral of a bounded continuous function on a space of finite measure, and thus may be used instead.

2.4.1.2 Characteristic Function

The **characteristic function** of a real-valued random variable X is the function $\varphi_X \colon \mathbb{R} \to \mathbb{C}$ given by

$$\varphi_X(t) = E\left[e^{itX}\right] = \int_R e^{itx} f_X(x)\,dx = \overline{\int_R e^{-itx} f_X(x)\,dx} = \overline{P(t)}\,,$$

where i is the imaginary unit and $t \in \mathbb{R}$ is the argument of the characteristic function. If X has a probability density function $f_X(x)$ and $P(t) = \int_R e^{-itx} f_X(x)\,dx$ denotes the continuous Fourier transform of the probability density function, then the characteristic function is the inverse Fourier transform of the probability density function, and is the complex conjugate of $P(t)$ (Billingsley 2008).[6] Likewise, f_X may be recovered from φ_X through the inverse Fourier transform, which will be given below. In addition, Oberhettinger (2014) provides extensive tables of characteristic functions.

A characteristic function is uniformly continuous on the entire space and bounded: $|\varphi(t)| \leq 1$. In particular, characteristic function is Hermitian, $\varphi(-t) = \varphi(t)$. Similar to moment generating function, characteristic function approach is particularly useful in analysis of linear combinations of independent random variables:

$$\varphi_{S_n}(t) = \varphi_{\sum_{j=1}^{n} a_j X_j}(t) = \varphi_{X_1}(a_1 t)\,\varphi_{X_2}(a_2 t)\cdots\varphi_{X_n}(a_n t)\ . \tag{2.4.2}$$

Similarly to the cumulative distribution function, the characteristic function completely determines the behavior and properties of the probability distribution of a random variable X.

First, there is a bijection between probability distributions and characteristic functions. That is, for any two random variables X_1, X_2, they both have the same probability distribution if and only if $\varphi_{X_1} = \varphi_{X_2}$. In addition, the bijection stated above is continuous according to Lévy's continuity theorem.[7]

Second, the formula in definition of characteristic function allows us to compute φ when we know the density function f. If, on the other hand, we know the characteristic function and want to find the corresponding distribution function, then one of the following inversion theorems can be used. Suppose characteristic function φ_X is integrable, then F_X is absolutely con-

[6] It should be noted though, that the convention for the constants appearing in the definition of the characteristic function differs from the usual convention for the Fourier transform (Pinsky 2002). For example some authors (Bochner 2012) define $\varphi_X(t) = E\left[e^{-2\pi itX}\right]$, which is essentially a change of parameter.

[7] Lévy's continuity theorem: A sequence X_j of n variate random variables converges in distribution to random variable X if and only if the sequence φ_{X_j} converges pointwise to a function φ which is continuous at the origin. Then φ is the characteristic function of X (Cuppens 2014). This theorem is frequently used to prove the law of large numbers, and the Central Limit Theorem.

tinuous and $f_X(x) = F'_X(x) = \frac{1}{2\pi}\int_R e^{-itx}\varphi_X(t)\,dt = \frac{1}{2\pi}\int_R e^{itx}\overline{\varphi_X(t)}dt$, and, $F_X(b) - F_X(a) = \frac{1}{2\pi}\lim_{T\to\infty}\int_{-T}^{+T}\frac{e^{-ita}-e^{-itb}}{it}\varphi_X(t)\,dt$. For a random variable bounded from below, one can obtain $F(b)$ by taking a such that $\mathrm{F}(a) = 0$; otherwise, if a random variable is not bounded from below, the limit for $a \to -\infty$ gives $F(b)$ but is numerically impractical (Shephard 1991). Another theorem introduced by Wendel (1961) states that, if x is a continuity point of F_X then $F_X(x) = \frac{1}{2} - \frac{1}{\pi}\int_0^{+\infty}\frac{\mathrm{Im}\left(e^{-itx}\varphi_X(t)\right)}{t}dt$, where the imaginary part of a complex number z is given by $\mathrm{Im}(z) = (z - z*)/2i$, but however, the integral may be not Lebesgue-integrable.

Finally, there is a one-to-one correspondence between cumulative distribution function, moment generating function and characteristic function, and it is always possible to find one of these functions if we know the other ones. For example, Lukacs (1970) stated that if a random variable X has a moment generating function, then the domain of the characteristic function can be extended to the complex plane[8], and $M_X(t) = \varphi_X(-it)$; Or oppositely, the characteristic function is the moment generating function of iX or the moment generating function of X evaluated on the imaginary axis. $\varphi_X(t) = M_{iX}(t) = M_X(it)$. In fact, characteristic function is a Wick rotation of the moment generating function when the latter exists. Characteristic functions can also be used to find moments of a random variable. Provided that the k th moment exists, characteristic function is k times continuously differentiable on the entire real line and, $m'_k = M_X^{(k)}(0) = (-i)^k \varphi_X^{(k)}(0)$. Oppositely, if a characteristic function $\varphi_X(t)$ has a k th derivative at zero, then the random variable X has all moments up to k if k is even, but only up to k–1 if k is odd (Lukacs 1970). Similarly, the logarithm of a characteristic function is a cumulant generating function, which is also useful for finding cumulants.

2.4.1.3 Cumulant Generating Function

In probability theory and statistics, the cumulants of a probability distribution are a set of quantities that provide an alternative to the moments of the distribution. In some cases theoretical treatments of problems in terms of cumulants are simpler than those using moments.

The n th **cumulants** κ_n of a random variable X are defined via the **cumulant generating function** $K_X(t)$, which is the natural logarithm of the moment generating function, and via its Maclaurin series expansion:

[8]As defined above, the argument of the characteristic function is treated as a real number; however, certain aspects of the theory of characteristic functions are advanced by extending the definition into the complex plane by analytical continuation, in cases where this is possible. (Lukacs 1970)

$$K_X(t) = \log E\left[e^{tx}\right] = \log M_X(t) = \sum_{i=1}^{\infty} \frac{\kappa_i t^i}{i!} = \kappa_1 t + \frac{\kappa_2 t^2}{2!} + \frac{\kappa_3 t^3}{3!} + \cdots + \frac{\kappa_n t^n}{n!} + \cdots , \tag{2.4.3}$$

so that differentiate the above expansion n times and we can get $K_X^{(n)}(t) = \sum_{i=n}^{\infty} \frac{\kappa_i t^{i-n}}{(i-n)!} = \kappa_n + \kappa_{n+1} t + \frac{\kappa_{n+2} t^2}{2!} + \cdots$ and evaluate the result at zero: $\kappa_n = K_X^{(n)}(0) = \left.\frac{\partial^n \log M_X(t)}{\partial t^n}\right|_{t=0} = \left.\frac{\partial^n \log\left(C_X(t) e^{\mu t}\right)}{\partial t^n}\right|_{t=0}$.[9] By Faà di Bruno's formula, this expression links the cumulants in terms of moments, and we can obtain for $n \geq 1$ $\kappa_n = \sum_{k=1}^{n} (-1)^{k-1} (k-1)! B_{n,k}\left(m_1', \ldots, m_{n-k+1}'\right)$, and for $n > 1$ $\kappa_n = \sum_{k=1}^{n} (-1)^{k-1} (k-1)! B_{n,k}(0, m_2, \ldots, m_{n-k+1})$, where $B_{n,k}$ are incomplete Bell polynomials. Therefore, the nth cumulant is an n th degree polynomial in the first n raw moments and to express the cumulants as functions of the central moments, drop from these polynomials all terms in which m_1' appears as a factor. Also need to mention that to express the cumulants κ_n for $n > 2$ as functions of the standardized central moments α_n, set $m_2' = 1$ together with $m_1' = 0$ in the polynomials.

Similar to n th standardized moment, another definition named n th **standardized cumulant** is set as $\rho_n = \frac{\kappa_n}{\sigma^n} = \frac{\kappa_n}{\kappa_2^{\frac{n}{2}}} = \frac{K^{(n)}(0)}{\left(K^{tt}(0)\right)^{\frac{n}{2}}}$.

There are some useful properties of cumulants and the cumulant generating function. For example, the cumulant generating function $K_X(t)$, if it exists, passes through the origin $K_X(0) = 0$. In addition, for a degenerate point mass at a, the cumulant generating function is the straight line: $K_a(t) = at$. This result also hints that $\kappa_1(a) = a$ while for $n > 1$, $\kappa_n(a) = 0$. Also working with cumulants can have an advantage over using moments because for statistically independent random variables, given their cumulant generating functions exist,

$$K_{S_n}(t) = K_{\sum_{j=1}^{n} a_j X_j}(t) = K_{X_1}(a_1 t) + K_{X_2}(a_2 t) + \cdots + K_{X_n}(a_n t) \ , \tag{2.4.4}$$

[9] Some writers (Kendall and Stuart 1969, Lukacs 1970) prefer to define the cumulant generating function as the natural logarithm of the characteristic function, which is sometimes also called the second characteristic function, $h_X(t) = \log E\left[e^{itX}\right] = \log \varphi_X(t) = \sum_{j=1}^{\infty} \frac{\kappa_j (it)^j}{j!}$. An advantage of $h_X(t)$ is that $\varphi_X(t)$ is well defined for all real values of t even when $M_X(t)$ is not. Although the function $h_X(t)$ will be well defined, it nonetheless may mimic $K_X(t)$ by not having a Maclaurin series beyond or, rarely even to linear order in the argument t. Thus, many cumulants may still not be well defined. Both the Cauchy distribution and stable distribution (related to the Lévy distribution) are examples of distributions for which the power-series expansions of the generating functions have only finitely many well-defined terms.

$$\kappa_n\left(\sum_{j=1}^{n}a_jX_j\right)=a_1^n\kappa_n\left(X_1\right)+a_2^n\kappa_n\left(X\right)+\cdots+a_n^n\kappa_n\left(X_n\right)\ . \tag{2.4.5}$$

This result also indicates that the nth cumulant is homogeneous of degree n. Finally, Lukacs (1970) stated that given the results for the cumulants of the normal distribution, it might be hoped to find families of distributions for which $\kappa_m=\kappa_{m+1}=\cdots=0$ for some $m>3$, with the lower-order cumulants orders 3 to m–1 being non-zero. However, there are no such distributions. The underlying result here is that the cumulant generating function cannot be a finite-order polynomial of degree greater than 2.

Example. For statistical contexts the cumulant generating function arises naturally and is directly available for exponential models. At section (2.2.1.2), we introduced the family of linear exponential models in canonical form with density (2.2.8). Still at that setting, we take a closer look at the moments and cumulants of the sufficient statistic in the exponential family. The moment generating function of sufficient statistic is

$$\begin{aligned}M_{\mathbf{S}}\left(\mathbf{t}\right)&=E\left[e^{\mathbf{t}'\mathbf{s}(\mathbf{y})}\right]\\&=\int e^{\mathbf{t}'\mathbf{s}(\mathbf{y})}u\left(\mathbf{y}\right)e^{\boldsymbol{\theta}'\mathbf{s}(\mathbf{y})-A(\boldsymbol{\theta})}d\mathbf{y}\\&=e^{A(\boldsymbol{\theta}+\mathbf{t})-A(\boldsymbol{\theta})}\int u\left(\mathbf{y}\right)e^{(\mathbf{t}+\boldsymbol{\theta})'\mathbf{s}(\mathbf{y})-A(\mathbf{t}+\boldsymbol{\theta})}d\mathbf{y}\\&=e^{A(\boldsymbol{\theta}+\mathbf{t})-A(\boldsymbol{\theta})}\ .\end{aligned}$$

Thus, the cumulant generating function becomes

$$K_{\mathbf{S}}\left(\mathbf{t}\right)=A\left(\boldsymbol{\theta}+\mathbf{t}\right)-A\left(\boldsymbol{\theta}\right)\ . \tag{2.4.6}$$

Therefore, $A\left(\boldsymbol{\theta}\right)$ is also called the nominal cumulant generating function. By (2.4.6), people can easily get the cumulants of sufficient statistics. For example $E\left[\mathbf{S}\right]=\kappa_1=\left.\frac{\partial K_{\mathbf{S}}(\mathbf{t})}{\partial\mathbf{t}}\right|_{\mathbf{t}=0}=\left.\frac{\mathrm{d}A(\boldsymbol{\theta}+\mathbf{t})}{\mathrm{d}(\boldsymbol{\theta}+\mathbf{t})}\right|_{\mathbf{t}=0}=\frac{\mathrm{d}A(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}$, $Var\left[\mathbf{S}\right]=\kappa_2=\frac{\mathrm{d}^2A(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\theta}'}$, etc. Need to mention that the canonical form is non-unique, but this will not affect the conclusion from the cumulants of the sufficient statistics.

For example, we consider a gamma distribution characterized by shape α and rate β. The probability density function is $f\left(y\right)=\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta y}$. It is in the family of linear exponential models with natural parameters $\boldsymbol{\theta}=\begin{pmatrix}\theta_1\\\theta_2\end{pmatrix}=\begin{pmatrix}\alpha-1\\-\beta\end{pmatrix}$ and sufficient statistics $\mathbf{S}=\begin{pmatrix}S_1\\S_2\end{pmatrix}=\begin{pmatrix}\log X\\X\end{pmatrix}$ and $A\left(\boldsymbol{\theta}\right)=\log\Gamma\left(\theta_1+1\right)-\left(\theta_1+1\right)\log\left(-\theta_2\right)$. Thus by (2.4.6) we can get the following solutions: $E\left[S_1\right]=E\left[\log X\right]=\frac{\partial A(\boldsymbol{\theta})}{\partial\theta_1}=\frac{\Gamma'(\theta_1+1)}{\Gamma(\theta1+1)}-\log\left(-\theta_2\right)=\psi\left(\alpha\right)-\log\beta$ and $E\left[S_2\right]=E\left[X\right]=\frac{\partial A(\boldsymbol{\theta})}{\partial\theta_2}=-\frac{\theta_1+1}{\theta_2}=-\frac{\alpha}{\beta}$ and $Var\left[S_2\right]=$

$Var[X] = \frac{\partial^2 A(\boldsymbol{\theta})}{\partial \theta_2^2} = \frac{\theta_1+1}{\theta_2^2} = \frac{\alpha}{\beta^2}$. All of these calculations can be done using integration, making use of various properties of the gamma function, but this requires significantly more work. □

2.4.2 The Edgeworth Expansion

2.4.2.1 Hermite polynomials

The **Hermite polynomials** are a classical orthogonal polynomial sequence and they have two versions of definition: for $n = 0, 1, 2, \ldots$, the normal version is $H_n(x) = (-1)^n e^{\frac{x^2}{2}} \frac{\mathrm{d}^n\left(e^{-\frac{x^2}{2}}\right)}{\mathrm{d}x^n}$ as well as the generalized version is $\phi(x) H_n(x) = (-1)^n \frac{\mathrm{d}^n(\phi(x))}{\mathrm{d}x^n}$ equivalently $(\phi(x) H_n(x))' = -\phi(x) H_{n+1}(x)$, where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ is the probability density function for the standard normal distribution. And the Hermite polynomials are given by the exponential generating function $e^{xt-\frac{t^2}{2}} = \sum_{n=0}^{\infty} \left(H_n(x) \frac{t^n}{n!}\right)$.

In particular, $H_n(x)$ is the Hermite polynomial of degree n and is also a polynomial of degree n with leading coefficient 1. The first few Hermite polynomials are: $H_0(x) = 1$, $H_1(x) = x$, $H_2(x) = x^2 - 1$, $H_3(x) = x^3 - 3x$, $H_4(x) = x^4 - 6x^2 + 3$, $H_5(x) = x^5 - 10x^3 + 15x$, $H_6(x) = x^6 - 15x^4 + 45x^2 - 15$, etc. In addition, the sequence of Hermite polynomials satisfies the recursion $H_{n+1}(x) = xH_n(x) - H_n'(x)$ (or $H_{n+1}(x) = xH_n(x) - nH_{n-1}(x)$) and constitute an Appell sequence $H_n^{(m)}(x) = \frac{n!}{(n-m)!} H_{n-m}(x) = m! \begin{pmatrix} n \\ m \end{pmatrix} H_{n-m}(x)$ (or$H_n'(x) = nH_{n-1}(x)$) and also follow Turán's inequalities $H_n^2(x) - H_{n+1}(x) H_{n-1}(x) > 0$. These relations, together with the initial polynomials $H_0(x)$ and $H_1(x)$, can be used in practice to compute the polynomials quickly. For example, the Hermite polynomials evaluated at zero argument are called Hermite numbers, and according to the recursion relation $H_n(0) = -(n-1)H_{n-2}(0)$, we can obtain

$$H_n(0) = \begin{cases} 0 & \text{if } n\text{is odd;} \\ (-1)^{\frac{n}{2}} (n-1)!! & \text{if } n\text{is even.} \end{cases}$$

On the other hand, assuming $H_n(x) = \sum_{k=0}^{n} a_{n,k} x^k$, individual coefficients are related by the following recursion formula, $a_{n+1,k} = a_{n,k-1} - na_{n-1,k}$ for $k > 0$, and $a_{n+1,k} = -na_{n-1,k}$ for $k = 0$, given the initial condition $a_{0,0} = 1$, $a_{1,0} = 0$, $a_{1,1} = 1$.

For the differential equation $\left(e^{-\frac{x^2}{2}} u'\right)' + \lambda e^{-\frac{x^2}{2}} u = 0$ or $u'' - xu' + \lambda u = 0$ with the boundary conditions that u should be polynomially bounded at infinity, the equation has solutions only if λ is a non-negative integer, and up to an overall scaling, the solution is uniquely given by $u(x) = H_\lambda(x)$. In

addition, this differential equation can be rewritten as an eigenvalue problem called the Hermite equation $L[u] = u'' - xu' = -\lambda u$. And the solutions of it are the eigenfunctions of the differential operator L. In addition, the Hermite polynomials also have other properties. For example, the Hermite polynomials can be on Taylor expanding, $H_n(x+z) = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} H_k(z)$ and they can be represented as moments $H_n(x) = \int_{-\infty}^{+\infty} (x+iz)^n \phi(z)\, dz$.

The Hermite polynomials defined above are orthogonal with respect to the weight function $e^{-\frac{x^2}{2}}$ and thus orthogonal with respect to the standard normal probability distribution which has expected value 0 and variance 1. In particular, $\int_{-\infty}^{+\infty} H_m(x) H_n(x) \phi(x)\, dx = n!\delta_{mn}$. Furthermore, the generalized Hermite polynomials of different variance α where α is any positive number, $H_n^{[\alpha]}(x)$, can also be defined to be orthogonal with respect to the normal probability distribution whose density function is $\frac{1}{\sqrt{2\pi\alpha}} e^{-\frac{x^2}{2\alpha}}$.

2.4.2.2 The Edgeworth Expansion

For this section we assume that $(Y_1, \ldots, Y_n)$ are a sequence of IID random variables from density function $f_Y(y)$ with mean μ and variance σ^2. Also, set the sample sum $S_n = \sum_{j=1}^{n} Y_j$ and the standardized sample sum $S_n^* = \frac{\bar{Y}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$.

Lemma. *According to the Central Limit Theorem, $S_n^* \xrightarrow{d} N(0,1)$ with relative errors of order $O\left(n^{-\frac{1}{2}}\right)$.*

Proof. According to (2.4.4), the cumulant generating function of S_n^* can be expressed as

$$K_{S_n^*}(t) = K_{\left(\frac{Y_1-\mu}{\sigma} + \frac{Y_2-\mu}{\sigma} + \cdots + \frac{Y_n-\mu}{\sigma}\right)\frac{1}{\sqrt{n}}}(t) = nK_{Y-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right). \tag{2.4.7}$$

Then by (2.4.3), we can expand the last cumulant generating function $K_{Y-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right)$ and obtain $K_{Y-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right) = \sum_{i=1}^{\infty} \frac{\kappa_i (Y-\mu)}{i!} \left(\frac{t}{\sigma\sqrt{n}}\right)^i$. In addition, from (2.4.5) we obtain $\kappa_n(Y-\mu) = \kappa_n(Y) = \kappa_n$ for $n > 1$, together with the case at $n = 1$, $\kappa_1(Y-\mu) = 0$. Now we can substitute all these results back to (2.4.7) and obtain

$$\begin{aligned} & nK_{Y-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right) \qquad (2.4.8) \\ = \; & n\sum_{i=2}^{\infty} \frac{\kappa_i}{i!} \left(\frac{t}{\sigma\sqrt{n}}\right)^i \sum_{i=2}^{\infty} \frac{t^i}{i! n^{\frac{i}{2}-1}} \left(\frac{\kappa_i}{\sigma^i}\right) = \sum_{i=2}^{\infty} \frac{\rho_i t^i}{i! n^{\frac{i}{2}-1}} = \frac{t^2}{2} + O\left(n^{-\frac{1}{2}}\right). \end{aligned}$$

Since $K_{S_n^*}(t) = \frac{t^2}{2} + O\left(n^{-\frac{1}{2}}\right) = K_{\phi^{-1}}(t) + O\left(n^{-\frac{1}{2}}\right)$, we prove that $S_n^* \xrightarrow{d} N(0,1)$ with order of convergence $O\left(n^{-\frac{1}{2}}\right)$. □

In order to obtain Edgeworth series, we need another conclusion shown in Barndorff-Nielsen and Cox (1989 p.18).

Lemma. *For a random variable S^* and $k = 0, 1, 2, \ldots$, we have*

$$\int_{-\infty}^{+\infty} e^{ts^*} \phi(s^*) H_k(s^*) \, ds^* = t^k e^{\frac{t^2}{2}} . \tag{2.4.9}$$

Proof. To prove this result is equivalent to prove $\int_{-\infty}^{+\infty} e^{ts^* - \frac{t^2}{2}} \phi(s^*) H_k(s^*) \, ds^* = t^k$ for $k = 0, 1, 2, \ldots$. In particular, the part $e^{ts^* - \frac{t^2}{2}}$ can be expanded by the Hermite polynomial generating function according to section (2.4.2.1). And finally, we can apply the orthogonal property of Hermite polynomial to get the conclusion.

$$\begin{aligned}
\int_{-\infty}^{+\infty} e^{ts^* - \frac{t^2}{2}} \phi(s^*) H_k(s^*) \, ds^* &= \int_{-\infty}^{+\infty} \left[\sum_{n=0}^{\infty} \left(H_n(s^*) \frac{t^n}{n!} \right) \right] \phi(s^*) H_k(s^*) \, ds^* \\
&= \sum_{n=0}^{\infty} \frac{t^n}{n!} \left(\int_{-\infty}^{+\infty} H_n(s^*) H_k(s^*) \phi(s^*) \, ds^* \right) \\
&= \sum_{n=0}^{\infty} \frac{t^n}{n!} n! \delta_{nk} = \sum_{n=0}^{\infty} t^n \delta_{nk} = t^k .
\end{aligned}$$

□

Having obtained (2.4.8) and (2.4.9), now we can finally proceed to Edgeworth series. Starting from (2.4.8), the moment generating function of S_n^* can be expressed as

$$M_{S_n^*}(t) = e^{K_{S_n^*}(t)} = e^{\frac{t^2}{2} + \sum_{i=3}^{\infty} \frac{\rho_i t^i}{i! n^{\frac{i}{2}-1}}} = e^{\frac{t^2}{2}} \cdot e^{\sum_{i=3}^{\infty} \frac{\rho_i t^i}{i! n^{\frac{i}{2}-1}}} . \tag{2.4.10}$$

Being expanded in a Taylor series about $\sum_{i=3}^{\infty} \frac{\rho_i t^i}{i! n^{\frac{i}{2}-1}} = 0$ and substituted with

identity of (2.4.9), (2.4.10) can continue to evolve as

$$\begin{aligned}(2.4.10) &= e^{\frac{t^2}{2}} \cdot \left(1+\sum_{i=3}^{\infty}\frac{\rho_i t^i}{i!n^{\frac{i}{2}-1}}+\frac{1}{2}\left(\sum_{i=3}^{\infty}\frac{\rho_i t^i}{i!n^{\frac{i}{2}-1}}\right)^2+\cdots\right)\\ &= e^{\frac{t^2}{2}} \cdot \left(1+\frac{\rho_3 t^3}{3!n^{\frac{1}{2}}}+\frac{\rho_4 t^4}{4!n}+\frac{\rho_3^2 t^6}{2\times(3!)^2 n}+O\left(n^{-\frac{3}{2}}\right)\right)\\ &= e^{\frac{t^2}{2}}+\frac{\rho_3 t^3 e^{\frac{t^2}{2}}}{3!n^{\frac{1}{2}}}+\frac{\rho_4 t^4 e^{\frac{t^2}{2}}}{4!n}+\frac{\rho_3^2 t^6 e^{\frac{t^2}{2}}}{2\times(3!)^2 n}+O\left(n^{-\frac{3}{2}}\right)\\ &= \int_{-\infty}^{+\infty} e^{ts^*}\phi(s^*)\left(H_0(s^*)+\frac{\rho_3 H_3(s^*)}{3!n^{\frac{1}{2}}}+\frac{\rho_4 H_4(s^*)}{4!n}+\frac{\rho_3^2 H_6(s^*)}{2\times(3!)^2 n}\right)ds^*\\ &+O\left(n^{-\frac{3}{2}}\right).\end{aligned}$$

Inverting the above expression term by term by the definition of moment generating function (2.4.1), we can finally find the **Edgeworth expansion** for the probability density function of S_n^*

$$f_{S_n^*}(s^*)=\phi(s^*)\left(1+\frac{\rho_3 H_3(s^*)}{6n^{\frac{1}{2}}}+\frac{\rho_4 H_4(s^*)}{24n}+\frac{\rho_3^2 H_6(s^*)}{72n}\right)+O\left(n^{-\frac{3}{2}}\right). \tag{2.4.11}$$

We can see from (2.4.11) that, the error of the leading term $\phi(s^*)$ is $O\left(n^{-\frac{1}{2}}\right)$ in general, provided that $\rho_3 \neq 0$. Thus the convergence for first order theory using the normal approximation is relatively slow, especially in the tails of the distribution where the value of $H_3(s^*)$ can be appreciable. Unfortunately, it is often the tails where one wants to get good estimates in order to construct confidence intervals and obtain p-values. Nonetheless, if we are concerned with behavior at or near the mean, i.e. $s^*=0$, we virtually have $f_{S_n^*}(0)=\frac{1}{\sqrt{2\pi}}\left(1+\frac{\rho_4}{8n}-\frac{5\rho_3^2}{24n}\right)+O\left(n^{-\frac{3}{2}}\right)$. In this circumstance, $H_3(s^*)$ vanishes since all the odd order Hermite polynomials vanish at $s^*=0$, and thus the term related to $\frac{1}{\sqrt{n}}$ will disappear, and the error of the standard normal approximation is $O\left(n^{-1}\right)$ rather than $O\left(n^{-\frac{1}{2}}\right)$.

The general version of the definition of Hermite polynomials leads to $\int_{-\infty}^{s^*}\phi(y)H_{n+1}(y)\,dy=-\phi(s^*)H_n(s^*)$. Integrating (2.4.11) term by term through this way, we obtain the Edgeworth series of the cumulative distribution function of S_n^* which completely determines behavior and properties

of the probability distribution of the random variable S_n^*.

$$\begin{aligned}
F_{S_n^*} & (s^*) \\
= & E\left[1_{\{S_n^* \leq s^*\}}\right] \\
= & \int_{-\infty}^{s^*} f_{S_n^*}(y)\, dy \\
= & \int_{-\infty}^{s^*} \left[\phi(y)\left(1 + \frac{\rho_3 H_3(y)}{6n^{\frac{1}{2}}} + \frac{\rho_4 H_4(y)}{24n} + \frac{\rho_3^2 H_6(y)}{72n}\right)\right] dy + O\left(n^{-\frac{3}{2}}\right) \\
= & \int_{-\infty}^{s^*} \phi(y)\, dy + \frac{\rho_3 \int_{-\infty}^{s^*} \phi(y) H_3(y)\, dy}{6n^{\frac{1}{2}}} + \frac{\rho_4 \int_{-\infty}^{s^*} \phi(y) H_4(y)\, dy}{24n} + \\
& \frac{\rho_3^2 \int_{-\infty}^{s^*} \phi(y) H_6(y)\, dy}{72n} + O\left(n^{-\frac{3}{2}}\right) \\
= & \Phi(s^*) - \frac{\rho_3 \phi(s^*) H_2(s^*)}{6n^{\frac{1}{2}}} - \frac{\rho_4 \phi(s^*) H_3(s^*)}{24n} - \frac{\rho_3^2 \phi(s^*) H_5(s^*)}{72n} + O\left(n^{-\frac{3}{2}}\right) \\
= & \Phi(s^*) - \phi(s^*)\left\{\frac{\rho_3 H_2(s^*)}{6n^{\frac{1}{2}}} + \frac{\rho_4 H_3(s^*)}{24n} + \frac{\rho_3^2 H_5(s^*)}{72n}\right\} + O\left(n^{-\frac{3}{2}}\right) \quad (2.4.12)
\end{aligned}$$

Thus, a distribution with given cumulants κ_n can be approximated through an Edgeworth series. For background on the Edgeworth expansion see Cramér (1999). The discussion in this section follows the treatment given in Barndorff-Nielsen and Cox (1989, ch.4). For a survey of the multivariate Edgeworth expansion see McCullaugh (1987, Ch.5).

2.4.3 The Saddlepoint Expansion

A modified version of the Edgeworth expansion is the saddlepoint method (Daniels, 1954; Barndorff-Nielsen and Cox, 1979). The saddlepoint method provides extremely accurate approximations to density functions based on corresponding cumulant generating functions. This subsection will introduce the saddlepoint expansion by using two different approaches. The first approach, from which the approximation takes its name, uses the saddlepoint technique from applied mathematics and it will be outlined in subsection (2.4.3.1). In the later subsections, an alternative more statistical approach is built through an Edgeworth expansion for a tilted exponential model centered on the data point in question.

2.4.3.1 The Saddlepoint Approximation

Before we derive the saddlepoint approximation, we need a result on Laplace approximation.

Lemma. *Consider a function $f(x)$ that is a smooth function on $[a, b]$, and the endpoints a and b could possibly be infinite. Assume that $\hat{x} \in [a, b]$ is the*

unique point for f such that $f(\hat{x}) = \max_{[a,b]} f(x)$ with $f'(\hat{x}) = 0$ and $f''(\hat{x}) < 0$. In addition, $h(x)$ is a positive smooth function and is relatively diffuse and flat over the neighborhood of $\hat{x}$. Then we have

$$\int_a^b h(x) e^{Mf(x)} dx = \sqrt{\frac{2\pi}{-Mf''(\hat{x})}} h(\hat{x}) e^{Mf(\hat{x})} + O\left(\frac{1}{M}\right). \tag{2.4.13}$$

The idea here is that the integral value is largely determined by the local properties of f and h at critical value $\hat{x}$ as expressed through the values $f(\hat{x})$, $h(\hat{x})$ and curvature $f''(\hat{x})$ at that point. When M is large, the significant contribution to the integral of these functions is essentially entirely originating from a neighborhood around $\hat{x}$.

Proof. We formalize this proof by Taylor expansion of the function f around $\hat{x}$: $f(x) \approx f(\hat{x}) + \frac{1}{2} f''(\hat{x}) (x - \hat{x})^2 + O\left((x - \hat{x})^3\right)$, and further approximate h linearly around $\hat{x}$: $h(x) \approx h(\hat{x}) + O((x - \hat{x}))$. Therefore we can obtain

$$\begin{aligned}
\int_a^b h(x) e^{Mf(x)} dx &\approx \int_a^b h(\hat{x}) e^{Mf(\hat{x}) + \frac{M}{2} f''(\hat{x})(x-\hat{x})^2} dx \\
&= h(\hat{x}) e^{Mf(\hat{x})} \int_a^b e^{\frac{M}{2} f''(\hat{x})(x-\hat{x})^2} dx \\
&\approx \sqrt{\frac{2\pi}{-Mf''(\hat{x})}} h(\hat{x}) e^{Mf(\hat{x})}.
\end{aligned}$$

The last approximation comes from a result: $\int_a^b e^{\frac{M}{2} f''(\hat{x})(x-\hat{x})^2} dx \approx \sqrt{\frac{2\pi}{-Mf''(\hat{x})}}$. The left integral is a Gaussian integral if the limits of integration go from $-\infty$ to $+\infty$, which can be assumed because the exponential decays very fast away from $\hat{x}$ when M is large. □

A generalization of Laplace method and extension to arbitrary precision is provided by Fog (2008).

For this section, we assume that the moment generating function exists in an open neighborhood around the origin. By the inversion theorem of characteristic function, the probability density function can be got as $f_Y(y) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-it_0 y} \varphi_Y(t_0) dt_0 = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{K_Y(it_0) - it_0 y} dt_0$. Then applying (2.4.13) on this result, we will have $\frac{1}{2\pi} \sqrt{\frac{2\pi}{-(i)^2 K_Y''(i\hat{t}_0)}} e^{K_Y(i\hat{t}_0) - i\hat{t}_0 y}$. Finally we take $\hat{t} = i\hat{t}_0$ and obtain the saddlepoint approximation for the probability density function of Y as

$$\hat{f}_Y(y) = \sqrt{\frac{1}{2\pi K_Y''(\hat{t})}} e^{K_Y(\hat{t}) - \hat{t}y}, \tag{2.4.14}$$

where $\hat{t} = \hat{t}(y)$ is known as the saddlepoint and is the solution to the saddlepoint equation $K'_Y\left(\hat{t}\right) = y$. We can see that it is never negative.

For a sample of IID random variables $(Y_1, \ldots, Y_n)$, the saddlepoint approximation to the probability density function of the mean of the random sample $\bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$ and the sample sum $S_n = \sum_{j=1}^{n} Y_j$ are

$$\hat{f}_{\bar{Y}}(\bar{y}) = \sqrt{\frac{n}{2\pi K''_Y\left(\hat{t}\right)}} e^{n\left[K_Y\left(\hat{t}\right) - \hat{t}\bar{y}\right]}, \tag{2.4.15}$$

where $\hat{t} = \hat{t}(\bar{y})$ satisfies $K'_Y\left(\hat{t}\right) = \bar{y}$, and

$$\hat{f}_{S_n}(s) = \sqrt{\frac{1}{2\pi n K''_Y\left(\hat{t}\right)}} e^{nK_Y\left(\hat{t}\right) - \hat{t}s}, \tag{2.4.16}$$

where $\hat{t} = \hat{t}(s)$ satisfies $K'_Y\left(\hat{t}\right) = \frac{s}{n}$. In addition, let $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_n$ be IID k-dimensional random vectors from a density $f_{\mathbf{V}}(\mathbf{v})$ on $\mathbb{R}^k$. Denote by $K_{\mathbf{V}}(\mathbf{t}) = \log E\left[e^{\mathbf{t}'\mathbf{V}}\right]$ the cumulant generating function for $\mathbf{V}_i$. Then the saddlepoint approximation to $f_{\mathbf{V}}(\mathbf{v})$ is given by

$$\hat{f}_{\mathbf{V}}(\mathbf{v}) = \sqrt{\frac{1}{(2\pi)^k \left|K''_{\mathbf{V}}\left(\hat{\mathbf{t}}\right)\right|}} e^{K_{\mathbf{V}}\left(\hat{\mathbf{t}}\right) - \hat{\mathbf{t}}'\mathbf{v}}, \tag{2.4.17}$$

where $\hat{\mathbf{t}} = \hat{\mathbf{t}}(\mathbf{v})$ is the solution to the saddlepoint equation $K'_{\mathbf{V}}\left(\hat{\mathbf{t}}\right) = \mathbf{v}$, and where $K'_{\mathbf{V}}\left(\hat{\mathbf{t}}\right) = \left.\frac{\partial K(\mathbf{t})}{\partial \mathbf{t}}\right|_{\hat{\mathbf{t}}}$ and $K''_{\mathbf{V}}\left(\hat{\mathbf{t}}\right) = \left.\frac{\partial^2 K(\mathbf{t})}{\partial \mathbf{t}\partial \mathbf{t}'}\right|_{\hat{\mathbf{t}}}$.

Example. For a set of IID normal random variables $(X_1, \ldots, X_n)$ where each individual point follows $N\left(\mu, \sigma^2\right)$, according to Bernstein's theorem (Lukacs & King 1954), $\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ also follow a normal distribution but with mean μ and variance $\frac{\sigma^2}{n}$. Then the density function for $\bar{X}$ should be $f_{\bar{X}}(\bar{x}) = \sqrt{\frac{n}{2\pi\sigma^2}} e^{-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}}$.

Now let's construct the saddlepoint approximation for the density function of $\bar{X}$. Since the cumulant generating function for a single normally distributed variable is $K_X(t) = \mu t + \frac{1}{2}\sigma^2 t^2$, then $\hat{t}$ satisfies $K'_X\left(\hat{t}\right) = \bar{x}$ resulting that $\hat{t} = \frac{\bar{x}-\mu}{\sigma^2}$. By (2.4.15), the resulting saddlepoint approximation is $\hat{f}_{\bar{x}}(\bar{x}) = \sqrt{\frac{n}{2\pi\sigma^2}} e^{-n\frac{(\bar{x}-\mu)^2}{2\sigma^2}}$. It is same as the result above. □

The original method of proof by Daniels (1954) involves an inversion of the characteristic function using a complex-plane path specially chosen in accord with general saddlepoint techniques. The process at this subsection are standing from the perspective of applied mathematics and in the following subsections we will show a more statistical version of the derivation on the saddlepoint expansion.

2.4.3.2 The Conjugate Exponential Family

We start with the normalization of the probability distribution. In general, an arbitrary function $p(y)$ that serves as the **kernel** of a probability distribution $f(y)$ can be made into the distribution by normalizing $f(y) = \frac{p(y)}{Z}$, where $Z = \int p(y)\, dy$ and is called the partition function or the normalizer.

Example. At section (2.2.1.2), we introduced the family of linear exponential models in canonical form $f(\mathbf{y};\boldsymbol{\theta}) = u(\mathbf{y})e^{\left(\boldsymbol{\theta}'\mathbf{s}(\mathbf{y})-A(\boldsymbol{\theta})\right)} = g(\boldsymbol{\theta})\cdot u(\mathbf{y})\, e^{\boldsymbol{\theta}'\mathbf{s}(\mathbf{y})}$. At this setting, $u(\mathbf{y})\, e^{\boldsymbol{\theta}'\mathbf{s}(\mathbf{y})}$ can be viewed as kernel and partition function is therefore $Z = \int u(\mathbf{y})\, e^{\boldsymbol{\theta}'\mathbf{s}(\mathbf{y})} d\mathbf{y}$. Since $\int f(\mathbf{y};\boldsymbol{\theta})\, d\mathbf{y} = \int g(\boldsymbol{\theta})\, u(\mathbf{y})\, e^{\boldsymbol{\theta}'\mathbf{s}(\mathbf{y})} d\mathbf{y}$ $= g(\boldsymbol{\theta}) \cdot Z = 1$, we obtain $A(\boldsymbol{\theta}) = -\log g(\boldsymbol{\theta}) = \log Z = \log\left(\int u(\mathbf{y})\, e^{\boldsymbol{\theta}'\mathbf{s}(\mathbf{y})} d\mathbf{y}\right)$. $A(\boldsymbol{\theta})$ is therefore called the log-partition function because it is the logarithm of a normalization factor, without which $f(\mathbf{y};\boldsymbol{\theta})$ would not be a probability distribution. □

Suppose we take kernel to be $e^{yt} f_Y(y)$, then partition function will be $\int e^{yt} f_Y(y)\, dy = M_Y(t)$. By doing this, we can define the **conjugate density** or **tilted distribution** $f_Y(y;t)$, where the first argument in the bracket denotes the random variable and the second values at which the density is evaluated.

$$f_Y(y;t) = \frac{e^{yt} f_Y(y)}{\int e^{yt} f_Y(y)\, dy} = \frac{e^{yt} f_Y(y)}{M_Y(t)} = e^{yt - K_Y(t)} f_Y(y)\ . \tag{2.4.18}$$

This definition embeds $f_Y(y)$ in a conjugate exponential family with condition $f_Y(y;0) = f_Y(y)$, and the operation of forming $f_Y(y;t)$ is called exponential tilting (Efron 1981).

In addition, from the conclusion of (2.4.6), we have $K_Y^*(u) = K_Y(t+u) - K_Y(t)$ where $K_Y^*(u)$ is the cumulant generating function of conjugate density. And thus the cumulants and standardized cumulants of conjugate density can be derived as $\kappa_n^* = \left.\frac{\partial^n K_Y^*(u)}{\partial u^n}\right|_{u=0} = \left.\frac{\partial^n K_Y(u+t)}{\partial (u+t)^n}\right|_{u=0} = \frac{\partial^n K_Y(t)}{\partial t^n} = K_Y^{(n)}(t)$ and $\rho_n^* = \frac{\kappa_n^*}{\left(\kappa_2^*\right)^{\frac{n}{2}}} = \frac{K_Y^{(n)}(t)}{\left(K_Y''(t)\right)^{\frac{n}{2}}} = \gamma_n(t)$, respectively.

In addition, let $(Y_1, \ldots, Y_n)$ be a sequence of IID random variables with density function $f_Y(y)$. From (2.4.18), we can obtain the tilted distribution of $S_n = \sum_{j=1}^{n} Y_j$, $f_{S_n}(s;t) = \prod_{j=1}^{n} f_Y(y_j;t) = e^{st - nK_Y(t)} f_{S_n}(s)$ and the tilted distribution of $\bar{Y} = \frac{S_n}{n}$, $f_{\bar{Y}}(\bar{y};t) = f_{S_n}(s;t)\frac{\mathrm{d}s}{\mathrm{d}\bar{y}} = e^{n(\bar{y}t - K_Y(t))} f_{\bar{Y}}(\bar{y})$. Likewise, for IID k-dimensional random vectors $(\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_n)$ from a density $f_{\mathbf{V}}(\mathbf{v})$ on $\mathbb{R}^k$, the conjugate density of $\bar{\mathbf{V}}$ is $f_{\bar{\mathbf{V}}}(\bar{\mathbf{v}};\mathbf{t}) = f_{\bar{\mathbf{V}}}(\bar{\mathbf{v}})\, e^{n\left[\mathbf{t}'\bar{\mathbf{v}} - K_{\mathbf{V}}(\mathbf{t})\right]}$.

2.4.3.3 Saddlepoint Expansion

To obtain an accurate expansion to the density function $f_{\bar{Y}}(\bar{y}) = f_{\bar{Y}}(\bar{y};0)$, we can first get the expansion of $f_{\bar{Y}}(\bar{y};t)$ from applying the Edgeworth ex-

pansion to the conjugate density $f_W(w;t)$, where $w = \frac{\bar{y}-\kappa_1^*}{\sqrt{\frac{\kappa_2^*}{n}}} = \frac{\bar{y}-K_Y'(t)}{\sqrt{\frac{K_Y''(t)}{n}}}$ and $\frac{\mathrm{d}w}{\mathrm{d}\bar{y}}$ is the Jacobian.

$$\begin{aligned}
& f_{\bar{Y}}(\bar{y};t) \\
= \; & f_W(w;t)\frac{\mathrm{d}w}{\mathrm{d}\bar{y}} \\
= \; & \sqrt{\frac{n}{K_Y''(t)}}\cdot \\
& \left\{\phi(w)\left(1+\frac{\rho_3^* H_3(w)}{6n^{\frac{1}{2}}}+\frac{\rho_4^* H_4(w)}{24n}+\frac{(\rho_3^*)^2 H_6(w)}{72n}\right)+O\left(n^{-\frac{3}{2}}\right)\right\} \\
= \; & \sqrt{\frac{n}{K_Y''(t)}}\cdot \\
& \left\{\phi(w)\left(1+\frac{\gamma_3(t) H_3(w)}{6n^{\frac{1}{2}}}+\frac{\gamma_4(t) H_4(w)}{24n}+\frac{(\gamma_3(t))^2 H_6(w)}{72n}\right)\right\} \\
& +O\left(n^{-\frac{3}{2}}\right).
\end{aligned}$$

Second, according to the last comment in section (2.4.2.2), if t is chosen such that w is close to zero or $\bar{y}$ is in the center of the distribution then the error incurred would be $O\left(n^{-1}\right)$ instead of $O\left(n^{-\frac{1}{2}}\right)$. In other words we will choose $t=\hat{t}$, such that $E^*\left[\bar{Y};\hat{t}\right]=\bar{y}$. On the other hand, the log-likelihood function of the tilted distribution of $\bar{Y}$ is $\ell(t;\bar{y}) = nt\bar{y} - nK_Y(t) + a$, where a is a constant and the maximum likelihood estimator of t can be obtained by solving the estimating equation $\ell'\left(\hat{t};\bar{y}\right) = n\bar{y} - nK_Y'\left(\hat{t}\right) = 0$. Since the Rao statistic has $E^*\left[\ell';t\right]=0$, we can obtain $E^*\left[\bar{Y};t\right]=K_Y'(t)$ and thus the maximum likelihood estimator, $\hat{t}$, corresponds to the value of t such that $\bar{y}$ is in the center if the distribution. Also need to note that this statistical interpretation as MLE coincides with the mathematical saddlepoint condition introduced at section (2.4.3.1). Now we estimate the expansion from first step at $\bar{y}=K_Y'\left(\hat{t}\right)$, that is where $w=0$.

$$\begin{aligned}
& f_{\bar{Y}}\left(\bar{y};\hat{t}\right) \\
= \; & \sqrt{\frac{n}{K_Y''\left(\hat{t}\right)}}\cdot \\
& \left\{\phi(0)\left(1+\frac{\gamma_3\left(\hat{t}\right) H_3(0)}{6n^{\frac{1}{2}}}+\frac{\gamma_4\left(\hat{t}\right) H_4(0)}{24n}+\frac{\left(\gamma_3\left(\hat{t}\right)\right)^2 H_6(0)}{72n}\right)+O\left(n^{-\frac{3}{2}}\right)\right\} \\
= \; & \sqrt{\frac{n}{2\pi K_Y''\left(\hat{t}\right)}}\left(1+\frac{\gamma_4\left(\hat{t}\right)}{8n}-\frac{5\left(\gamma_3\left(\hat{t}\right)\right)^2}{24n}+O\left(n^{-2}\right)\right)
\end{aligned}$$

Note that the order of convergence for this step is supposed to be $O\left(n^{-\frac{3}{2}}\right)$. However, all the $n^{-\frac{3}{2}}$ terms at the origin in the edgeworth expansion contain odd degree of Hermite numbers and the resulting zero therefore enhance the order of convergence to $O\left(n^{-2}\right)$.

Finally we obtain the **saddlepoint expansion** (or the tilted Edgeworth expansion) for the probability density function of the sample mean $\bar{Y}$,

$$\begin{aligned} f_{\bar{Y}}(y) &= f_{\bar{Y}}\left(\bar{y};\hat{t}\right)\cdot e^{n\left(K_Y\left(\hat{t}\right)-\bar{y}\hat{t}\right)} \\ &= \sqrt{\frac{n}{2\pi K_Y''\left(\hat{t}\right)}}e^{n\left(K_Y\left(\hat{t}\right)-\bar{y}\hat{t}\right)}\left(1+\frac{\gamma_4\left(\hat{t}\right)}{8n}-\frac{5\left(\gamma_3\left(\hat{t}\right)\right)^2}{24n}\right) \\ &\quad +O\left(n^{-2}\right)\ . \end{aligned} \tag{2.4.19}$$

By the same means, we can also get the saddlepoint expansion for the probability density function of the sample sum S_n. For this case, $w=\frac{s-n\kappa_1^*}{\sqrt{n\kappa_2^*}}=\frac{s-nK_Y'(t)}{\sqrt{nK_Y''(t)}}$. Also choosing $t=\hat{t}$ we have $E^*\left[S_n;\hat{t}\right]=s$ which leads to $K_Y'\left(\hat{t}\right)=\frac{s}{n}$.

$$f_{S_n}(s)=\sqrt{\frac{1}{2\pi nK_Y''\left(\hat{t}\right)}}e^{nK_Y\left(\hat{t}\right)-s\hat{t}}\left(1+\frac{\gamma_4\left(\hat{t}\right)}{8n}-\frac{5\left(\gamma_3\left(\hat{t}\right)\right)^2}{24n}\right)+O\left(n^{-2}\right)\ . \tag{2.4.20}$$

An important feature of (2.4.19) and (2.4.20) is that these formulas hold generally for large deviation regions in the form of $|s-nE[Y]|\leq bn$ for fixed b, and in certain cases even for all s or $\bar{y}$ (Daniels 1954, Jensen 1988 and Barndorff-Nielsen and Cox 1989).

Kolassa (2006 Ch.4) displayed general form for the saddlepoint method which can be seen as a correspondence to (2.4.14). Suppose we want to approximate the density function $f_U(u)$ and the random variable U may be any targeted variable such as sample mean or sample sum. The corresponding cumulant generating function is $K_U(t)$ which can be derived from the cumulant generating function of the original model. For example, the cumulant generating function for $\bar{Y}$ is $nK_Y\left(\frac{t}{n}\right)$. Thus the exponential tilting can be written as $f_U(u;t)=e^{tu-K_U(t)}f_U(u)$. Saddlepoint is set as $\hat{t}$ such that $K_U'\left(\hat{t}\right)=u$ and we obtain $f_U(u)=\sqrt{\frac{1}{2\pi K_U''\left(\hat{t}\right)}}e^{K_U\left(\hat{t}\right)-\hat{t}u}+O\left(n^{-1}\right)$.

For the vector form of the saddlepoint expansion, see Barndorff-Nielsen and Cox (1989).

2.4.3.4 Normalized Saddlepoint Approximation

Note that the leading terms of (2.4.19) and (2.4.20) in saddlepoint expansion are exactly the same as the saddlepoint approximation of (2.4.15) and (2.4.16). According to Barndorff-Nielsen and Cox (1989 p.107) and Reid (1996), these leading terms, considered as functions of $\bar{y}$ or s, are in gen-

eral not exactly normalized, i.e. does not integrate to one. This raises the possibility that we modify (2.4.15) and (2.4.16) to

$$f_{\bar{Y}}(\bar{y}) = c_n \sqrt{\frac{n}{2\pi K_Y''(\hat{t})}} e^{n[K_Y(\hat{t}) - \hat{t}\bar{y}]} + O\left(n^{-1}\right) , \tag{2.4.21}$$

$$f_{S_n}(s) = c_n \sqrt{\frac{1}{2\pi n K_Y''(\hat{t})}} e^{nK_Y(\hat{t}) - s\hat{t}} + O\left(n^{-1}\right) , \tag{2.4.22}$$

where c_n is chosen to normalize the leading terms.

A necessary and sufficient condition that normalization produces a uniform improvement from $O\left(n^{-1}\right)$ to $O\left(n^{-\frac{3}{2}}\right)$, i.e. from second order to third order approximation, is that the n^{-1} term $\left(\frac{\gamma_4(\hat{t})}{8n} - \frac{5(\gamma_3(\hat{t}))^2}{24n}\right)$ is constant for all members of the exponential family associated with $f_Y(y)$. The reason is that, if the n^{-1} term is constant, i.e. does not depend on $\hat{t}$, hence on s or $\bar{y}$, then that term will be absorbed into the normalizing constant. In the one-dimensional case, there are just three families for which the n^{-1} term does not depend on $\hat{t}$: the normal, gamma, and inverse Gaussian (Blaesild and Jensen 1985).

In great majority of cases, however, the n^{-1} term will vary with s or $\bar{y}$ and the resulting normalization will in effect depend on the n^{-1} term in the central region of the distribution and will produce an error that is $O\left(n^{-\frac{3}{2}}\right)$ only in the normal deviation region, i.e. $|s - nE[Y]| \leq c\sqrt{n}$ for some fixed constant c.

A detailed discussion of the error properties of the saddlepoint approximation is given in Kolassa (2006, Ch4).

Even though first introduced by Daniels in 1954, and afterwards enhanced by De Bruijn (1970) and Bleistein and Handelsman (1975), the saddlepoint approximation was not well recognized until a paper by Barndorff-Nielsen and Cox (1979) discussed it for general statistical applications. A thorough review and detailed discussion of it were given by Barndorff-Nielsen (1986a, 1991), Daniels (1987), Reid (1988, 1996), Barndorff-Nielsen and Cox (1989, Ch4), and Jensen (1995).

In particular, Reid (1996) stated the comparison result between Edgeworth approximation and saddlepoint approximation to $\bar{Y}$. In both approximation, the relative errors are not uniform in $\bar{y}$, the value at which the density function is evaluated. But an advantage of the saddlepoint approximation beyond its improved relative error is that the error is nearly uniform in $\bar{y}$, which means that in practice the relative errors remain small far out in the tails. The two-term Edgeworth expansion for the density of $\bar{Y}$, by contrast, has a relative error of $O\left(n^{-\frac{3}{2}}\right)$ that fluctuates substantially with $\bar{y}$, and in finite samples the Edgeworth approximation tends to perform poorly in the tails of the distribution. In addition, Fraser (1988)

mentioned that saddlepoint approximation is a good approximation for the normal like case, but may have a poor accuracy for the case far from normal such as the uniform distribution.

2.4.3.5 Saddlepoint Approximation to the Cumulative Distribution Function

For practical use in statistical inference, we are more often interested in approximating the cumulative distribution of a statistic in order to compute p-values or confidence intervals. The cumulative distribution function for a continuous random variable Y is defined as $F_Y(y) = \int_{-\infty}^{y} f_Y(s)\,ds$, however, numerically integrating the saddlepoint approximation of the density function of Y will introduce significant error for the cases that an analytical solution does not exist.

Suppose a real continuous random variable Y has cumulative distribution function $F_Y(y)$ and cumulant generating function $K_Y(t)$. Lugannani and Rice (1980) applied the saddlepoint methods to approximate $F_Y(y)$, obtaining the **Lugannani-Rice formula**:

$$F_Y(y) = \Phi(R) + \phi(R)\left(\frac{1}{R} - \frac{1}{Q}\right) + O\left(n^{-\frac{3}{2}}\right), \tag{2.4.23}$$

where

$$R = \operatorname{sgn}\left(\hat{t}\right)\sqrt{2\left(\hat{t}y - K_Y\left(\hat{t}\right)\right)}, \tag{2.4.24}$$

$$Q = \hat{t}\sqrt{K_Y''\left(\hat{t}\right)}, \tag{2.4.25}$$

and $\hat{t} = \hat{t}(y)$ is the saddlepoint satisfying $K_Y'\left(\hat{t}\right) = y$**.** Daniels (1987) and Butler (2007 P.12) showed that (2.4.23) works only under $Y \neq E[Y]$, because when it is in the **singularity condition**, that is $Y = E[Y]$ or $\hat{t} = 0$, we have $R = Q = 0$ and the approximation breaks down with the last factor in (2.4.23) undefined. As $Y \to E[Y]$, the approximation (2.4.23) should be replaced by the limiting value of (2.4.23).

$$\hat{F}_Y(y) = \frac{1}{2} + \frac{\rho_3}{6\sqrt{2\pi}} = \frac{1}{2} + \frac{1}{6\sqrt{2\pi}}\frac{K_Y'''(0)}{K_Y''(0)^{\frac{3}{2}}}. \tag{2.4.26}$$

Thus, the entire expression is now continuous and, more generally, continuously differentiable or "smooth". Apart from the theoretical smoothness, any practical computation that uses software is vulnerable to numerical instability when making the computation of $F_Y(y)$ in the neighborhood of $E[Y]$.

Assume that there exists a sample of IID random variables $Y_1, \ldots, Y_n$, and Lugannani and Rice approximation can also be applied to $F_{\bar{Y}}(\bar{y})$, the cumulative distribution function of $\bar{Y}$, with R and Q taking the following

forms

$$R = \text{sgn}\left(\hat{t}\right)\sqrt{2n\left(\hat{t}\bar{y} - K_Y\left(\hat{t}\right)\right)}\,,$$

and

$$Q = \hat{t}\sqrt{nK_Y''\left(\hat{t}\right)}\,.$$

Similar to the discussion in last subsection, (2.4.23) has third-order accuracy, $O\left(n^{-\frac{3}{2}}\right)$, in a moderate deviation region that is sequences of sets of bounded central tendency, or sets of $\bar{Y}$ for which $\sqrt{n}\left(\bar{Y} - E\left[\bar{Y}\right]\right)$ remains bounded with increasing n, i.e. $\sqrt{n}|\bar{Y} - E\left[\bar{Y}\right]| \leq c$ for fixed c. However, it will be second-order accuracy, $O\left(n^{-1}\right)$, when the values of $\bar{Y}$ are in large deviation region that is sequences of sets for which $\bar{Y} - E\left[\bar{Y}\right]$ remains bounded as $n \to \infty$, i.e. $|\bar{Y} - E\left[\bar{Y}\right]| \leq b$ for fixed b (Butler 2007 P.53).

Daniels (1987) discussed two explicit approximation formulae for the tail probability of a sample mean, which are shown to arise from different approaches to the saddlepoint approximation. The first is the classical one based on the Edgeworth expansion of the exponentially shifted density recentred at the mean. The second is just the Lugannani & Rice formula which controls the relative error uniformly over the whole range of the mean. The derivation of (2.4.23) using saddlepoint techniques is reviewed and numerical comparisons with the exact tail probabilities for typical continuous and discrete distribution are made in this paper. In addition, uniformity properties of the approximation are discussed in Jensen (1988, 1995). Accuracy in a variety of simple bootstrap and randomization applications is discussed in Davison and Hinkley (1988).

Alternatively Barndorff-Nielsen(1986a;1990) proposed an alternative approximation that incorporates the correction term into the quartile of the normal cumulative distribution, giving the **Barndorff-Nielsen formula**:

$$F_Y(y) = \Phi(R^*) + O\left(n^{-\frac{3}{2}}\right)\,, \tag{2.4.27}$$

where

$$R^* = R - \frac{1}{R}\log\frac{R}{Q}\,, \tag{2.4.28}$$

where R and Q are same defined as that in the Lugannani and Rice Approximation.

Since R and R^* are asymptotically monotone functions of Y, the approximations (2.4.23) and (2.4.27) also give approximations to the distribution of R and R^*. It is interesting to note that Barndorff-Nielsen's method adjusts R such that $\Phi(R^*)$ is close to the exact cumulative function; whereas the Lugannani and Rice's method adjusts $\Phi(R)$ such that it is close to the true cumulative distribution function. The equivalence of these two methods to third-order accuracy can be verified by expanding R^* about R and was rigorously established in Jensen (1992) for exponential family models. However, a shortcoming with Lugannani and Rice approximation is that it

may result in having the calculated cumulative distribution function outside the allowable $[0,1]$ range, while Barndorff-Nielsen formula avoids so.

Example. To illustrate the accuracy of the saddlepoint approximation, we consider the Gamma distribution for sample size $n=1$. For $y \in (0,\infty)$, the basic functions for Gamma(α,β) distribution are

$$f_Y(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} , \tag{2.4.29}$$

$$F_Y(y) = \frac{\gamma(\alpha,\beta y)}{\Gamma(\alpha)} , \tag{2.4.30}$$

$$K_Y(t) = -\alpha \log\left(1-\frac{t}{\beta}\right) , \tag{2.4.31}$$

where the shape parameter $\alpha > 0$ and the rate parameter $\beta > 0$ and $t < \beta$. In addition, $\Gamma(s) = \int_0^\infty t^{s-1}e^{-t}dt$ $(\mathrm{Re}(s)>0)$ is the ordinary gamma function. The upper incomplete gamma function is defined as $\Gamma(s,x) = \int_x^\infty t^{s-1}e^{-t}dt$ $(\mathrm{Re}(s)>0)$; whereas the lower incomplete gamma function is defined as $\gamma(s,x) = \int_0^x t^{s-1}e^{-t}dt$ $(\mathrm{Re}(s)>0)$. From the definitions, we have that $\Gamma(s) = \Gamma(s,0)$ and $\gamma(s,x)+\Gamma(s,x) = \Gamma(s)$.

Given the cumulant generating function listed above, we can obtain the saddlepoint $\hat{t}$ by solving $K_Y'(\hat{t}) = y$. Hence $\hat{t} = \hat{t}(y) = \beta - \frac{\alpha}{y}$. Therefore

$$R = \mathrm{sgn}(\hat{t})\sqrt{2(\hat{t}y - K_Y(\hat{t}))} = \mathrm{sgn}(\beta y-\alpha)\sqrt{2\left(\beta y-\alpha+\alpha\log\frac{\alpha}{\beta y}\right)} , \tag{2.4.32}$$

$$Q = \hat{t}\sqrt{K_Y''(\hat{t})} = \frac{\beta y-\alpha}{\sqrt{\alpha}} .$$

With these values, expressions (2.4.23) and (2.4.27) are explicit in y and yield very simple approximations to the gamma cumulative distribution function.

Table (2.1) compares different approximations of cumulative distribution function with the exact one for the very skewed Gamma distribution, $Gamma\left(\frac{1}{2},1\right)$. In particular, signed log-likelihood approximation which is a first order approximation is set here as a contrast, illustrating the accuracy of Lugannani and Rice approximation and Barndorff-Nielsen approximation which are both third order approximation. In the parenthesis, the percentage relative errors are presented and we can find that relative errors of (2.4.23) and (2.4.27) remain very small for the computations in the right tail and accuracy appears quite good but not quite as accurate in the left tail.

For larger values of α with $\beta = 1$, the gamma distribution approaches a normal shape. Since the approximation is exact in the normal setting which is the limit for large α according to the example at subsection (2.4.3.1),

one might expect the saddlepoint approximation to gain in accuracy with increasing α. This can be seen numerically by reproducing Table (2.1) with $\alpha = 2$ in Table (2.2).

The rate of convergence for relative error is $O\left(\alpha^{-\frac{3}{2}}\right)$ as $\alpha \to \infty$ for a fixed value of the standardized variable $Z = \frac{Y-\alpha}{\sqrt{\alpha}}$ (Daniels, 1987). This $O\left(\alpha^{-\frac{3}{2}}\right)$ relative error statement means that $\alpha^{\frac{3}{2}}\left\{\frac{\hat{F}\left(\sqrt{\alpha}z+\alpha\right)}{F\left(\sqrt{\alpha}z+\alpha\right)}-1\right\}$ remains bounded as $\alpha \to \infty$ for fixed Z.

Finally, to illustrate the accuracy of the saddlepoint approximation methods, Figure (2.1) and Figure (2.2) plotted the exact cumulative distribution function of $Gamma\left(\frac{1}{2},1\right)$ and $Gamma\left(2,1\right)$ along with two third-order approximations given by Lugannani and Rice (2.4.23) and Barndorff-Nielsen (2.4.27) and one first-order signed log-likelihood approximation. The plots illustrates that the two forms of third-order approximation and the exact cumulative distribution function are indistinguishable. Actually Table (2.1) and Table (2.2) recorded some selected values from these two figure.

□

2.5 The p^* Formula

A feature of the saddlepoint approximation is that if the underlying distribution is a member of an exponential family, then the saddlepoint approximation can be reformulated entirely in terms of the likelihood function. In its likelihood formulation, this approximation can be shown to apply to more general families of distributions, where it provides an approximation to the density of the maximum likelihood estimator.

Suppose that the distribution of $\mathbf{Y}$ is in the family of linear exponential model in canonical form. By marginalizing from $\mathbf{Y}$ to $\mathbf{S}\left(\mathbf{Y}\right)$, we finally get the density of minimal sufficient statistic $\mathbf{S}$ from the density of $\mathbf{Y}$ at (2.2.8):

$$f_{\mathbf{S}}\left(\mathbf{s};\boldsymbol{\theta}\right) = e^{\boldsymbol{\theta}'\mathbf{s}-A(\boldsymbol{\theta})}h\left(\mathbf{s}\right) \ . \tag{2.5.1}$$

$h\left(\mathbf{s}\right)$ is generally hard to calculate, however, the density of $f_{\mathbf{S}}\left(\mathbf{s};\boldsymbol{\theta}\right)$ can be accurately approximated by the saddlepoint approximation. Before that, we need an important result between the maximum likelihood estimate of the canonical parameter, $\hat{\boldsymbol{\theta}}$, and the saddlepoint, $\hat{\mathbf{t}}$.

Lemma. *The maximum likelihood estimate of the canonical parameter, $\hat{\boldsymbol{\theta}}$, and the saddlepoint, $\hat{\mathbf{t}}$ are linked by the following result:*

$$\hat{\boldsymbol{\theta}} = \hat{\mathbf{t}} + \boldsymbol{\theta} \, . \tag{2.5.2}$$

Proof. On the one hand, from (2.5.1) the log likelihood function for $\mathbf{s}$ can be written as $\ell\left(\boldsymbol{\theta}\right) = a + \boldsymbol{\theta}'\mathbf{s} - A\left(\boldsymbol{\theta}\right)$. Thus, the maximum likelihood estimator,

Table 2.1: Tail probabilities of cdf and its approximations for $Gamma\left(\frac{1}{2},1\right)$

Tail	Values of y	Exact Tail Probabilities	Signed Log-likelihood Approximation	Lugannani and Rice Approximation	Barndorff-Nielsen Approximation
Left Tail $P(Y \leq y)$	0.00005	0.0080 (0.00%)	0.0021 (73.90%)	0.0091 (-13.92%)	0.0087 (-9.38%)
	0.0001	0.0113 (0.00%)	0.0031 (72.92%)	0.0128 (-13.61%)	0.0123 (-9.06%)
	0.0005	0.0252 (0.00%)	0.0075 (70.14%)	0.0284 (-12.62%)	0.0273 (-8.13%)
	0.001	0.0357 (0.00%)	0.0112 (68.64%)	0.0400 (-12.03%)	0.0384 (-7.61%)
Right Tail $P(Y \geq y)$	2	0.0455 (0.00%)	0.1020 (-124.14%)	0.0458 (-0.58%)	0.0471 (-3.45%)
	3	0.0143 (0.00%)	0.0366 (-156.08%)	0.0145 (-1.63%)	0.0150 (-4.66%)
	4	0.0047 (0.00%)	0.0133 (-183.68%)	0.0048 (-2.46%)	0.0049 (-5.59%)
	5	0.0016 (0.00%)	0.0048 (-208.39%)	0.0016 (-3.13%)	0.0017 (-6.32%)

Table 2.2: Tail probabilities of cdf and its approximations for $Gamma\left(2,1\right)$

Tail	Values of y	Exact Tail Probabilities	Signed Log-likelihood Approximation	Lugannani and Rice Approximation	Barndorff-Nielsen Approximation
Left Tail $P(Y \leq y)$	0.1	0.0047 (0.00%)	0.0021 (54.81%)	0.0047 (-1.45%)	0.0047 (-0.68%)
	0.5	0.0902 (0.00%)	0.0553 (38.68%)	0.0906 (-0.47%)	0.0902 (0.00%)
	1	0.2642 (0.00%)	0.1897 (28.21%)	0.2647 (-0.16%)	0.2639 (0.14%)
	1.3	0.3732 (0.00%)	0.2849 (23.66%)	0.3735 (-0.09%)	0.3726 (0.15%)
Right Tail $P(Y \geq y)$	4	0.0916 (0.00%)	0.1340 (-46.27%)	0.0917 (-0.17%)	0.0921 (-0.52%)
	5	0.0404 (0.00%)	0.0633 (-56.46%)	0.0405 (-0.26%)	0.0407 (-0.65%)
	6	0.0174 (0.00%)	0.0288 (-65.94%)	0.0174 (-0.34%)	0.0175 (-0.78%)
	7	0.0073 (0.00%)	0.0128 (-74.84%)	0.0073 (-0.42%)	0.0074 (-0.89%)

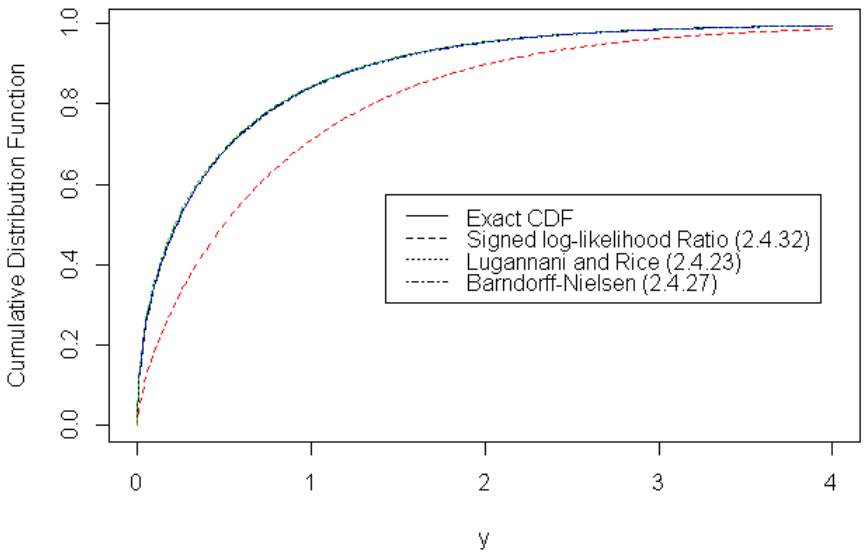


Figure 2.1: Exact and approximated cumulative distribution functions for $Gamma\left(\frac{1}{2}, 1\right)$

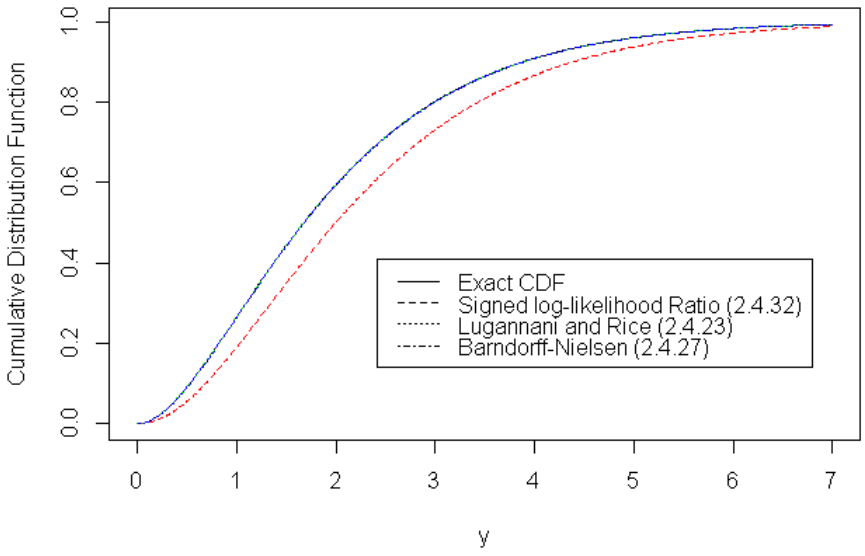


Figure 2.2: Exact and approximated cumulative distribution functions for $Gamma\,(2,1)$

$\hat{\boldsymbol{\theta}}$, satisfies

$$A_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}\right)=\mathbf{s}\,. \tag{2.5.3}$$

And the observed information matrix at $\hat{\boldsymbol{\theta}}$ is $\mathbf{j}\left(\hat{\boldsymbol{\theta}}\right)=-\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}=A_{\boldsymbol{\theta}\boldsymbol{\theta}'}\left(\hat{\boldsymbol{\theta}}\right)$. Note that between the minimum sufficient statistic $\mathbf{S}\left(\mathbf{Y}\right)$ and the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$, the Jacobian matrix is also $\frac{\mathbf{ds}}{\mathbf{d}\hat{\boldsymbol{\theta}}'}=A_{\boldsymbol{\theta}\boldsymbol{\theta}'}\left(\hat{\boldsymbol{\theta}}\right)=\mathbf{j}\left(\hat{\boldsymbol{\theta}}\right)$.

On the other hand, recall that we have $K_{\mathbf{S}}\left(\mathbf{t}\right)=A\left(\boldsymbol{\theta}+\mathbf{t}\right)-A\left(\boldsymbol{\theta}\right)$ as (2.4.6). Therefore, we can obtain

$$K'_{\mathbf{S}}\left(\mathbf{t}\right)=\frac{\partial K_{\mathbf{S}}\left(\mathbf{t}\right)}{\partial\mathbf{t}}=\frac{\partial A\left(\boldsymbol{\theta}+\mathbf{t}\right)}{\partial\left(\boldsymbol{\theta}+\mathbf{t}\right)}=A_{\boldsymbol{\theta}+\mathbf{t}}\left(\boldsymbol{\theta}+\mathbf{t}\right)\,, \tag{2.5.4}$$

and $K''_{\mathbf{S}}\left(\mathbf{t}\right)=\frac{\partial^2 K_{\mathbf{S}}(\mathbf{t})}{\partial\mathbf{t}\partial\mathbf{t}'}=\frac{\partial^2 A(\boldsymbol{\theta}+\mathbf{t})}{\partial(\boldsymbol{\theta}+\mathbf{t})\partial(\boldsymbol{\theta}+\mathbf{t})'}=A_{(\boldsymbol{\theta}+\mathbf{t})(\boldsymbol{\theta}+\mathbf{t})'}\left(\boldsymbol{\theta}+\mathbf{t}\right)$.

Finally, the saddlepoint, $\hat{\mathbf{t}}$, satisfies the saddlepoint equation $K'_{\mathbf{S}}\left(\hat{\mathbf{t}}\right)=\mathbf{s}$. Together with (2.5.3) and (2.5.4), we have $K'_{\mathbf{S}}\left(\hat{\mathbf{t}}\right)=A_{\boldsymbol{\theta}+\hat{\mathbf{t}}}\left(\boldsymbol{\theta}+\hat{\mathbf{t}}\right)=\mathbf{s}=A_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}\right)$ and thus obtain $\hat{\boldsymbol{\theta}}=\hat{\mathbf{t}}+\boldsymbol{\theta}$. □

After collect the preparation above, we can now apply the saddlepoint approximation for (2.5.1) and obtain

$$\begin{aligned}
f_{\mathbf{S}}\left(\mathbf{s};\boldsymbol{\theta}\right)&=\sqrt{\frac{1}{(2\pi)^p\left|K''_{\mathbf{S}}\left(\hat{\mathbf{t}}\right)\right|}}e^{\left(K_{\mathbf{S}}\left(\hat{\mathbf{t}}\right)-\hat{\mathbf{t}}'\mathbf{s}\right)}+O\left(n^{-1}\right)\\
&=\sqrt{\frac{1}{(2\pi)^p\left|A_{(\boldsymbol{\theta}+\mathbf{t})(\boldsymbol{\theta}+\mathbf{t})'}\left(\boldsymbol{\theta}+\hat{\mathbf{t}}\right)\right|}}e^{\left(A\left(\boldsymbol{\theta}+\hat{\mathbf{t}}\right)-A(\boldsymbol{\theta})-\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)'\mathbf{s}\right)}+O\left(n^{-1}\right)\\
&=\sqrt{\frac{1}{(2\pi)^p\left|\mathbf{j}\left(\hat{\boldsymbol{\theta}}\right)\right|}}e^{\left(\ell(\boldsymbol{\theta})-\ell\left(\hat{\boldsymbol{\theta}}\right)\right)}+O\left(n^{-1}\right)\,.
\end{aligned} \tag{2.5.5}$$

Example. Consider the exponential model (2.2.9) in which the sufficient statistic $\mathbf{S}$ is decomposed into two components, one component, S_1, of dimension one and the second component, $\mathbf{S}_2$ of dimension $(p-1)$. Skovgaard (1987) and Fraser and Reid (1993) showed the following extension based on (2.2.9) and (2.5.5).

While the saddlepoint approximation to the density function of $\mathbf{S}$ has the form (2.5.5), the same approximation to the marginal density function of S_2 would have the form:

$$f\left(\mathbf{s}_2;\boldsymbol{\theta}\right)=\sqrt{\frac{1}{(2\pi)^{p-1}\left|\mathbf{j}_{\boldsymbol{\lambda}\boldsymbol{\lambda}'}\left(\hat{\boldsymbol{\theta}}_\psi\right)\right|}}e^{\left(\ell(\boldsymbol{\theta})-\ell\left(\hat{\boldsymbol{\theta}}_\psi\right)\right)}+O\left(n^{-1}\right)\,.$$

This leads to the conditional density function of $\mathbf{S}$ given $\mathbf{S}_2$

$$f\left(\mathbf{s}|\mathbf{s}_2,\psi\right)=c\sqrt{\frac{|\mathbf{j}_{\boldsymbol{\lambda}\boldsymbol{\lambda}'}\left(\hat{\boldsymbol{\theta}}_\psi\right)|}{(2\pi)\,|\mathbf{j}\left(\hat{\boldsymbol{\theta}}\right)|}}e^{\left(\ell\left(\hat{\boldsymbol{\theta}}_\psi\right)-\ell\left(\hat{\boldsymbol{\theta}}\right)\right)}+O\left(n^{-\frac{3}{2}}\right)\text{ ,}$$

where c is a normalizing constant. Skovgaard (1987) called this expression for the conditional density, the **double saddlepoint approximation**. The corresponding likelihood function is given as below

$$\ell(\psi)=\ell(\hat{\boldsymbol{\theta}}_\psi)+\frac{1}{2}\log|\mathbf{j}_{\boldsymbol{\lambda}\boldsymbol{\lambda}'}(\hat{\boldsymbol{\theta}}_\psi)|\text{ .} \tag{2.5.6}$$

It is actually a third order approximation to the likelihood function from the conditional distribution (2.2.13). And inference concerning the interest parameter ψ in the presence of the nuisance parameter $\boldsymbol{\lambda}$ can be constructed from the conditional likelihood function $\ell\left(\psi\right)$ at the observed value $\mathbf{y}^0$. □

Barndorff-Nielsen (1980) shows that (2.5.5) can be transformed to a corresponding approximation for $\hat{\boldsymbol{\theta}}$:

$$\begin{aligned} f_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}};\boldsymbol{\theta}\right) &= f_{\mathbf{S}}\left(\mathbf{s};\boldsymbol{\theta}\right)\left|\frac{\mathbf{ds}}{\mathbf{d}\hat{\boldsymbol{\theta}}'}\right| \\ &= (2\pi)^{-\frac{p}{2}}\,|\mathbf{j}\left(\hat{\boldsymbol{\theta}}\right)|^{\frac{1}{2}}e^{\left(\ell(\boldsymbol{\theta})-\ell\left(\hat{\boldsymbol{\theta}}\right)\right)}+O\left(n^{-1}\right) \\ &= (2\pi)^{-\frac{p}{2}}\,|\mathbf{j}\left(\hat{\boldsymbol{\theta}}\right)|^{\frac{1}{2}}\frac{\mathcal{L}\left(\boldsymbol{\theta}\right)}{\mathcal{L}\left(\hat{\boldsymbol{\theta}}\right)}+O\left(n^{-1}\right)\text{ .} \end{aligned} \tag{2.5.7}$$

In addition, Durbin (1980) shows that, similar to subsection (2.4.3.4), a renormalizing process, $\int_{\hat{\boldsymbol{\theta}}}f_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}};\boldsymbol{\theta}\right)\mathbf{d}\hat{\boldsymbol{\theta}}=1$, can be conducted to (2.5.7) and we thus get the normalizing constant $c\left(\boldsymbol{\theta}\right)=\left(\int_{\hat{\boldsymbol{\theta}}}|\mathbf{j}\left(\hat{\boldsymbol{\theta}}\right)|^{\frac{1}{2}}e^{\left(\ell(\boldsymbol{\theta})-\ell\left(\hat{\boldsymbol{\theta}}\right)\right)}\mathbf{d}\hat{\boldsymbol{\theta}}\right)^{-1}$. This renormalizing process make the error term of the saddlepoint approximation reduce to $O\left(n^{-\frac{3}{2}}\right)$ and we obtain a more accurate approximation for the density of the maximum likelihood estimator.

$$\begin{aligned} f_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}};\boldsymbol{\theta}\right) &= c\left(\boldsymbol{\theta}\right)|\mathbf{j}\left(\hat{\boldsymbol{\theta}}\right)|^{\frac{1}{2}}e^{\left(\ell(\boldsymbol{\theta})-\ell\left(\hat{\boldsymbol{\theta}}\right)\right)}+O\left(n^{-\frac{3}{2}}\right) \\ &= c\left(\boldsymbol{\theta}\right)|\mathbf{j}\left(\hat{\boldsymbol{\theta}}\right)|^{\frac{1}{2}}\frac{\mathcal{L}\left(\boldsymbol{\theta}\right)}{\mathcal{L}\left(\hat{\boldsymbol{\theta}}\right)}+O\left(n^{-\frac{3}{2}}\right)\text{ .} \end{aligned} \tag{2.5.8}$$

Barndorff-Nielsen (1980) shows that for exponential models (2.5.8) actually work in the same way as the saddlepoint expansion for the density of $\hat{\boldsymbol{\theta}}$. In addition, this expression is invariant under reparameterization: $\boldsymbol{\theta}$ need not be the canonical parameter (Fraser 1990). This approximation also has

been found to be extremely accurate in the general context for a general statistical model $f(\mathbf{y};\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ has dimension p (Barndorff-Nielsen 1983).

If the dimension of $\mathbf{S}$ is greater than p, then $f_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}};\boldsymbol{\theta}\right)$ needs to be interpreted as a conditional density $f_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}|\mathbf{a};\boldsymbol{\theta}\right)$ given some exact or approximate ancillary $\mathbf{A}=\mathbf{A}(\mathbf{y})$. It has been assumed that there is a one-to-one correspondence between the minimum sufficient statistic $\mathbf{S}(\mathbf{Y})$ and a new statistic $\left(\hat{\boldsymbol{\theta}},\mathbf{A}\right)$, according to subsection (2.2.2.2) as well as Barndorff-Nielsen (1980). Thus we can write the log-likelihood as a function of $\left(\hat{\boldsymbol{\theta}},\mathbf{A}\right)$ in its dependence on the data; that is, $\ell(\boldsymbol{\theta})=\ell\left(\boldsymbol{\theta};\hat{\boldsymbol{\theta}},\mathbf{a}\right)$; and obtain

$$\begin{aligned} f_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}|\mathbf{a};\boldsymbol{\theta}\right) &= c(\boldsymbol{\theta},\mathbf{a})\,|\mathbf{j}\left(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}},\mathbf{a}\right)|^{\frac{1}{2}}e^{\left(\ell\left(\boldsymbol{\theta};\hat{\boldsymbol{\theta}},\mathbf{a}\right)-\ell\left(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}},\mathbf{a}\right)\right)}+O\left(n^{-\frac{3}{2}}\right) \\ &= c(\boldsymbol{\theta},\mathbf{a})\,|\mathbf{j}\left(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}},\mathbf{a}\right)|^{\frac{1}{2}}\frac{\mathcal{L}\left(\boldsymbol{\theta};\hat{\boldsymbol{\theta}},\mathbf{a}\right)}{\mathcal{L}\left(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}},\mathbf{a}\right)}+O\left(n^{-\frac{3}{2}}\right)\ , \end{aligned} \tag{2.5.9}$$

where $c(\boldsymbol{\theta},\mathbf{a})$ is a normalizing constant in a sense similar to above.

This p^{*} formula holds quite generally, subject to specification of an approximate ancillary (Barndorff-Nielsen 1980, 1983). The usual presentation of the formula, however, does not include a general prescription for determining the ancillary $\mathbf{A}(\mathbf{Y})$, and thus it does not lead directly to a plot of the density of $\hat{\boldsymbol{\theta}}$ for particular $\boldsymbol{\theta}$ values, unless $\hat{\boldsymbol{\theta}}$ is minimal sufficient or the ancillary $\mathbf{A}(\mathbf{Y})$ is otherwise available. An affine ancillary has been suggested by Barndorff-Nielsen (1980) on asymptotic grounds. A computer implementable procedure for calculating a preferred ancillary $\mathbf{A}(\mathbf{Y})$ based on differential likelihood is discussed by Fraser and Reid (1988).

In the special case that $f(\mathbf{y};\boldsymbol{\theta})$ is an exponential family model, the sufficient statistic $\mathbf{S}(\mathbf{Y})$ is a one-to-one function of $\hat{\boldsymbol{\theta}}$; thus the likelihood function depends on the data only through $\hat{\boldsymbol{\theta}}$. In this case (2.5.9) coincides with the approximation to $f_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}};\boldsymbol{\theta}\right)$ with no conditioning involved and it is back to (2.5.8). Another important feature of this formula is that it gives the exact conditional density in the case of transformation model (Fraser 1968 Ch2; Barndorff-Nielsen 1980). It is because in the special case that $f(\mathbf{y};\boldsymbol{\theta})$ is a transformation model, a maximal ancillary exists and the factorization (2.2.10) holds; and with $\mathbf{A}$ fixed, $\hat{\boldsymbol{\theta}}$ is a one-to-one function of $\mathbf{S}$.

Some discussion and analysis of the p^{*} formula is given by Barndorff-Nielsen (1980,1983,1986b, 2012), McCullagh(1984) and Reid (1988); they use asymptotic calculations based on sample space geometry and cumulants or use transformation model theory. An alternative interpretation of the approximation using the Laplace-integral method is discussed by Fraser (1988): a transformation of $\boldsymbol{\theta}$ and of $\hat{\boldsymbol{\theta}}$ is defined to yield constant observed information and an approximating exponential model then sup-

ports a local saddlepoint calculation. Detailed discussion and review of the literature are given in Reid (1988, 1995) and Barndorff-Nielsen and Cox (1994 Ch6).

Example. The sufficient statistic of a set of IID data observations is simply the sum of individual sufficient statistics, and encapsulates all the information needed to describe the posterior distribution of the parameters, given the data. Thus it is useful to know a good approximation of the following densities of minimal sufficient statistic from the exponential family, $f_{\mathbf{w}}(\mathbf{w};\boldsymbol{\theta}) = e^{\left(\boldsymbol{\theta}'\mathbf{w}-nA(\boldsymbol{\theta})\right)}g(\mathbf{w})$ and $f_{\bar{\mathbf{s}}}(\bar{\mathbf{s}};\boldsymbol{\theta}) = e^{\left(n\boldsymbol{\theta}'\bar{\mathbf{s}}-nA(\boldsymbol{\theta})\right)}v(\bar{\mathbf{s}})$, where $\mathbf{W}(\mathbf{Y}) = \sum_{i=1}^{n}\mathbf{S}(\mathbf{Y}_i)$ and $\bar{\mathbf{S}}(\mathbf{Y}) = \frac{1}{n}\mathbf{W}(\mathbf{Y})$ for IID random vectors $\mathbf{Y}_i$.

The saddlepoint approximation for $f_{\mathbf{w}}(\mathbf{w};\boldsymbol{\theta})$ and the corresponding p^* formula are

$$f_{\mathbf{w}}(\mathbf{w};\boldsymbol{\theta}) = \sqrt{\frac{1}{(2\pi)^p\left|\mathbf{j}^{\mathbf{w}}\left(\hat{\boldsymbol{\theta}}\right)\right|}}e^{\left(\ell^{\mathbf{w}}(\boldsymbol{\theta})-\ell^{\mathbf{w}}(\hat{\boldsymbol{\theta}})\right)} + O\left(n^{-1}\right),$$

$$f_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}};\boldsymbol{\theta}\right) = c(\boldsymbol{\theta})\left|\mathbf{j}^{\mathbf{w}}\left(\hat{\boldsymbol{\theta}}\right)\right|^{\frac{1}{2}}e^{\left(\ell^{\mathbf{w}}(\boldsymbol{\theta})-\ell^{\mathbf{w}}(\hat{\boldsymbol{\theta}})\right)} + O\left(n^{-\frac{3}{2}}\right),$$

$$f_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}|\mathbf{a};\boldsymbol{\theta}\right) = c(\boldsymbol{\theta},\mathbf{a})\left|\mathbf{j}^{\mathbf{w}}\left(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}},\mathbf{a}\right)\right|^{\frac{1}{2}}e^{\left(\ell^{\mathbf{w}}(\boldsymbol{\theta};\hat{\boldsymbol{\theta}},\mathbf{a})-\ell^{\mathbf{w}}(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}},\mathbf{a})\right)} + O\left(n^{-\frac{3}{2}}\right),$$

where $\ell^{\mathbf{w}}(\boldsymbol{\theta}) = a + \boldsymbol{\theta}'\mathbf{w} - nA(\boldsymbol{\theta})$ and the MLE, $\hat{\boldsymbol{\theta}}$, satisfies $nA_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}\right) = \mathbf{w}$ and $\mathbf{j}^{\mathbf{w}}\left(\hat{\boldsymbol{\theta}}\right) = nA_{\boldsymbol{\theta}\boldsymbol{\theta}'}\left(\hat{\boldsymbol{\theta}}\right)$.

In addition, the saddlepoint approximation for $f_{\bar{\mathbf{s}}}(\bar{\mathbf{s}};\boldsymbol{\theta})$ and the corresponding p^* formula are

$$f_{\bar{\mathbf{s}}}(\bar{\mathbf{s}};\boldsymbol{\theta}) = \sqrt{\frac{1}{(2\pi)^p\left|\mathbf{j}^{\bar{\mathbf{s}}}\left(\hat{\boldsymbol{\theta}}\right)\right|}}e^{\left(\ell^{\bar{\mathbf{s}}}(\boldsymbol{\theta})-\ell^{\bar{\mathbf{s}}}(\hat{\boldsymbol{\theta}})\right)} + O\left(n^{-1}\right),$$

$$f_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}};\boldsymbol{\theta}\right) = c(\boldsymbol{\theta})\left|\mathbf{j}^{\bar{\mathbf{s}}}\left(\hat{\boldsymbol{\theta}}\right)\right|^{\frac{1}{2}}e^{\left(\ell^{\bar{\mathbf{s}}}(\boldsymbol{\theta})-\ell^{\bar{\mathbf{s}}}(\hat{\boldsymbol{\theta}})\right)} + O\left(n^{-\frac{3}{2}}\right),$$

$$f_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}|\mathbf{a};\boldsymbol{\theta}\right) = c(\boldsymbol{\theta},\mathbf{a})\left|\mathbf{j}^{\bar{\mathbf{s}}}\left(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}},\mathbf{a}\right)\right|^{\frac{1}{2}}e^{\left(\ell^{\bar{\mathbf{s}}}(\boldsymbol{\theta};\hat{\boldsymbol{\theta}},\mathbf{a})-\ell^{\bar{\mathbf{s}}}(\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}},\mathbf{a})\right)} + O\left(n^{-\frac{3}{2}}\right),$$

where $\ell^{\bar{\mathbf{s}}}(\boldsymbol{\theta}) = a + n\boldsymbol{\theta}'\bar{\mathbf{s}} - nA(\boldsymbol{\theta})$ and the MLE, $\hat{\boldsymbol{\theta}}$, satisfies $A_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}\right) = \bar{\mathbf{s}}$ and $\mathbf{j}^{\bar{\mathbf{s}}}\left(\hat{\boldsymbol{\theta}}\right) = nA_{\boldsymbol{\theta}\boldsymbol{\theta}'}\left(\hat{\boldsymbol{\theta}}\right)$.

□

2.6 Third-Order Likelihood Inference for a Scalar Parameter of Interest of a General Statistical Model

Theoretically, the third-order likelihood inference for a scalar parameter of interest is based on the third-order approximation on p-value function by using two methods, Lugannani and Rice, and Barndorff-Nielsen. The strength of these two methods are their prominent accuracy for small or medium data size and applicability on testing any parameter of interest. However, the associated limitation is the requirement on the existence of the complete likelihood function or an approximation to the complete likelihood function at least.

2.6.1 Single Parameter Model

Let us first discuss a simple situation where both data and the full parameter are essentially one dimensional. The applications, however, may incorporate the models that permit a reduction of the data to a one dimensional sufficient statistic. In exponential models the reduction is by means of minimal sufficiency, while in location models the reduction is by means of ancillarity. For information on dimension reduction, see section (2.2).

2.6.1.1 Canonical Exponential Family Model with Single Parameter

In section (2.4.3.5), we introduced the saddlepoint approximation to the cumulative distribution function, and in this section we start with its likelihood formulation relative to exponential families, which turns out to apply much more generally.

Lemma. *When the underlying density for random variable Y arises from univariate canonical exponential family with single parameter, the likelihood formulation of statistics R and Q originated from (2.4.24) and (2.4.25) becomes*

$$R = R(\theta) = sgn(\hat{\theta} - \theta)\sqrt{2\left(\ell(\hat{\theta}) - \ell(\theta)\right)}\,. \tag{2.6.1}$$

This is actually the signed log-likelihood ratio statistic.

$$Q = Q(\theta) = (\hat{\theta} - \theta)\sqrt{j\left(\hat{\theta}\right)} = q\,. \tag{2.6.2}$$

This is actually the Wald statistic.

Proof. To obtain (2.6.1), we start from (2.4.24). In addition, the derivation will need the following conclusions: $\hat{\boldsymbol{\theta}} = \hat{\mathbf{t}} + \boldsymbol{\theta}$ at (2.5.2), $K(\mathbf{t}) = A(\boldsymbol{\theta} + \mathbf{t}) -$

$A(\boldsymbol{\theta})$ at (2.4.6) and $\ell(\boldsymbol{\theta}) = a + \boldsymbol{\theta}'\mathbf{y} - A(\boldsymbol{\theta})$ from (2.5.1).

$$\begin{aligned} R &= \operatorname{sgn}(\hat{t})\sqrt{2\left(\hat{t}y - K_Y(\hat{t})\right)} \\ &= \operatorname{sgn}\left(\hat{\theta}-\theta\right)\sqrt{2\left\{\left(\hat{\theta}-\theta\right)y - \left(A(\theta+\hat{t}) - A(\theta)\right)\right\}} \\ &= \operatorname{sgn}\left(\hat{\theta}-\theta\right)\sqrt{2\left\{\left(\hat{\theta}y - A\left(\hat{\theta}\right)\right) - \left(\theta y - A(\theta)\right)\right\}} \\ &= \operatorname{sgn}\left(\hat{\theta}-\theta\right)\sqrt{2\left(\ell\left(\hat{\theta}\right) - \ell(\theta)\right)}. \end{aligned}$$

To obtain (2.6.2), we start from (2.4.25). In addition, the derivation will need the following conclusions: $\hat{\boldsymbol{\theta}} = \hat{\mathbf{t}} + \boldsymbol{\theta}$ at (2.5.2) and $K''(\mathbf{t}) = A_{(\boldsymbol{\theta}+\mathbf{t})(\boldsymbol{\theta}+\mathbf{t})'}(\boldsymbol{\theta}+\mathbf{t})$ from (2.4.6) and $A_{\boldsymbol{\theta}\boldsymbol{\theta}'}\left(\hat{\boldsymbol{\theta}}\right) = \mathbf{j}\left(\hat{\boldsymbol{\theta}}\right)$ from (2.5.3).

$$Q = \hat{t}\sqrt{K_Y''(\hat{t})} = \left(\hat{\theta}-\theta\right)\sqrt{A''(\theta+\hat{t})} = \left(\hat{\theta}-\theta\right)\sqrt{A''\left(\hat{\theta}\right)} = \left(\hat{\theta}-\theta\right)\sqrt{j\left(\hat{\theta}\right)} = q.$$

□

Hence, the approximations of p-value function of θ with relative error $O\left(n^{-\frac{3}{2}}\right)$ can be expressed in the Lugannani and Rice type formula (2.4.23) and the Barndorff-Nielsen type formula (2.4.27).

$$p(\theta) = F\left(\hat{\theta},\theta\right) = F(y,\theta) = \Phi(R) + \phi(R)\left(\frac{1}{R} - \frac{1}{Q}\right) + O\left(n^{-\frac{3}{2}}\right), \quad (2.6.3)$$

$$p(\theta) = F\left(\hat{\theta},\theta\right) = F(y,\theta) = \Phi(R^*) + O\left(n^{-\frac{3}{2}}\right) = \Phi\left(R - \frac{1}{R}\log\frac{R}{Q}\right) + O\left(n^{-\frac{3}{2}}\right), \quad (2.6.4)$$

where R^* is the same defined in (2.4.28) and is generally referred to as the modified signed log-likelihood ratio statistic. Barndorff-Nielsen (1990 , 1990b, 1991) discusses the derivation of (2.6.4) from the p^* approximation. In addition, a positive one-to-one relationship $\hat{\theta} = \hat{\theta}(y)$ is the saddlepoint satisfying $A'\left(\hat{\theta}\right) = y$, and $A''\left(\hat{\theta}\right) > 0$.

It is interesting to note that both R and Q have standard normal distributions to the first order, so the modification and correction implicit in (2.6.3) and (2.6.4) provides the large improvement from $O\left(n^{-\frac{1}{2}}\right)$ to $O\left(n^{-\frac{3}{2}}\right)$.

Corresponding to the singularity condition $Y = E[Y]$ in (2.4.26), both approximations also have the singularity point $\theta = \hat{\theta}$ here. Daniels (1987) and Reid (1996) derived the limiting value $\hat{p}\left(\hat{\theta}\right)$.

Lemma. *Daniels (1987) and Reid (1996) obtained the following limiting value* $\hat{p}\left(\hat{\theta}\right)$*:*

$$\hat{p}\left(\hat{\theta}\right)=\hat{F}_{\bar{Y}}\left(\bar{y};\hat{\theta}\right)=\frac{1}{2}+\frac{\rho_3}{6\sqrt{2\pi n}}=\frac{1}{2}+\frac{1}{6\sqrt{2\pi n}}\frac{\ell'''\left(\hat{\theta}\right)/n}{\left(j\left(\hat{\theta}\right)/n\right)^{\frac{3}{2}}}.$$

Proof. Our derivation starts from (2.4.26). Recall that at singularity condition, we have $\hat{t}=0$.

$$\begin{aligned}\hat{F}_Y(y) &= \frac{1}{2}+\frac{\rho_3}{6\sqrt{2\pi}}=\frac{1}{2}+\frac{1}{6\sqrt{2\pi}}\frac{K_Y'''(0)}{K_Y''(0)^{\frac{3}{2}}}\\ &= \frac{1}{2}+\frac{1}{6\sqrt{2\pi}}\frac{K_Y'''\left(\hat{t}\right)}{K_Y''\left(\hat{t}\right)^{\frac{3}{2}}}=\frac{1}{2}+\frac{1}{6\sqrt{2\pi}}\frac{A'''\left(\theta+\hat{t}\right)}{A''\left(\theta+\hat{t}\right)^{\frac{3}{2}}}\\ &= \frac{1}{2}+\frac{1}{6\sqrt{2\pi}}\frac{A'''\left(\hat{\theta}\right)}{A''\left(\hat{\theta}\right)^{\frac{3}{2}}}=\frac{1}{2}+\frac{1}{6\sqrt{2\pi}}\frac{\ell'''\left(\hat{\theta}\right)}{j\left(\hat{\theta}\right)^{\frac{3}{2}}}=\hat{p}\left(\hat{\theta}\right).\end{aligned}$$

□

In addition, Fraser, Reid, Li and Wong (2003) also proposed a bridging method for dealing with this singularity problem. However, Reid (1996) pointed out that for many applications it is only of interest to compute p-values in the left or right tail of the distribution, well away from the fifty percent point. Since the aim of this dissertation is just to provide accurate approximations to the two tails of $p(\theta)$, the singularity problem will not be examined.

Example. In the example at subsection (2.4.3.5), we illustrated the saddlepoint approximation to the cumulative distribution function of $Gamma(\alpha,\beta)$ by Lugannani and Rice method and Barndorff-Nielsen method. To illustrate the application and accuracy of the third order likelihood methods discussed up to this point, we continue that example and consider a special case of gamma distribution. In particular, when $\alpha=1$, the distribution $Gamma(1,\theta)$ will just be the exponential distribution with rate parameter θ and its probability density function is $f_Y(y;\theta)=\theta e^{-\theta y}$.

Firstly, the exact p-value function of θ for the exponential distribution at $\hat{\theta}=\frac{n}{\sum_{i=1}^{n}y_i}$ is $1-F_{Gamma(n,\theta)}\left(\sum_{i=1}^{n}y_i\right)$, where $F_{Gamma(n,\theta)}(\cdot)$ is the cumulative distribution function of $Gamma(n,\theta)$ in the form of (2.4.30). It comes from the fact that if $Y_i\sim Exp(\theta)$ (Exponential distribution with rate parameter θ), then the sum $\sum_{i=1}^{n}Y_i\sim Gamma(n,\theta)$; and thus

$$
\begin{aligned}
p(\theta) &= P\left(\hat{\Theta} \leq \hat{\theta}; \theta\right) = P\left(\frac{n}{\sum_{i=1}^{n} Y_i} \leq \frac{n}{\sum_{i=1}^{n} y_i}\right). \\
&= 1 - P\left(\sum_{i=1}^{n} Y_i \leq \sum_{i=1}^{n} y_i\right) = 1 - F_{Gamma(n,\theta)}\left(\sum_{i=1}^{n} y_i\right)
\end{aligned}
$$

Secondly, suppose that a sample $(y_1, \ldots, y_n)$ from $Exp(\theta)$ is available. Then we can obtain the following basic likelihood results: $\ell(\theta) = n \log\theta - \theta \cdot \sum_{i=1}^{n} y_i = n(\log\theta - \theta\bar{y})$, $\ell_\theta(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} y_i$ with $\hat{\theta} = \frac{n}{\sum_{i=1}^{n} y_i} = \frac{1}{\bar{y}}$, $\ell_{\theta\theta}(\theta) = -\frac{n}{\theta^2}$ and $j\left(\hat{\theta}\right) = \frac{n}{\hat{\theta}^2} = \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = n\bar{y}^2$. From these results, we can obtain the following approximations of $p(\theta)$:

(1) $\Phi(R)$, where R is the signed log-likelihood ratio statistic from (2.6.1) $R = R(\theta) = \text{sgn}(\hat{\theta}-\theta)\sqrt{2\left(\ell\left(\hat{\theta}\right) - \ell(\theta)\right)} = \text{sgn}(1-\theta\bar{y})\sqrt{2n(\theta\bar{y} - \log\theta\bar{y} - 1)}$;

(2) $\Phi(Q)$, where Q is the Wald statistic from (2.6.2) and $Q = Q(\theta) = (\hat{\theta} - \theta)\sqrt{j\left(\hat{\theta}\right)} = \sqrt{n}(1 - \theta\bar{y})$;

(3) Lugannani and Rice approximation in (2.6.3);

(4) Barndorff-Nielsen approximation in (2.6.4).

Finally, to compare the accuracy of these asymptotic methods, we randomly generated 3 data sets from $Exp(3)$ with sample sizes 1, 3 and 10 respectively. The codes in R of this random number generating process are listed below and the resulting simulated data are recorded in Table (2.3).

```
> rexp(1,3)
> rexp(3,3)
> rexp(10,3)
```

Here is the result. Figures (2.3), (2.4) and (2.5) display the exact and the approximated p-values for a grid of θ values from 0 to 20 for the three data sets. The horizontal lines indicate the two nominal levels, 0.025 and 0.975, for the 95% central confidence interval; and the resulting intervals for θ are given in Table 2.4. We can see that both plots and Table 2.4 show the outstanding performance of the third-order methods over the first-order methods, and the third-order methods have remarkable accuracy even when sample size reaches its admissible lower bound of 1.

□

Table 2.3: Three simulated data sets from $Exp(3)$

Data Set	Sample Size n	Observations $(y_1, \ldots, y_n)$
1	1	0.2222027
2	3	0.03838707, 0.27538165, 0.29540766
3	10	0.12836713, 0.02309837, 0.26046243, 0.20793721, 0.30842606, 1.18434839, 0.87649635, 0.03954462, 0.18146885, 0.28150354

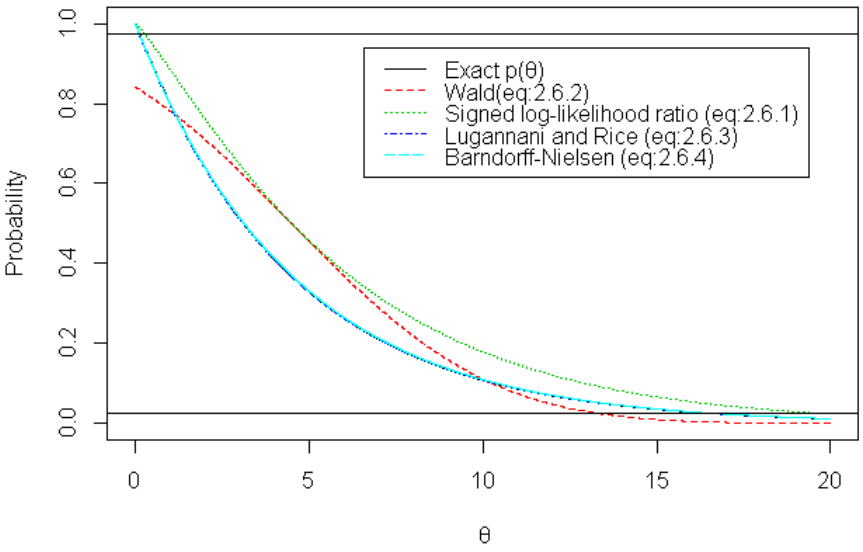


Figure 2.3: $p(\theta)$ for Data Set 1 $(n = 1)$

Table 2.4: 95% central confidence intervals for θ

Method	n=1	n=3	n=10
Exact	(0.1139,16.6014)	(1.0156,11.8598)	(1.3734,4.8930)
Signed log-likelihood ratio	(0.2568,19.8153)	(1.2247,12.7702)	(1.4351,5.0232)
Wald statistic	(-4.3202,13.3210)	(-0.6480,10.4974)	(1.0889,4.6390)
Lugannani and Rice	(0.1095,16.6342)	(1.0142,11.8635)	(1.3734,4.8932)
Barndorff-Nielsen	(0.1114,16.6838)	(1.0156,11.8685)	(1.3735,4.8934)

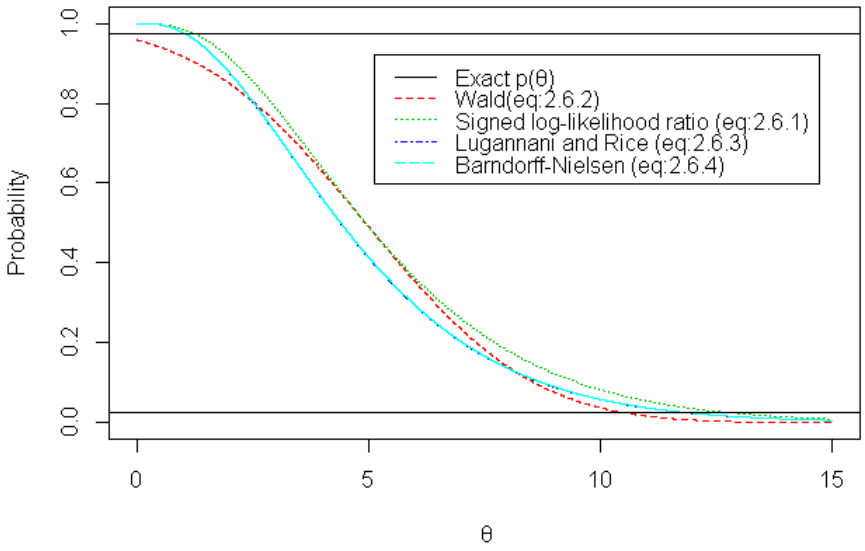


Figure 2.4: $p(\theta)$ **for Data Set 2** $(n = 3)$

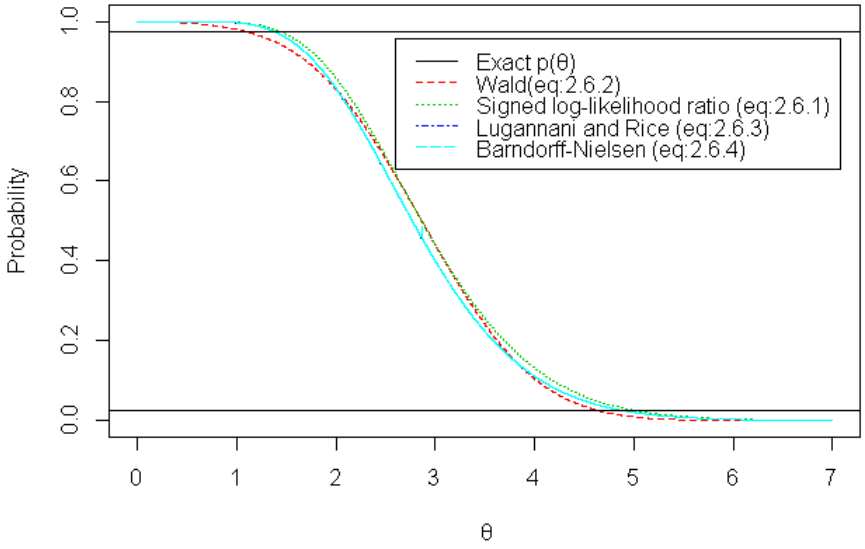


Figure 2.5: $p(\theta)$ for Data Set 3 $(n = 10)$

2.6.1.2 General Model with Single Parameter

For most general application R is still the signed likelihood ratio statistic as (2.6.1); whereas the quantity Q has various forms depending on the model and the procedures being used.

In a sample of size n from a one parameter location family, $f(y;\theta) = f(y-\theta)$, the one-dimensional variable can be taken as $\hat{\theta}$ (or any other location estimate of θ), and (2.6.3) and (2.6.4) give approximations to the conditional distribution of $\hat{\theta}$, given the location-model ancillary statistic $\mathbf{A} = (A_1, \ldots, A_n)$, $A_i = Y_i - \hat{\theta}$. In this case $\ell\left(\theta;\hat{\theta},\mathbf{a}\right) = \sum \log f\left(a_i + \hat{\theta} - \theta\right)$, $\frac{\partial \ell\left(\theta;\hat{\theta},\mathbf{a}\right)}{\partial \hat{\theta}} = -\frac{\partial \ell(\theta)}{\partial \theta}$, and Q simplifies to Rao statistic. (Reid 1996)

$$Q = Q(\theta) = \frac{s(\theta)}{\sqrt{j\left(\hat{\theta}\right)}} = S\,.$$

Lemma. *When the underlying density for random variable Y is a general density $f(y;\theta)$, by using a tangent exponential model approximation, Fraser (1990) derived*

$$Q \quad = (\hat{\varphi} - \varphi)\,\hat{j}^{\frac{1}{2}}_{\hat{\varphi}\varphi} = \quad \left\{\ell_{;y}\left(\hat{\theta}\right) - \ell_{;y}(\theta)\right\}\left\{\frac{\partial \ell_{;y}(\theta)}{\partial \theta}\bigg|_{\theta=\hat{\theta}}\right\}^{-1} j\left(\hat{\theta}\right)^{\frac{1}{2}}\,.$$

In addition, an alternative expression for Q was obtained by Barndorff-Nielsen(1990) by integrating the p^* approximation directly.

$$Q = \left\{\ell_{;\hat{\theta}}\left(\hat{\theta}\right) - \ell_{;\hat{\theta}}(\theta)\right\} j\left(\hat{\theta}\right)^{-\frac{1}{2}}\,,$$

where $\ell_{;\hat{\theta}}(\theta) = \frac{\partial \ell(\theta;y)}{\partial \hat{\theta}} = \frac{\partial \ell\left(\theta;\hat{\theta}\right)}{\partial \hat{\theta}}$ and $\hat{\theta}$ is typically a one-to-one function of y.

Proof. In this proof, we first develop a procedure for an exponential model not in standard form that uses only an observed likelihood function and its first sample space derivative at a data point $\mathbf{y}^0$ and produces directly the cumulant generating function, the canonical parameter and the local canonical variable; in effect, the procedure gives a characterization of an exponential model in terms of likelihood properties local on the sample space. This result is used later to modify the Lugannani and Rice formula and Barndorff-Nielsen formula so that it is independent of the parameterization of the model. Then for a general continuous statistical model we derive an approximating or tangent exponential model at a point y and then use the modified Lugannani and Rice formula and Barndorff-Nielsen formula to obtain a general model version of tail probability approximation. The main reference for this proof is Fraser (1990).

1. Consider a continuous exponential family model, with some one-one equivalent canonical variable $\mathbf{t}(\mathbf{y})$, a function of minimal sufficient statistic

$\mathbf{y}$, and one-one equivalent canonical parameter $\boldsymbol{\varphi}(\boldsymbol{\theta})$ and the scalar parameter of interest being $\psi(\boldsymbol{\theta}) = \psi$.

$$f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta}) = \exp\left(\boldsymbol{\varphi}'(\boldsymbol{\theta})\mathbf{t}(\mathbf{y}) - A\left(\boldsymbol{\varphi}(\boldsymbol{\theta})\right)\right) u(\mathbf{y}) \ . \tag{2.6.5}$$

The canonical parameter $\boldsymbol{\varphi}(\boldsymbol{\theta})$, the canonical variable $\mathbf{t}(\mathbf{y})$, the nominal cumulant generating function $A\left(\boldsymbol{\varphi}(\boldsymbol{\theta})\right)$ and the underlying $u(\mathbf{t}(\mathbf{y}))$ are not uniquely determined. To eliminate the indeterminacy, we standardize the model with respect to a sample point $\mathbf{y}^0$ having the maximum likelihood estimate $\boldsymbol{\theta}^0 = \hat{\boldsymbol{\theta}}\left(\mathbf{y}^0\right)$ and we require: $\mathbf{t}\left(\mathbf{y}^0\right) = \mathbf{0}$, $\boldsymbol{\varphi}\left(\boldsymbol{\theta}^0\right) = \mathbf{0}$, $A(\mathbf{0}) = 0$ and $\frac{\partial \boldsymbol{\varphi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} = \mathbf{I}$. Thus, the first derivative behavior of $\boldsymbol{\varphi}$ and $\boldsymbol{\theta}$ coincide at $\boldsymbol{\theta}^0$.

We will have the following conclusions

$$\boldsymbol{\varphi}'(\boldsymbol{\theta}) = \ell_{;\mathbf{y}'}\left(\boldsymbol{\theta};\mathbf{y}^0\right)\mathbf{s}_{;\mathbf{y}'}^{-1}\left(\boldsymbol{\theta}^0;\mathbf{y}^0\right) = \ell_{;\mathbf{y}'}\left(\boldsymbol{\theta};\mathbf{y}^0\right)\ell_{\boldsymbol{\theta};\mathbf{y}'}^{-1}\left(\boldsymbol{\theta}^0;\mathbf{y}^0\right) \ , \tag{2.6.6}$$

$$A(\boldsymbol{\varphi}) = -\ell\left(\boldsymbol{\varphi}^{-1};\mathbf{y}^0\right) \ , \tag{2.6.7}$$

$$f_{\mathbf{s}}(\mathbf{s};\boldsymbol{\theta}) = f_{\mathbf{s}}(\mathbf{s};\boldsymbol{\theta}^0)e^{\boldsymbol{\varphi}'(\boldsymbol{\theta})\mathbf{s} - A(\boldsymbol{\varphi}(\boldsymbol{\theta}))} \ . \tag{2.6.8}$$

Here are some derivations to get these results. Since $\log f\left(\mathbf{y};\boldsymbol{\theta}^0\right) = \log f\left(\mathbf{y};\hat{\boldsymbol{\theta}}\left(\mathbf{y}^0\right)\right)$ is a function of $\mathbf{y}$ with $\boldsymbol{\theta}^0$ now taken as fixed, we can let

$$\ell(\boldsymbol{\theta};\mathbf{y}) = \log f(\mathbf{y};\boldsymbol{\theta}) - \log f\left(\mathbf{y};\boldsymbol{\theta}^0\right) \tag{2.6.9}$$

be the likelihood function normed with respect to the value $\boldsymbol{\theta}^0$.

Also let $\mathbf{s} = \mathbf{s}\left(\boldsymbol{\theta}^0;\mathbf{y}\right)$ be the $\boldsymbol{\theta}^0$ score taken as a function of $\mathbf{y}$. From (2.6.5) and $\mathbf{s}\left(\boldsymbol{\theta}^0;\mathbf{y}^0\right) = \mathbf{0}$, we have:

$$\begin{aligned}
\mathbf{s}\left(\boldsymbol{\theta}^0;\mathbf{y}^0\right) &= \frac{\partial \ell\left(\boldsymbol{\theta};\mathbf{y}^0\right)}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} = \frac{\partial\left(\boldsymbol{\varphi}'(\boldsymbol{\theta})\mathbf{t}(\mathbf{y}^0) - A\left(\boldsymbol{\varphi}(\boldsymbol{\theta})\right)\right)}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} \\
&= \frac{\partial \boldsymbol{\varphi}'(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} \cdot \mathbf{t}\left(\mathbf{y}^0\right) - \frac{\partial \boldsymbol{\varphi}'(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} \cdot \frac{\partial A}{\partial \boldsymbol{\varphi}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} \\
&= \frac{\partial A}{\partial \boldsymbol{\varphi}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} = \mathbf{0} \ .
\end{aligned}$$

Thus, we can derive $\mathbf{s} = \mathbf{s}\left(\boldsymbol{\theta}^0;\mathbf{y}\right)$ as:

$$\begin{aligned}
\mathbf{s} &= \mathbf{s}\left(\boldsymbol{\theta}^0;\mathbf{y}\right) = \frac{\partial \ell(\boldsymbol{\theta};\mathbf{y})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} = \frac{\partial\left(\boldsymbol{\varphi}'(\boldsymbol{\theta})\mathbf{t}(\mathbf{y}) - A\left(\boldsymbol{\varphi}(\boldsymbol{\theta})\right)\right)}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} \\
&= \frac{\partial \boldsymbol{\varphi}'(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} \cdot \mathbf{t}(\mathbf{y}) - \frac{\partial \boldsymbol{\varphi}'(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} \cdot \frac{\partial A}{\partial \boldsymbol{\varphi}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0} = \mathbf{t}(\mathbf{y}) \ .
\end{aligned}$$

Rewrite (2.6.5) as $\ell(\boldsymbol{\theta};\mathbf{y}) = \boldsymbol{\varphi}'(\boldsymbol{\theta})s\left(\boldsymbol{\theta}^0;\mathbf{y}\right) - A\left(\boldsymbol{\varphi}(\boldsymbol{\theta})\right)$, express $\boldsymbol{\varphi}$ with the

sample space derivatives and we obtain (2.6.6) and (2.6.7).

$$\begin{aligned}\boldsymbol{\varphi}'(\boldsymbol{\theta}) &= \frac{\partial \ell(\boldsymbol{\theta};\mathbf{y})}{\partial \mathbf{s}'} = \frac{\partial \ell(\boldsymbol{\theta};\mathbf{y})}{\partial (\mathbf{y}^0)'}\frac{\partial \mathbf{y}^0}{\partial \mathbf{s}'} = \frac{\partial \ell(\boldsymbol{\theta};\mathbf{y})}{\partial (\mathbf{y}^0)'}\cdot\left\{\frac{\partial \mathbf{s}}{\partial (\mathbf{y}^0)'}\right\}^{-1} \\ &= \ell_{;\mathbf{y}'}(\boldsymbol{\theta};\mathbf{y}^0)\,\mathbf{s}^{-1}_{;\mathbf{y}'}(\boldsymbol{\theta}^0;\mathbf{y}^0) = \ell_{;\mathbf{y}'}(\boldsymbol{\theta};\mathbf{y}^0)\,\ell^{-1}_{\boldsymbol{\theta};\mathbf{y}'}(\boldsymbol{\theta}^0;\mathbf{y}^0)\ ,\end{aligned}$$

$$A(\boldsymbol{\varphi}) = \boldsymbol{\varphi}'(\boldsymbol{\theta})s(\boldsymbol{\theta}^0;\mathbf{y}^0) - \ell(\boldsymbol{\theta};\mathbf{y}^0) = -\ell(\boldsymbol{\theta};\mathbf{y}^0) = -\ell(\boldsymbol{\varphi}^{-1};\mathbf{y}^0)\ .$$

Since $f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta}^0) = e^{\boldsymbol{\varphi}'(\boldsymbol{\theta}^0)\mathbf{t}(\mathbf{y}) - A\left(\boldsymbol{\varphi}(\boldsymbol{\theta}^0)\right)}u(\mathbf{y}) = u(\mathbf{y})$, thus the exponential model $f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})$ can then be written as

$$\begin{aligned} & f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta}) \\ = \ & f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta}^0)e^{\boldsymbol{\varphi}'(\boldsymbol{\theta})\mathbf{s}\left(\boldsymbol{\theta}^0;\mathbf{y}\right) - A(\boldsymbol{\varphi}(\boldsymbol{\theta}))} = f_{\mathbf{s}}(\mathbf{s};\boldsymbol{\theta}^0)e^{\boldsymbol{\varphi}'(\boldsymbol{\theta})\mathbf{s} - A(\boldsymbol{\varphi}(\boldsymbol{\theta}))}\left|\frac{\partial \mathbf{s}}{\partial \mathbf{y}'}\right| \quad (2.6.10) \\ = \ & f_{\mathbf{s}}(\mathbf{s};\boldsymbol{\theta}^0)e^{\boldsymbol{\varphi}'(\boldsymbol{\theta})\mathbf{s} - A(\boldsymbol{\varphi}(\boldsymbol{\theta}))}\left|\mathbf{s}_{;\mathbf{y}'}(\boldsymbol{\theta}^0;\mathbf{y})\right| = f_{\mathbf{s}}(\mathbf{s};\boldsymbol{\theta})\left|\mathbf{s}_{;\mathbf{y}'}(\boldsymbol{\theta}^0;\mathbf{y})\right| \quad (2.6.11)\end{aligned}$$

From the results above, we can see that the observed likelihood $\ell(\boldsymbol{\theta};\mathbf{y}^0)$ and its first sample space derivative $\ell_{;\mathbf{y}'}(\boldsymbol{\theta};\mathbf{y}^0)$ entirely determine the natural parameter $\boldsymbol{\varphi}$ by (2.6.6) and the nominal cumulant generating function $A(\boldsymbol{\varphi})$ by (2.6.7) and thus fully determine $f_{\mathbf{s}}(\mathbf{s};\boldsymbol{\theta})$, the exponential model for the score variable $\mathbf{s}$, by (2.6.8) and also the density $f_{\mathbf{Y}}(\mathbf{y}^0;\boldsymbol{\theta})$ at the observed data point $\mathbf{y}^0$. The density $f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})$ elsewhere on the sample space for $\mathbf{y}$ (2.6.11) requires $\left|\mathbf{s}_{;\mathbf{y}'}(\boldsymbol{\theta}^0;\mathbf{y})\right|$ and would not be available from the $\mathbf{y}^0$ likelihood information. In addition, for computation, $\boldsymbol{\varphi}$ and $A(\boldsymbol{\varphi})$ are directly available but the null density $f_{\mathbf{s}}(\mathbf{s};\boldsymbol{\theta}^0)$ would generally require a Fourier inversion from $A(\boldsymbol{\varphi})$; for statistical purposes, however, an accurate approximation for $f_{\mathbf{s}}(\mathbf{s};\boldsymbol{\theta}^0)$ is directly available by the saddlepoint method.

2. We generalize the results above to the more general case where $\mathbf{y}$ is not minimal sufficient or the exponential structure is somehow disguised. For this, consider a continuous model $f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})$ where the parameter $\boldsymbol{\theta}$ and the minimal sufficient statistic have dimension p but the variable $\mathbf{y}$ has dimension n. Thus (2.6.5) is more similar to (2.2.7). In this setting, the minimal sufficient statistic locally at the point $\mathbf{y}$ (Fraser 1966) takes values in a p dimensional vector space of $\boldsymbol{\theta}$ functions, $L_{MS}\{\ell_1(\boldsymbol{\theta};\mathbf{y}),\ldots,\ell_n(\boldsymbol{\theta};\mathbf{y})\}$, where $\ell(\boldsymbol{\theta};\mathbf{y})$ is given in (2.6.9) and $\ell_i(\boldsymbol{\theta};\mathbf{y}) = \frac{\partial \ell(\boldsymbol{\theta};\mathbf{y})}{\partial y_i}$. And in the result above the derivative $\ell_{;\mathbf{y}'}(\boldsymbol{\theta};\mathbf{y})$ is replaced by $\ell_{;\mathbf{V}}(\boldsymbol{\theta};\mathbf{y})$, which is the directional derivatives span the space L_{MS}. This would happen typically, unless a chosen $\mathbf{V}_i$ fell in the $n-p$ dimensional null space of the linear forms $d\ell$ (Frsaer 1990). In addition, $\mathbf{s}_{;\mathbf{y}'}(\boldsymbol{\theta};\mathbf{y})$ is replaced by $\mathbf{s}_{;\mathbf{V}}(\boldsymbol{\theta};\mathbf{y})$.

3. The Lugannani and Rice formula and Barndorff-Nielson formula give left-tail probability approximation $p(\theta) = F\left(\hat{\theta};\theta\right) = F(\hat{\varphi};\varphi)$ for a one-parameter exponential model. The accompanying definition (2.6.1) for R

uses likelihood drop from $\hat{\varphi}$ to φ and is invariant under reparameterization. The definition (2.6.2) for Q, however, uses the canonical parameter φ. A parameterization invariant version of Q is recorded as

$$\begin{aligned} Q &= (\hat{\varphi}-\varphi)\sqrt{j(\hat{\varphi})} \\ &= \left\{\ell_{;V}\left(\theta^0;\mathbf{y}^0\right)s_{;V}^{-1}\left(\theta^0;\mathbf{y}^0\right)-\ell_{;V}\left(\theta;\mathbf{y}^0\right)s_{;V}^{-1}\left(\theta^0;\mathbf{y}^0\right)\right\}\sqrt{j(\hat{\varphi})} \\ &= \left\{\ell_{;V}\left(\hat{\theta};\mathbf{y}\right)-\ell_{;V}(\theta;\mathbf{y})\right\}\left\{\ell_{\theta;V}\left(\hat{\theta};\mathbf{y}\right)\right\}^{-1}j\left(\hat{\theta}\right)^{\frac{1}{2}}. \end{aligned}$$

In addition, we can total differentiate the estimating equation $s\left(\hat{\theta};\mathbf{y}\right)=0$, obtaining $\frac{\partial s}{\partial\hat{\theta}}d\hat{\theta}+\frac{\partial s}{\partial V}dV=0$ and $\frac{d\hat{\theta}}{dV}=-\frac{\frac{\partial s}{\partial V}}{\frac{\partial s}{\partial\hat{\theta}}}=\frac{\ell_{\theta;V}\left(\hat{\theta};\mathbf{y}\right)}{-\ell_{\theta\theta}\left(\hat{\theta};\mathbf{y}\right)}=j\left(\hat{\theta}\right)^{-1}\ell_{\theta;V}\left(\hat{\theta};\mathbf{y}\right)$. Therefore, Q can also be expressed as

$$\begin{aligned} Q &= \left\{\ell_{;V}\left(\hat{\theta};\mathbf{y}\right)-\ell_{;V}(\theta;\mathbf{y})\right\}\left\{\ell_{\theta;V}\left(\hat{\theta};\mathbf{y}\right)\right\}^{-1}j\left(\hat{\theta}\right)^{\frac{1}{2}} \\ &= \left\{\ell_{;\hat{\theta}}\left(\hat{\theta};\mathbf{y}\right)-\ell_{;\hat{\theta}}(\theta;\mathbf{y})\right\}\frac{d\hat{\theta}}{dV}\left\{\ell_{\theta;V}\left(\hat{\theta};\mathbf{y}\right)\right\}^{-1}j\left(\hat{\theta}\right)^{\frac{1}{2}} \\ &= \left\{\ell_{;\hat{\theta}}\left(\hat{\theta};\mathbf{y}\right)-\ell_{;\hat{\theta}}(\theta;\mathbf{y})\right\}j\left(\hat{\theta}\right)^{-\frac{1}{2}}. \end{aligned}$$

In these expressions we have used $\mathbf{y}$ for the data point and $\hat{\theta}$ for the corresponding maximum likelihood estimate and θ remains as the parameter value for which the tail probability is being calculated. We also note that normalization of the likelihood function is unnecessary: that $\log f(\mathbf{y};\theta)$ can be used in place of $\ell(\theta;\mathbf{y})$ (Frsaer 1990).

4. Without the exponential assumption, we find it reasonable for inference to use an approximating exponential model in preference to an approximating normal model. The approximating exponential model is called a tangent exponential model. For the general continuous model $f(\mathbf{y};\boldsymbol{\theta})$ we take the tangent exponential model at the point $\mathbf{y}^0$ to be the exponential model (2.6.5) that coincides with the given model at $\mathbf{y}^0$. □

The comparison between (2.6.3) and (2.6.4) on a variety of models has been conducted extensively in literature. Barndorff-Nielsen (1990a) for exponential, inverse Gaussian and von Mises distributions; Fraser (1990) for the location log-gamma, the gamma, the logistic, and Cauchy distribution; Barndorff-Nielsen (1991) for Cauchy distribution; Barndorff-Nielsen and Chamberlin (1991) for the location log-gamma, the inverse Gaussian, and a specially constructed (2,1) curved exponential family; Fraser and Reid (1993) for a tilted and shifted logistic model which has one parameter but is neither an exponential nor a location model; DiCiccio and Martin (1993) for an exponential model with censoring. In addition, DiCiccio, Field and Fraser (1990) compare the Lugannani and Rice approximation with the

mean and variance corrected version of Fisher's hyperbola model, which is a (2,1) curved exponential family. Kolassa (2006), especially in the exercises, considers several numerical illustrations, and provides Mathematica code for computing many of them.

2.6.2 Exponential family Model and Transformation Model with Multiple Parameters

This section will study the cases of the exponential family model and the transformation model with the presence of nuisance parameter λ, and at next section we will generalize to any statistical model with multiple parameters. The proposed methodology only depends on the likelihood function and its first sample space derivative at the data points.

2.6.2.1 Canonical Exponential Model with Multiple Parameters

Consider a canonical exponential family model in (2.2.8) or (2.2.9) with the canonical parameter $\theta = \left(\psi, \boldsymbol{\lambda}'\right)'$. For any given random sample from this model, $R = R(\psi)$ remains the signed log-likelihood ratio statistic in the form of (2.3.15). Taken into the consideration of eliminating the nuisance parameter $\boldsymbol{\lambda}$, $Q = Q(\psi) = q(\psi)$ is the standardized maximum likelihood departure statistic in the canonical parameter space in the form of (2.3.13).

Hence, we can apply either Lugannani and Rice method or Barndorff-Nielsen method to obtain $p(\psi)$ with R and Q defined above.

$$p(\psi) = \Phi(R) + \phi(R)\left(\frac{1}{R} - \frac{1}{Q}\right) + O\left(n^{-\frac{3}{2}}\right), \tag{2.6.12}$$

$$p(\psi) = \Phi(R^*) + O\left(n^{-\frac{3}{2}}\right) = \Phi\left(R - \frac{1}{R}\log\frac{R}{Q}\right) + O\left(n^{-\frac{3}{2}}\right). \tag{2.6.13}$$

2.6.2.2 General Exponential Model with Multiple Parameters

Again, since the signed log-likelihood ratio statistic, $R = R(\psi)$, is invariant to reparameterization, it remains unchanged and is defined in (2.3.15). However, the quantity $Q = Q(\psi)$ has to be re-expressed in the canonical parameter, $\boldsymbol{\varphi}(\boldsymbol{\theta})$, scale. Here are the steps:

1. For general exponential model, we can obtain canonical parameter by (2.6.6).

2. Denote $\boldsymbol{\varphi}^{\psi}(\boldsymbol{\theta})$ to be the row of $\boldsymbol{\varphi}_{\boldsymbol{\theta}'}^{-1}(\boldsymbol{\theta})$ that corresponds to ψ, and $\left\|\boldsymbol{\varphi}^{\psi}(\boldsymbol{\theta})\right\|^2$ is the squared length of the vector.

3. Let $\chi(\boldsymbol{\theta}) = \frac{\boldsymbol{\varphi}^{\psi}(\hat{\boldsymbol{\theta}}_{\psi})}{\left\|\boldsymbol{\varphi}^{\psi}(\hat{\boldsymbol{\theta}}_{\psi})\right\|}\boldsymbol{\varphi}(\boldsymbol{\theta})$. And $\chi(\boldsymbol{\theta})$ can be viewed operationally as the scalar parameter of interest $\psi(\boldsymbol{\theta})$ in $\boldsymbol{\varphi}(\boldsymbol{\theta})$ scale, and it is actually a rotated coordinate of $\boldsymbol{\varphi}(\boldsymbol{\theta})$ that agrees with $\psi(\boldsymbol{\theta})$ at $\hat{\boldsymbol{\theta}}_{\psi}$. Basically, the calibrated version $\chi(\boldsymbol{\theta})$ of $\psi(\boldsymbol{\theta})$ is a vector from the space spanned by the

columns of $\boldsymbol{\varphi}(\boldsymbol{\theta})$ and its direction depends on the constrained MLE $\hat{\boldsymbol{\theta}}_{\psi}$ for given $\boldsymbol{\varphi}(\boldsymbol{\theta})$;

4. From above, we can see $\left|\chi\left(\hat{\boldsymbol{\theta}}\right)-\chi\left(\hat{\boldsymbol{\theta}}_{\psi}\right)\right|$ is a measure of departure of $\hat{\psi}$ from ψ in $\boldsymbol{\varphi}(\boldsymbol{\theta})$ scale. In addition, Fraser, Reid and Wu (1999) obtained an estimated variance for $\left\{\chi\left(\hat{\boldsymbol{\theta}}\right)-\chi\left(\hat{\boldsymbol{\theta}}_{\psi}\right)\right\}$ in $\boldsymbol{\varphi}(\boldsymbol{\theta})$ scale: $\widehat{\text{var}}\left(\chi\left(\hat{\boldsymbol{\theta}}\right)-\chi\left(\hat{\boldsymbol{\theta}}_{\psi}\right)\right)$ $=\frac{\left|\mathbf{j}_{(\boldsymbol{\lambda}\boldsymbol{\lambda}')}\left(\hat{\boldsymbol{\theta}}_{\psi}\right)\right|}{\left|\mathbf{j}_{(\boldsymbol{\theta}\boldsymbol{\theta}')}\left(\hat{\boldsymbol{\theta}}\right)\right|}$. By chain rule in differentiation, we have the full information defined on the canonical parameter apace $\left|\mathbf{j}_{(\boldsymbol{\theta}\boldsymbol{\theta}')}\left(\hat{\boldsymbol{\theta}}\right)\right|=\left|\mathbf{j}_{\boldsymbol{\theta}\boldsymbol{\theta}'}\left(\hat{\boldsymbol{\theta}}\right)\right|\left|\boldsymbol{\varphi}_{\boldsymbol{\theta}'}\left(\hat{\boldsymbol{\theta}}\right)\right|^{-2}$ and nuisance information on the canonical parameter space [10] $\left|\mathbf{j}_{(\boldsymbol{\lambda}\boldsymbol{\lambda}')}\left(\hat{\boldsymbol{\theta}}_{\psi}\right)\right|=$ $\left|\mathbf{j}_{\boldsymbol{\lambda}\boldsymbol{\lambda}'}\left(\hat{\boldsymbol{\theta}}_{\psi}\right)\right|\left|\boldsymbol{\varphi}'_{\boldsymbol{\lambda}'}\left(\hat{\boldsymbol{\theta}}_{\psi}\right)\boldsymbol{\varphi}_{\boldsymbol{\lambda}'}\left(\hat{\boldsymbol{\theta}}_{\psi}\right)\right|^{-1}$

5. Finally, the standardized maximum likelihood departure of ψ in $\boldsymbol{\varphi}(\boldsymbol{\theta})$ scale is Q.

$$\begin{aligned} Q &= Q(\psi)=\frac{\text{sgn}\left(\hat{\psi}-\psi\right)\left|\chi\left(\hat{\boldsymbol{\theta}}\right)-\chi\left(\hat{\boldsymbol{\theta}}_{\psi}\right)\right|}{\sqrt{\widehat{\text{var}}\left(\chi\left(\hat{\boldsymbol{\theta}}\right)-\chi\left(\hat{\boldsymbol{\theta}}_{\psi}\right)\right)}} \\ &= \text{sgn}\left(\hat{\psi}-\psi\right)\left|\chi\left(\hat{\boldsymbol{\theta}}\right)-\chi\left(\hat{\boldsymbol{\theta}}_{\psi}\right)\right|\left\{\frac{\left|\mathbf{j}_{(\boldsymbol{\lambda}\boldsymbol{\lambda}')}\left(\hat{\boldsymbol{\theta}}_{\psi}\right)\right|}{\left|\mathbf{j}_{(\boldsymbol{\theta}\boldsymbol{\theta}')}\left(\hat{\boldsymbol{\theta}}\right)\right|}\right\}^{-\frac{1}{2}} . \quad (2.6.14) \end{aligned}$$

2.6.3 General Model

Fraser(1988)(1990), Fraser and Reid(1995), Fraser, Reid and Wu(1998) developed a general tail probability formula or called the Tangent Exponential Model. In their method, the statistic R is same as signed log-likelihood ratio statistic defined at (2.3.15). In order to get the statistic Q, we will have to process the following in advance:

First, we obtain the canonical parameter $\boldsymbol{\varphi}(\boldsymbol{\theta})$ by taking the sample space gradient at the observed data point $\mathbf{y}^0$ calculated in the directions given by a set of vectors $\mathbf{V}$:

$$\boldsymbol{\varphi}'(\boldsymbol{\theta})=\frac{\partial}{\partial \mathbf{y}'}\ell(\boldsymbol{\theta})\Big|_{\mathbf{y}^0}\cdot\mathbf{V} , \quad (2.6.15)$$

$$\mathbf{V}=\frac{\partial \mathbf{y}}{\partial \boldsymbol{\theta}'}\Big|_{\left(\mathbf{y}^0,\hat{\boldsymbol{\theta}}\right)}=-\left(\frac{\partial \mathbf{z}}{\partial \mathbf{y}'}\right)^{-1}\frac{\partial \mathbf{z}}{\partial \boldsymbol{\theta}'}\Big|_{\left(\mathbf{r}^0,\hat{\boldsymbol{\theta}^0}\right)} . \quad (2.6.16)$$

The set of vectors in $\mathbf{V}$ are referred to as ancillary directions and capture how the data is influenced by parameter change near the maximum like-

[10] Note that $\boldsymbol{\varphi}_{\boldsymbol{\lambda}'}\left(\hat{\boldsymbol{\theta}}_{\psi}\right)$ is a $p\times(p-1)$ dimensional matrix.

lihood value. The differentiation in (2.6.16) is taken for fixed values of a full-dimensional pivotal quantity and is defined from the total differentiation of this pivotal. A pivotal statistic $\mathbf{z}(\boldsymbol{\theta}, \mathbf{y})$ is a function of the variable $\mathbf{y}$ and the parameter $\boldsymbol{\theta}$ that has a fixed distribution (independent of $\boldsymbol{\theta}$) and is required component of the methodology.

Implicit in (2.6.15) is the necessary conditioning that reduces the dimension of the problem from n to p. This is done through the vectors in $\mathbf{V}$ which are based on the pivotal quantity $\mathbf{z}(\boldsymbol{\theta}, \mathbf{y})$ which in (2.6.15) serve to condition on an approximate ancillary statistic. This is a very technical point and the reader is referred to Fraser and Reid (1995) for full technical details.

Second, replace the parameter of interest $\psi(\boldsymbol{\theta})$ by a linear function of the $\boldsymbol{\varphi}(\boldsymbol{\theta})$ coordinates. This newly calibrated parameter $\chi(\boldsymbol{\theta})$ is given by the following:

$$\chi(\boldsymbol{\theta}) = \psi_{\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_\psi)\boldsymbol{\varphi}^{-1}_{\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_\psi)\boldsymbol{\varphi}(\boldsymbol{\theta})\ . \tag{2.6.17}$$

In order to apply Lugannani and Rice approximation in (2.6.12) or the Barndorff-Nielsen approximation in (2.6.13) to make inference on our parameter of interest $\psi(\boldsymbol{\theta})$, we need to calculate both R and Q in $\boldsymbol{\varphi}(\boldsymbol{\theta})$ scale. Since R calculated in the original $\boldsymbol{\theta}$ scale is equivalent to that calculated in $\boldsymbol{\varphi}(\boldsymbol{\theta})$ scale, the key step is to derive Q, the standardized maximum likelihood departure in the canonical parameter $\boldsymbol{\varphi}(\theta)$scale:

$$Q \equiv Q(\psi) = sgn(\hat{\psi} - \psi)\frac{\left|\chi(\hat{\boldsymbol{\theta}}) - \chi(\hat{\boldsymbol{\theta}}_\psi)\right|}{\sqrt{\hat{var}(\chi(\hat{\boldsymbol{\theta}}) - \chi(\hat{\boldsymbol{\theta}}_\psi))}}\ , \tag{2.6.18}$$

where $\left|\chi(\hat{\boldsymbol{\theta}}) - \chi(\hat{\boldsymbol{\theta}}_\psi)\right|$measures the departure of $\left|\hat{\psi} - \psi\right|$in the canonical parameter $\boldsymbol{\varphi}(\boldsymbol{\theta})$ scale. In addition the estimated variance of $(\chi(\hat{\boldsymbol{\theta}}) - \chi(\hat{\boldsymbol{\theta}}_\psi))$ is given by Fraser and Reid(1995) as:

$$\hat{var}(\chi(\hat{\boldsymbol{\theta}})-\chi(\hat{\boldsymbol{\theta}}_\psi)) = \frac{\psi_{\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_\psi)\tilde{\mathbf{j}}^{-1}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_\psi)\psi'_{\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_\psi)\left|\tilde{\mathbf{j}}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_\psi)\right|\left|\boldsymbol{\varphi}_{\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_\psi)\right|^{-2}}{\left|\mathbf{j}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}})\right|\left|\boldsymbol{\varphi}_{\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}})\right|^{-2}}\ . \tag{2.6.19}$$

In Chapter 2, we review the development of asymptotic likelihood-based methods, and in the next chapter, we will apply the third-order likelihood methodology into the reference on Sharpe ratio. The performance of the proposed method for normally distributed underlying returns is also examined through both real-life data set and simulated small data sets. Comparison between proposed method and existing method in literature will be conducted to prove the advantages of our proposed method.

Chapter 3

Asymptotic Likelihood Inference for Sharpe Ratio under IID Normal Log Return

3.1 Inference for Standard Sharpe Ratio under IID Normal Log Return

We emphasize that the third-order methodology discussed in the previous chapter is applicable under any parametric distributional assumptions, and thus in this chapter we will demonstrate the use of the method under the assumption of IID normal log returns, or equivalently, of IID lognormal gross returns.

3.1.1 Likelihood Methodology for One Sample Sharpe Ratio

Consider a fund with log-return at time t denoted by r_t, $t = 1, 2, \ldots, T$. Under IID normal assumption, we have $r_t \sim N(\mu, \sigma^2)$ and $\boldsymbol{\theta}' = (\mu, \sigma^2)$, meanwhile the canonical parameter of normal distribution and its first order derivative can be written as:

$$\boldsymbol{\varphi}'(\boldsymbol{\theta}) = \left(\frac{\mu}{\sigma^2}, \frac{1}{\sigma^2} \right) , \tag{3.1.1}$$

$$\varphi_{\boldsymbol{\theta}'}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{\sigma^2} & -\frac{\mu}{\sigma^4} \\ 0 & -\frac{1}{\sigma^4} \end{pmatrix} . \tag{3.1.2}$$

In addition, the determinant of this first order derivatives $|\boldsymbol{\varphi}_{\boldsymbol{\theta}'}(\boldsymbol{\theta})| = -\frac{1}{\sigma^6}$ and inverse matrix:$\boldsymbol{\varphi}^{-1}_{\boldsymbol{\theta}'}(\boldsymbol{\theta}) = \begin{pmatrix} \sigma^2 & -\mu\sigma^2 \\ 0 & \sigma^4 \end{pmatrix}$ will be used later at calculating Q.

Let r_f be a constant representing mean return for the risk-free asset. The parameter of interest is Sharpe ratio and we can explicitly list out its parametric form and first order derivatives:

$$\psi(\boldsymbol{\theta}) = SR = \frac{\mu - r_f}{\sigma} \,, \tag{3.1.3}$$

$$\psi_{\boldsymbol{\theta}'}(\boldsymbol{\theta}) = \left(\frac{1}{\sigma}, -\frac{\mu - r_f}{2\sigma^3}\right) \,. \tag{3.1.4}$$

Then, we start with unrestricted maximum likelihood estimation, which maximize the log-likelihood function $l(\boldsymbol{\theta})$.

$$l(\boldsymbol{\theta}) = l(\mu, \sigma^2) = a + \log\left(\prod f(r_t; \boldsymbol{\theta})\right) = a - \frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum(r_t - \mu)^2 \,. \tag{3.1.5}$$

Its first order and second order derivatives are calculated as follows:

$$l_{\mu}(\boldsymbol{\theta}) = \frac{\sum(r_t - \mu)}{\sigma^2} \,, \tag{3.1.6}$$

$$l_{\sigma^2}(\boldsymbol{\theta}) = -\frac{T}{2\sigma^2} + \frac{\sum(r_t - \mu)^2}{2\sigma^4} \,, \tag{3.1.7}$$

$$l_{\mu\mu}(\boldsymbol{\theta}) = -\frac{T}{\sigma^2} \,, \tag{3.1.8}$$

$$l_{\mu\sigma^2}(\boldsymbol{\theta}) = l_{\sigma^2\mu}(\boldsymbol{\theta}) = -\frac{\sum(r_t - \mu)}{\sigma^4} \,, \tag{3.1.9}$$

$$l_{\sigma^2\sigma^2}(\boldsymbol{\theta}) = \frac{T}{2\sigma^4} - \frac{\sum(r_t - \mu)^2}{\sigma^6} \,. \tag{3.1.10}$$

To obtain the unrestricted maximum likelihood estimator $\hat{\boldsymbol{\theta}}$, we solve first order conditions, that is to solve both (3.1.6) and (3.1.7) to zero. The result is:

$$\hat{\boldsymbol{\theta}}' = (\hat{\mu}, \hat{\sigma}^2) = (\frac{\sum r_t}{T}, \frac{\sum(r_t - \hat{\mu})^2}{T}) \,. \tag{3.1.11}$$

Knowing MLE, we can achieve other important variable for future use, such as the estimated Sharpe ratio $\hat{\psi} = \frac{\hat{\mu} - r_f}{\hat{\sigma}}$, the estimated unrestricted likelihood function $l(\hat{\boldsymbol{\theta}}) = a - \frac{T}{2}\log\hat{\sigma}^2 - \frac{T}{2}$, the observed information matrix evaluated at $\hat{\boldsymbol{\theta}}$, $\mathbf{j}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} -l_{\mu\mu}(\hat{\boldsymbol{\theta}}) & -l_{\mu\sigma^2}(\hat{\boldsymbol{\theta}}) \\ -l_{\sigma^2\mu}(\hat{\boldsymbol{\theta}}) & -l_{\sigma^2\sigma^2}(\hat{\boldsymbol{\theta}}) \end{pmatrix} = \begin{pmatrix} \frac{T}{\hat{\sigma}^2} & 0 \\ 0 & \frac{T}{2\hat{\sigma}^4} \end{pmatrix}$, and its

determinant $\frac{T^2}{2\hat{\sigma}^6}$.

The constrained maximum likelihood estimator can be solved by the Lagrange multiplier method (see (2.3.7)). The Lagrangian function is given at (3.1.12) and its first order derivatives are listed from (3.1.13) to (3.1.15). The tilted log-likelihood function can be obtained by replacing α by $\hat{\alpha}$ on the Lagrangian function, and its second order derivatives are given by (3.1.16) to (3.1.18).

$$H(\boldsymbol{\theta}, \alpha) = l(\boldsymbol{\theta}) + \alpha(\psi(\boldsymbol{\theta}) - \psi) = l(\boldsymbol{\theta}) + \alpha(\frac{\mu - r_f}{\sigma} - \psi) \ , \tag{3.1.12}$$

$$H_{\mu}(\boldsymbol{\theta}, \alpha) = l_{\mu}(\boldsymbol{\theta}) + \frac{\alpha}{\sigma} \ , \tag{3.1.13}$$

$$H_{\sigma^2}(\boldsymbol{\theta}, \alpha) = l_{\sigma^2}(\boldsymbol{\theta}) - \frac{\alpha(\mu - r_f)}{2\sigma^3} \ , \tag{3.1.14}$$

$$H_{\alpha}(\boldsymbol{\theta}, \alpha) = \frac{\mu - r_f}{\sigma} - \psi \ , \tag{3.1.15}$$

$$\tilde{l}_{\mu\mu}(\boldsymbol{\theta}) = H_{\mu\mu}(\boldsymbol{\theta}, \hat{\alpha}) = l_{\mu\mu}(\boldsymbol{\theta}) \ , \tag{3.1.16}$$

$$\tilde{l}_{\mu\sigma^2}(\boldsymbol{\theta}) = H_{\mu\sigma^2}(\boldsymbol{\theta}, \hat{\alpha}) = l_{\mu\sigma^2}(\boldsymbol{\theta}) - \frac{\hat{\alpha}}{2\sigma^3} \ , \tag{3.1.17}$$

$$\tilde{l}_{\sigma^2\sigma^2}(\boldsymbol{\theta}) = H_{\sigma^2\sigma^2}(\boldsymbol{\theta}, \hat{\alpha}) = l_{\sigma^2\sigma^2}(\boldsymbol{\theta}) + \frac{3}{4}\frac{\hat{\alpha}(\mu - r_f)}{\sigma^5} \ . \tag{3.1.18}$$

Solving first order derivatives, from (3.1.13) to (3.1.15), equal to zero, we will obtain the restricted maximum likelihood estimator. [1]

$$\hat{\boldsymbol{\theta}}'_{\psi} = (\tilde{\mu}, \tilde{\sigma}^2) = \left(r_f + \psi\tilde{\sigma}, \left(\frac{-\psi\bar{w} + \sqrt{(\psi\bar{w})^2 + \frac{4\sum w_t^2}{T}}}{2} \right)^2 \right) \ , \tag{3.1.19}$$

$$\hat{\alpha} = T\left(\psi - \frac{\bar{w}}{\tilde{\sigma}}\right) \ , \tag{3.1.20}$$

$$\bar{w} = \frac{\sum w_t}{T} = \frac{\sum(r_t - r_f)}{T} = \bar{r} - r_f \ . \tag{3.1.21}$$

Using constrained MLE, we are able to obtain some other important results, such as estimated restricted likelihood function: $l(\hat{\boldsymbol{\theta}}_{\psi}) = -\frac{T}{2}\log\tilde{\sigma}^2 - \frac{T}{2} - \frac{\hat{\alpha}\psi}{2}$, the tilted observed information matrix evaluated at $\hat{\boldsymbol{\theta}}_{\psi}$, $\tilde{\mathbf{j}}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_{\psi}) =$

[1] During the derivation, we have four important results: $\sum(r_t - \tilde{\mu}) = -\hat{\alpha}\tilde{\sigma}$, $\sum(r_t - r_f) = T\bar{w}$, $\tilde{\mu} - r_f = \psi\tilde{\sigma}$, $\sum(r_t - \tilde{\mu})^2 = (T + \hat{\alpha}\psi)\tilde{\sigma}^2$.

$-\frac{\partial^2 \tilde{\mathbf{j}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \mid_{\theta=\hat{\theta}_\psi} = \begin{pmatrix} \frac{T}{\tilde{\sigma}^2} & -\frac{\hat{\alpha}}{2\tilde{\sigma}^3} \\ -\frac{\hat{\alpha}}{2\tilde{\sigma}^3} & \frac{\hat{\alpha}\psi+2T}{4\tilde{\sigma}^4} \end{pmatrix}$, and its inverse can be calculated as $\tilde{\mathbf{j}}^{-1}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_\psi) = \left|\tilde{\mathbf{j}}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_\psi)\right|^{-1} \begin{pmatrix} \frac{\hat{\alpha}\psi+2T}{4\tilde{\sigma}^4} & \frac{\hat{\alpha}}{2\tilde{\sigma}^3} \\ \frac{\hat{\alpha}}{2\tilde{\sigma}^3} & \frac{T}{\tilde{\sigma}^2} \end{pmatrix}$.

Having prepared both the constrained MLE and unconstrained MLE, then we can obtain the signed log-likelihood ratio statistic $R(\psi) = \text{sgn}(\hat{\psi} - \psi)\sqrt{2(l(\hat{\theta}) - l(\hat{\theta}_\psi))} = \text{sgn}(\hat{\psi} - \psi)\sqrt{-T\log(\hat{\sigma}^2) + T\log(\tilde{\sigma}^2) + \hat{\alpha}\psi}$, the newly calibrated parameter $\chi(\boldsymbol{\theta}) = \frac{\mu}{\sigma^2}\tilde{\sigma} - \frac{1}{\sigma^2}\frac{(\tilde{\mu}+r_f)\tilde{\sigma}}{2}$, the numerator of $\hat{var}(\chi(\hat{\boldsymbol{\theta}}) - \chi(\hat{\boldsymbol{\theta}}_\psi))$ as $\frac{2T\tilde{\sigma}+\psi T\bar{w}}{4}\tilde{\sigma}^5$ and the denominator of $\hat{var}(\chi(\hat{\boldsymbol{\theta}}) - \chi(\hat{\boldsymbol{\theta}}_\psi))$ as $\frac{T^2\tilde{\sigma}^6}{2}$. Finally, we get all of them to the proposed third order likelihood approximation-the Barndorff-Nielsen method in (2.6.13).

3.1.2 Simulations for One Sample Sharpe Ratio

3.1.2.1 Reference Group of Existing Methodology

We introduced a couple of existing literature and their methodology to make inference of Sharpe ratio at Chapter One. Here in order to illustrate the exceptional accuracy of our proposed methodology, we construct the following ones as our reference group on numerical studies.

1. Jobson and Korkie (1981) and Lo (2002) derived the asymptotic distribution of estimated Sharpe ratio given IID returns (1.2.4):

$$\sqrt{n}(\widehat{SR} - SR) \xrightarrow{d} N(0, 1 + \frac{1}{2}\widehat{SR}^2) ;$$

2. Mertens (2002) enhanced Lo's result and obtained (1.2.5):

$$\sqrt{n}(\widehat{SR} - SR) \xrightarrow{d} N(0, 1 + \frac{\widehat{SR}^2}{2} - \hat{\alpha}_3\widehat{SR} + \frac{\hat{\alpha}_4 - 3}{4}\widehat{SR}^2) ;$$

3. The signed log-likelihood ratio statistic in (2.3.15):

$$\begin{aligned} R(\psi) &= \text{sgn}(\hat{\psi} - \psi)\sqrt{2(l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_\psi))} \\ &= \text{sgn}(\hat{\psi} - \psi)\sqrt{-T\log(\hat{\sigma}^2) + T\log(\tilde{\sigma}^2) + \hat{\alpha}\psi} ; \end{aligned}$$

4. We also apply the analysis introduced by Abramowitz and Stegun (1964 p949) who showed a very neat normality result (1.2.6) coming from the Akahira's approximation (1995) up to $O\left(n^{-1}\right)$.

$$\widehat{SR}\left(1 - \frac{1}{4\left(n-1\right)}\right) \xrightarrow{d} N\left(SR, \frac{1}{n} + \frac{\widehat{SR}^2}{2\left(n-1\right)}\right) .$$

Note that results 1, 2 and 3 are all first order approximation $O(n^{-\frac{1}{2}})$,

result 4 is second order approximation $O(n^{-1})$, while our proposed method is the third order approximation $O(n^{-\frac{3}{2}})$, meaning that theoretically proposed method is more valid and accurate than the above members in reference group. In addition, Jensen(1992) proved the asymptotical equivalence between Lugannani and Rice and Barndorff-Nielsen's approximation, and the results from the two methods are very much close to each other. Thus for all our examples and simulations in this thesis the proposed method will mainly focus on Barndorff-Nielsen's approximation.

3.1.2.2 Numerical Studies

Our first simulation study is to compare the accuracy of the confidence intervals obtained from the reference group of existing methodology and those obtained by proposed methods. For each combinations of $n = 4, 12$, $\mu = -1, 0, 1$, $\sigma^2 = 0.05, 1$ and $r_f = 0$, ten thousand Monte Carlo replications are performed. For each generated sample, the 95% confidence interval for Sharpe ratio is calculated.

The performance of a method is judged using the following criteria 1-6. The desired values are 0.95, 0.025, 0.025,0, 0 and 1 respectively. These values reflect the desired properties of the accuracy and symmetry of the interval estimates of Sharpe ratio.

1. The central coverage probability (CP): Proportion of the true Sharpe ratio falls within the 95% confidence interval;
2. The lower tail error rate (LE): Proportion of the true Sharpe ratio falls below the lower limit of the 95% confidence interval;
3. The upper tail error rate (UE): Proportion of the true Sharpe ratio falls above the upper limit of the 95% confidence interval;
4. The average bias (AB): It is defined as $AB = \frac{|LE-0.025|+|UE-0.025|}{2}$.
5. The average bias per unit of standard error (AB/SE): AB can also be quantified by taking reference of the standard error (SE), which comes from the belief that a value falling into any interval follows Bernoulli distribution and hence for our simulation the standard error is $\sqrt{\frac{0.025(1-0.025)}{10000}} = 0.0016$.
6. The degree of symmetry (SY): It is defined as $SY = \max\left\{\frac{LE}{UE}, \frac{UE}{LE}\right\}$

Results are recorded in Table (3.1) and Table (3.2). We can conclude from the simulation that the proposed modified signed log likelihood ratio method gives excellent results and outperforms the other four methods in all six criteria even for extreme sample size case:

- Average Bias and Central Coverage: For the case of $n = 4$, the proposed method produces results that are uniformly within 1.25 units of

Table 3.1: Simulation Result for One Sample Sharpe Ratio under IID Normal Return $n = 4$

Setting	Method	CP	LE	UE	AB	AB/SE	SY
$\mu = -1, \sigma = 0.05$	Lo	0.9345	0.0110	0.0545	0.0218	13.59	4.95
	Mertens	0.9961	0.0027	0.0012	0.0231	14.41	2.25
	Likelihood Ratio	0.8757	0.0086	0.1157	0.0536	33.47	13.45
	Abramowitz&Stegun	0.9866	0.0102	0.0032	0.0183	11.44	3.19
	Proposed	**0.9467**	**0.0249**	**0.0284**	**0.0018**	**1.09**	**1.14**
$\mu = 0, \sigma = 0.05$	Lo	0.9024	0.0467	0.0509	0.0238	14.88	1.09
	Mertens	0.9556	0.0210	0.0234	0.0028	1.75	1.11
	Likelihood Ratio	0.8879	0.0538	0.0583	0.0311	19.41	1.08
	Abramowitz&Stegun	0.9716	0.0133	0.0151	0.0108	6.75	1.14
	Proposed	**0.9496**	**0.0240**	**0.0264**	**0.0012**	**0.75**	**1.10**
$\mu = 1, \sigma = 0.05$	Lo	0.9362	0.0554	0.0084	0.0235	14.69	6.60
	Mertens	0.9969	0.0005	0.0026	0.0235	14.66	5.20
	Likelihood Ratio	0.8772	0.1166	0.0062	0.0552	34.50	18.81
	Abramowitz&Stegun	0.9891	0.0030	0.0079	0.0196	12.22	2.63
	Proposed	**0.9487**	**0.0272**	**0.0241**	**0.0016**	**0.97**	**1.13**
$\mu = -1, \sigma = 1$	Lo	0.9196	0.0250	0.0554	0.0152	9.50	2.22
	Mertens	0.9738	0.0251	0.0011	0.0120	7.50	22.82
	Likelihood Ratio	0.8840	0.0250	0.0910	0.0330	20.63	3.64
	Abramowitz&Stegun	0.9704	0.0250	0.0046	0.0102	6.38	5.43
	Proposed	**0.9471**	**0.0253**	**0.0276**	**0.0015**	**0.91**	**1.09**
$\mu = 0, \sigma = 1$	Lo	0.9036	0.0468	0.0496	0.0232	14.50	1.06
	Mertens	0.9518	0.0243	0.0239	0.0009	0.56	1.02
	Likelihood Ratio	0.8887	0.0553	0.0560	0.0307	19.16	1.01
	Abramowitz&Stegun	0.9698	0.0146	0.0156	0.0099	6.19	1.07
	Proposed	**0.9488**	**0.0248**	**0.0264**	**0.0008**	**0.50**	**1.06**
$\mu = 1, \sigma = 1$	Lo	0.9199	0.0554	0.0247	0.0154	9.59	2.24
	Mertens	0.9743	0.0011	0.0246	0.0122	7.59	22.36
	Likelihood Ratio	0.8856	0.0897	0.0247	0.0325	20.31	3.63
	Abramowitz&Stegun	0.9713	0.0038	0.0249	0.0107	6.66	6.55
	Proposed	**0.9460**	**0.0287**	**0.0253**	**0.0020**	**1.25**	**1.13**

Table 3.2: Simulation Result for One Sample Sharpe Ratio under IID Normal Return $n = 12$

Setting	Method	CP	LE	UE	AB	AB/SE	SY
$\mu = -1, \sigma = 0.05$	Lo	0.9433	0.0159	0.0408	0.0125	7.78	2.57
	Mertens	0.9960	0.0023	0.0017	0.0230	14.38	1.35
	Likelihood Ratio	0.9303	0.0127	0.0570	0.0222	13.84	4.49
	Abramowitz&Stegun	0.9592	0.0167	0.0241	0.0046	2.88	1.44
	Proposed	**0.9489**	**0.0251**	**0.0260**	**0.0005**	**0.34**	**1.04**
$\mu = 0, \sigma = 0.05$	Lo	0.9356	0.0325	0.0319	0.0072	4.50	1.02
	Mertens	0.9522	0.0254	0.0224	0.0015	0.94	1.13
	Likelihood Ratio	0.9341	0.0331	0.0328	0.0080	4.97	1.01
	Abramowitz&Stegun	0.9439	0.0291	0.0270	0.0031	1.91	1.08
	Proposed	**0.9506**	**0.0252**	**0.0242**	**0.0005**	**0.31**	**1.04**
$\mu = 1, \sigma = 0.05$	Lo	0.9406	0.0414	0.0180	0.0117	7.31	2.30
	Mertens	0.9960	0.0017	0.0023	0.0230	14.38	1.35
	Likelihood Ratio	0.9312	0.0540	0.0148	0.0196	12.25	3.65
	Abramowitz&Stegun	0.9576	0.0240	0.0184	0.0038	2.38	1.30
	Proposed	**0.9469**	**0.0261**	**0.0270**	**0.0016**	**0.97**	**1.03**
$\mu = -1, \sigma = 1$	Lo	0.9432	0.0200	0.0368	0.0084	5.25	1.84
	Mertens	0.9761	0.0189	0.0050	0.0131	8.16	3.78
	Likelihood Ratio	0.9361	0.0199	0.0440	0.0121	7.53	2.21
	Abramowitz&Stegun	0.9535	0.0212	0.0253	0.0021	1.28	1.19
	Proposed	**0.9515**	**0.0252**	**0.0233**	**0.0009**	**0.59**	**1.08**
$\mu = 0, \sigma = 1$	Lo	0.9367	0.0312	0.0321	0.0067	4.16	1.03
	Mertens	0.9524	0.0234	0.0242	0.0012	0.75	1.03
	Likelihood Ratio	0.9355	0.0317	0.0328	0.0073	4.53	1.03
	Abramowitz&Stegun	0.9441	0.0279	0.0280	0.0030	1.84	1.00
	Proposed	**0.9510**	**0.0241**	**0.0249**	**0.0005**	**0.31**	**1.03**
$\mu = 1, \sigma = 1$	Lo	0.9417	0.0378	0.0205	0.0087	5.41	1.84
	Mertens	0.9740	0.0069	0.0191	0.0120	7.50	2.77
	Likelihood Ratio	0.9354	0.0441	0.0205	0.0118	7.38	2.15
	Abramowitz&Stegun	0.9516	0.0266	0.0218	0.0024	1.50	1.22
	Proposed	**0.9492**	**0.0251**	**0.0257**	**0.0004**	**0.25**	**1.02**

standard deviation while Abramowitz and Stegun's method produces around 6-12 units of standard deviation, and other three first order method produce even less satisfactory results. For the case of $n = 12$, our proposed method produces average bias within even one unit of standard deviation, while Abramowitz and Stegun produces 1-3 units of standard deviation and other three method give worse results. On the other hand, the proposed method create decent central coverage probability even for extreme sample size $n = 4$. The existing methodology in reference group give much less unsatisfactory coverage probability for $n = 4$, however, their results improve as sample size rises.

- Sample Size Effect on Central Coverage: In order to make this size effect clearly visible, a second round of simulation is being conducted, setting $\mu = 1$, $\sigma = 1$, and the sample size to rise from $n = 2$ to $n = 500$. The result is recorded on Figure (3.1). Suppose we take 3 units of standard deviation as our acceptance level on AB, we can find that our proposed result can achieve this level even for extreme sample size $n = 4$ while the reference group may need at least a sample of $n = 100$. Another result that needs to be discussed is Mertens method which show a weak decaying trend as sample size rises. Theoretically for normally distributed returns, the skewness and historical kurtosis of the returns distribution are both zero, and so Mertens' form reduces to Lo's form. These are, however, unknown in practice, and have to be estimated from the data, which results in some mis-estimation of the standard error of Sharpe ratio when skew is extreme.

- The Degree of Symmetry: We can conclude from Table (3.1) and Table (3.2) that, besides unsatisfactory coverage probability, the existing methodology in reference group also give asymmetric intervals. Figure (3.2) is recording the sample size effect on SY, setting $\mu = 1$, $\sigma = 1$, and the sample size to rise from $n = 2$ to $n = 500$. We can see from this figure that even at very extreme case $n = 3$, our proposed method already obtains a good degree of symmetry of intervals, about 11% of relative difference between LE and RE (or $\frac{\max\{LE, RE\} - \min\{LE, RE\}}{\min\{LE, RE\}}$=11%). While all the other reference group may need at least about 100 sample size to make LE and RE have 50% of such relative difference.

- The Central Effect: If we read Table (3.1) and Table (3.2) carefully, we can find that whenever $\mu = 0$, the result of reference group could improve significantly. What is this effect? We conduct another round of simulation to treat this problem and the result is in Figure(3.3). We can see that for the methodology in reference group, when Sharpe ratio approaches its central position around "0", the average bias decreases a lot and the result of simulation improves significantly. On the other hand, However, our proposed method looks pretty flat across

Figure 3.1: The Effect of Sample Size on AB/ER, $\mu = 1$ and $\sigma = 1$

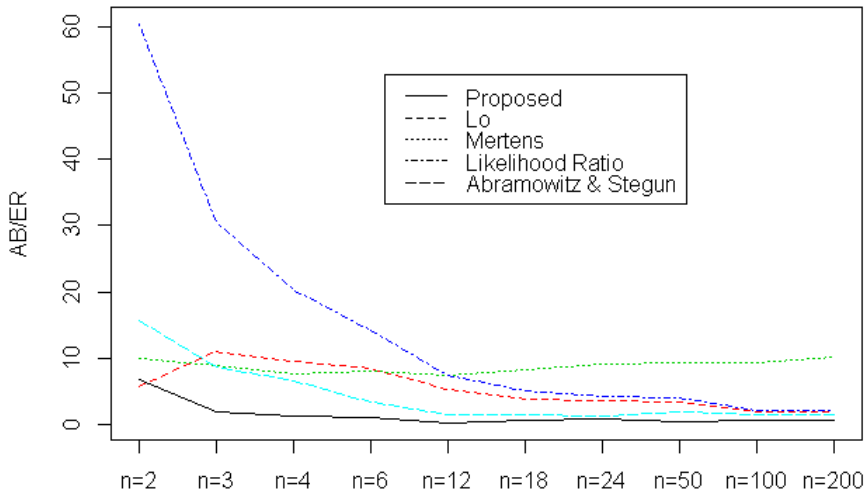
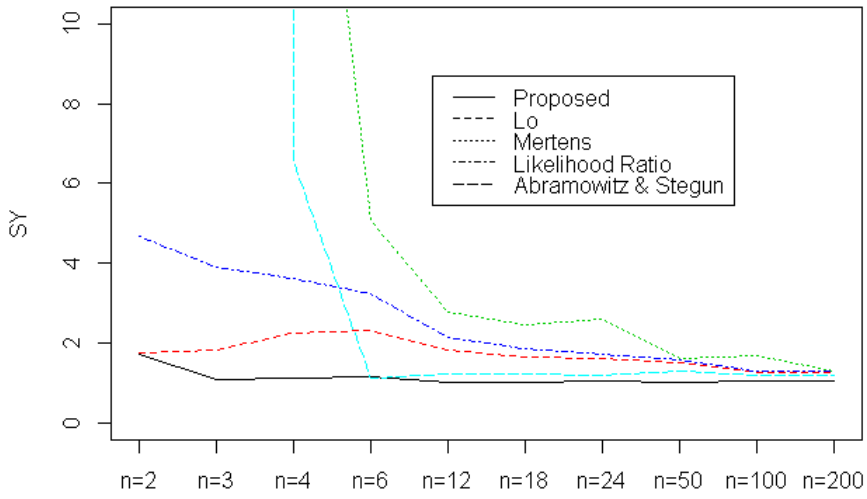


Figure 3.2: The Effect of Sample Size on SY, $\mu = 1$ and $\sigma = 1$



the whole domain, proving its great efficiency in making inference about Sharpe ratio.

3.1.3 Examples for One Sample Sharpe Ratio

In this subsection, we provide examples for inference on Sharpe ratio under IID normal return. The data set used at this chapter consists of two time series of monthly return from Aug 2014 to July 2015 and they are listed at Table (3.3). The first series represents the return for a large-cap mutual fund (Fund), a measure of the average return of all hedge funds (excepting Funds of Funds) in the Barclay Hedge Fund database[2]. The second series is the market index (Market), the monthly return of the S&P 500 index [3]. And finally we will use the average monthly return of 3-Month Treasury Bill during the above period as the risk-free rate, and its value is $r_f = 0.000242$. [4]

The likelihood methodology constructed in this Chapter is based on the normality of return, therefore, before conducting any statistical inference, we need to make sure the normality of the given data. A couple of tests are being conducted to achieve this goal and their corresponding p-value are shown in Table (3.4). Since the null hypothesis for all these tests is the data are normal, and all of the listed p-values are great than 0.1, thus we conclude a failure to reject the null and there is no enough evidence to go against the normal assumption for both Fund and Market.

Although a general simulation has been performed at last subsection, here we do another round of simulation under the setting of our example in order to validate our statistical inference. We mimic the mutual fund return data and market index return with Fund: $N(0.0009050833, 0.00001932047)$ and Market: $N(0.002995667, 0.0001244846)$, and $r_f = 0.000242$. 90%, 95%, and 99% confidence intervals for the Sharpe ratio were obtained for each sample. Table (3.5) and Table (3.6) report the results from these simulations for the fund and market returns, respectively. From these results tables it is clear that, again, our proposed method outperforms the other methods based on the criteria we examined.

Our first application focuses on confidence intervals for Sharpe ratio. Table (3.7) reports 95% confidence intervals for Sharpe ratio separately for the large cap mutual fund and the market index. We can find that the confidence intervals obtained from the five methods produce different results. In particular, our proposed method gives more accurate confidence interval compared with the reference methodology, which can be seen from:

[2]The index is simply the arithmetic average of the net returns of all the funds that have reported that month and it has been converted to the log-return for our calculation. Source: http://www.barclayhedge.com/research/indices/ghs/Hedge_Fund_Index.html.

[3]Source: http://finance.yahoo.com/q/hp?s=%5EGSPC&a=11&b=1&c=2012&d=00&e=1&f=2014&g=m

[4]Source: Board of Governors of the Federal Reserve System (US), 3-Month Treasury Bill: Secondary Market Rate [TB3MS], retrieved from FRED, Federal Reserve Bank of St.Louis https://research.stlouisfed.org/fred2/series/TB3MS, March 4, 2016.

Figure 3.3: The Central Effect of One Sample Sharpe Ratio under IID Normal Return, $n = 12$

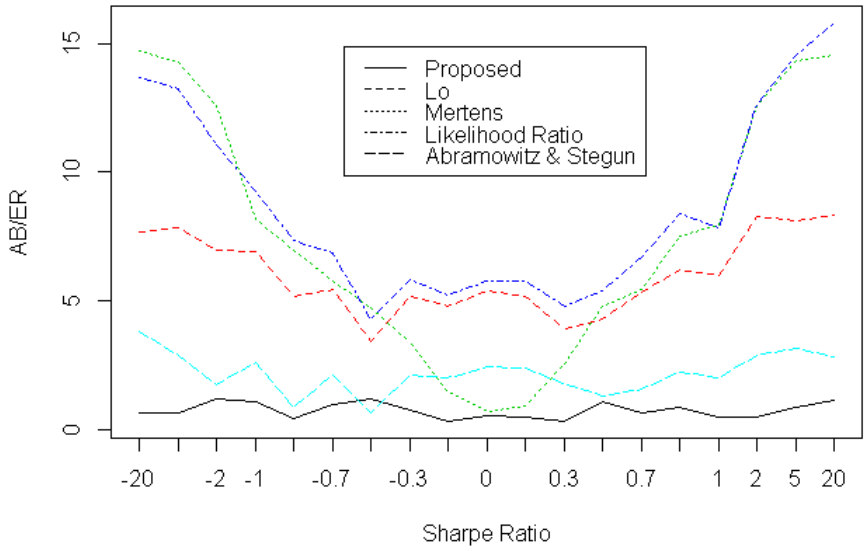


Table 3.3: Monthly return for Fund and Market

Month	logreturn Fund	logreturn Market
1	0.005009	0.016249
2	-0.005507	-0.006942
3	-0.001261	0.010148
4	0.002598	0.010492
5	-0.001828	-0.001449
6	-0.000435	-0.013695
7	0.009451	0.022843
8	0.001820	-0.007772
9	0.004665	0.003739
10	0.003504	0.004143
11	-0.004760	-0.009480
12	-0.002395	0.007672

Table 3.4: p-value of the test for normality on Fund and Market

Test	Fund	Market
Shapiro-Wilk test for normality	0.9212	0.8919
Anderson-Darling test for normality	0.8973	0.8486
Cramer-von Mises test for normality	0.8946	0.8275
Lilliefors (Kolmogorov-Smirnov) test for normality	0.9054	0.6753
Pearson chi-square test for normality	0.5724	0.8013
Shapiro-Francia test for normality	0.9042	0.9272

Table 3.5: Simulation Result for One Sample Sharpe Ratio under IID Normal Fund Return $n = 12$

CI	Method	CP	LE	UE	AB	AB/SE	SY
95%	Lo	0.9342	0.0337	0.0321	0.0079	4.94	1.05
	Mertens	0.9531	0.0213	0.0256	0.0022	1.34	1.20
	Likelihood Ratio	0.9325	0.0350	0.0325	0.0088	5.47	1.08
	Abramowitz&Stegun	0.9426	0.0287	0.0287	0.0037	2.31	1.00
	Proposed	**0.9489**	**0.0253**	**0.0258**	**0.0005**	**0.34**	**1.02**
90%	Lo	0.8761	0.0658	0.0581	0.0120	5.43	1.13
	Mertens	0.8971	0.0500	0.0529	0.0015	0.66	1.06
	Likelihood Ratio	0.8750	0.0669	0.0581	0.0125	5.68	1.15
	Abramowitz&Stegun	0.8866	0.0585	0.0549	0.0067	3.05	1.07
	Proposed	**0.8992**	**0.0511**	**0.0497**	**0.0007**	**0.32**	**1.03**
99%	Lo	0.9848	0.0067	0.0085	0.0026	3.71	1.27
	Mertens	0.9936	0.0018	0.0046	0.0018	2.57	2.56
	Likelihood Ratio	0.9832	0.0080	0.0088	0.0034	4.86	1.10
	Abramowitz&Stegun	0.9878	0.0052	0.0070	0.0011	1.57	1.35
	Proposed	**0.9889**	**0.0049**	**0.0062**	**0.0007**	**0.93**	**1.27**

Table 3.6: Simulation Result for One Sample Sharpe Ratio under IID Normal Market Return $n = 12$

Setting	Method	CP	LE	UE	AB	AB/SE	SY
95%	Lo	0.9406	0.0372	0.0222	0.0075	4.69	1.68
	Mertens	0.9696	0.0078	0.0226	0.0098	6.13	2.90
	Likelihood Ratio	0.9356	0.0422	0.0222	0.0100	6.25	1.90
	Abramowitz&Stegun	0.9490	0.0274	0.0236	0.0019	1.19	1.16
	Proposed	**0.9482**	**0.0261**	**0.0257**	**0.0009**	**0.56**	**1.02**
90%	Lo	0.8848	0.0720	0.0432	0.0144	6.55	1.67
	Mertens	0.9298	0.0274	0.0428	0.0149	6.77	1.56
	Likelihood Ratio	0.8798	0.077	0.0432	0.0169	7.68	1.78
	Abramowitz&Stegun	0.8976	0.057	0.0454	0.0058	2.64	1.26
	Proposed	**0.8997**	**0.0491**	**0.0512**	**0.0011**	**0.48**	**1.04**
99%	Lo	0.9869	0.0075	0.0056	0.0016	2.21	1.34
	Mertens	0.994	0.0004	0.0056	0.0026	3.71	14.00
	Likelihood Ratio	0.9830	0.0114	0.0056	0.0035	5.00	2.04
	Abramowitz&Stegun	0.9902	0.0042	0.0056	0.0007	1.00	1.33
	Proposed	**0.9887**	**0.0055**	**0.0058**	**0.0007**	**0.93**	**1.05**

(1)theoretically, the proposed method has third-order accuracy whereas the remaining four methods do not. This result has also been borne out in the simulations as well; (2) our proposed method produces narrower confidence interval compared with the other method on average.

The p-value functions calculated from each methods are plotted in Figures (3.4). These significance functions can be used to obtain p-values for specific hypothesized values of the Sharpe ratio, which are shown in Table (3.8) and (3.9). We can see that the p-values vary across the methods and people can result in different conclusions from the application of difference methods. For example, for Fund data, we want to test:

$$H_0 : SR = -0.6 \text{ versus } H_1 : SR \neq -0.6$$

The corresponding p-values for our proposed method is $2 \times (1 - 0.9949) > 0.01$, and we do not reject the null at 99% significance level, while all the other methods reject the null at 99% significance level.[5] As we are typically interested in tail probabilities which tend to be very small, it is important to estimate such probabilities with precision, and our proposed method is believable to give more convincing value.

Liu, Y., Rekkas, M., & Wong, A. (2012) conduct similar research based on the Fund and Market data from Matlab Financial ToolboxTMUser's Guide.

3.1.4 Likelihood Methodology for Two Independent Sample Comparison of Sharpe Ratio

Suppose one is interested in testing hypotheses concerning the comparison of Sharpe ratio of two independent funds X and Y. For instance, one may be interested in testing the null hypothesis $SR_X \geq SR_Y$ against $SR_X < SR_Y$; or, testing the null $SR_X = SR_Y$ against the alternative hypothesis $SR_X \neq SR_Y$. In fact, this is more useful and practical to know, because in most applicable conditions people need to identify their preferable investment from two or more available strategies at hand. In this subsection we apply the third order methodology introduced in Chapter 2 to test the difference of Sharpe ratio of independent funds X and Y.

Consider two funds with sample log-returns $(x_1, \ldots, x_m)$ and $(y_1, \ldots, y_n)$. Assume that these returns are identically and independently distributed as $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$, respectively, and thus $\boldsymbol{\theta}' = (\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2)$. The canonical parameter, augmented from (3.1.1), and its first order derivative will be the followings:

$$\boldsymbol{\varphi}'(\boldsymbol{\theta}) = \left(\frac{\mu_X}{\sigma_X^2}, \frac{1}{\sigma_X^2}, \frac{\mu_Y}{\sigma_Y^2}, \frac{1}{\sigma_Y^2} \right) , \tag{3.1.22}$$

[5]However, all of the methods indicate H_1 is significant at 5% level of significance

Table 3.7: 95% Confidence Intervals for Sharpe Ratio

Method	CI for SR of Fund	CI Length	CI for SR of Market	CI Length
Lo	(-0.6386 0.4945)	1.1331	(-0.3173 0.8329)	1.1502
Mertens	(-0.6485 0.5045)	1.1530	(-0.3163 0.8319)	1.1482
Likelihood Ratio	(-0.6386 0.4945)	1.1331	(-0.3171 0.8331)	1.1502
Abramowitz&Stegun	(-0.6370 0.4962)	1.1332	(-0.3240 0.8279)	1.1519
Proposed	**(-0.6340 0.4989)**	**1.1329**	**(-0.3331 0.8168)**	**1.1498**

Table 3.8: p-values for One Sample Sharpe Ratio under IID Normal Fund Return

ψ	-0.6	-0.55	-0.5	-0.45	-0.4	-0.35	-0.3	-0.25
Lo	0.9954	0.9926	0.9882	0.9818	0.9725	0.9597	0.9424	0.9197
Mertens	0.9960	0.9933	0.9892	0.9832	0.9744	0.9620	0.9452	0.9229
Likelihood Ratio	0.9955	0.9926	0.9882	0.9818	0.9726	0.9597	0.9424	0.9197
A&S	0.9953	0.9923	0.9878	0.9812	0.9717	0.9586	0.9409	0.9177
Proposed	**0.9949**	**0.9919**	**0.9871**	**0.9802**	**0.9704**	**0.9567**	**0.9384**	**0.9146**

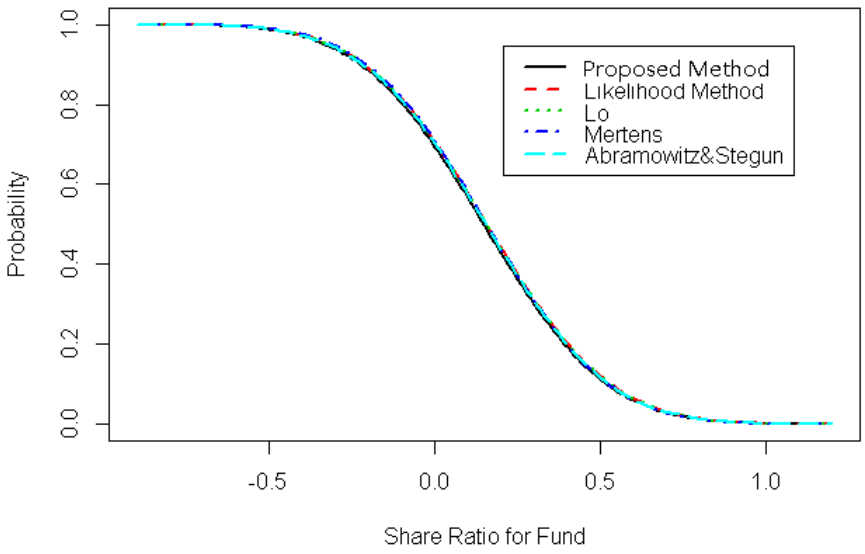
ψ	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9
Lo	0.0883	0.0639	0.0450	0.0309	0.0207	0.0135	0.0086	0.0053
Mertens	0.0851	0.0610	0.0426	0.0290	0.0192	0.0124	0.0077	0.0047
Likelihood Ratio	0.0883	0.0639	0.0450	0.0309	0.0207	0.0135	0.0086	0.0053
A&S	0.0865	0.0624	0.0439	0.0301	0.0201	0.0131	0.0083	0.0051
Proposed	**0.0830**	**0.0597**	**0.0419**	**0.0286**	**0.0191**	**0.0124**	**0.0078**	**0.0048**

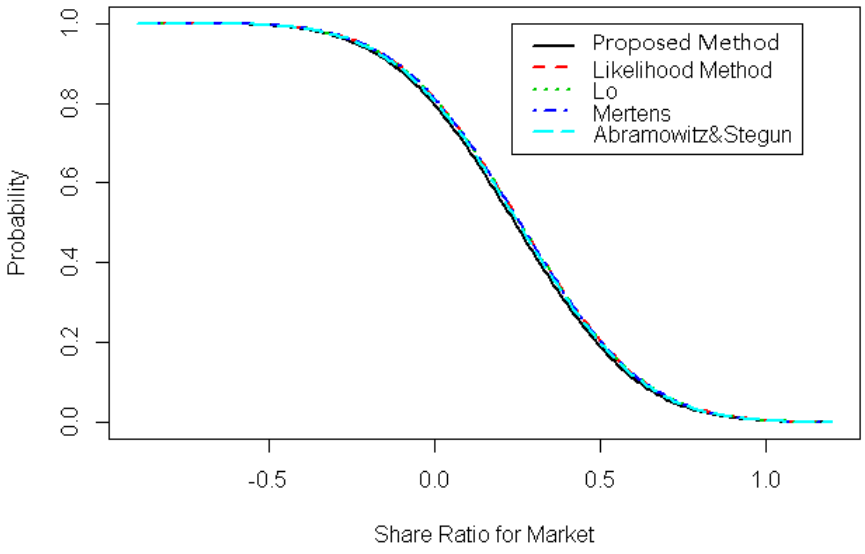
Table 3.9: p-values for One Sample Sharpe Ratio under IID Normal Market Return

ψ	-0.5	-0.45	-0.4	-0.35	-0.3	-0.25	-0.2	-0.15
Lo	0.9951	0.9921	0.9875	0.9808	0.9713	0.9582	0.9406	0.9177
Mertens	0.9952	0.9922	0.9876	0.9810	0.9716	0.9585	0.9410	0.9181
Likelihood Ratio	0.9951	0.9921	0.9875	0.9809	0.9714	0.9583	0.9407	0.9177
A&S	0.9947	0.9915	0.9867	0.9797	0.9698	0.9562	0.9380	0.9143
Proposed	**0.9943**	**0.9908**	**0.9857**	**0.9782**	**0.9676**	**0.9532**	**0.9340**	**0.9091**

ψ	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1
Lo	0.0907	0.0659	0.0467	0.0323	0.0218	0.0143	0.0092	0.0057
Mertens	0.0903	0.0655	0.0464	0.0321	0.0216	0.0142	0.0091	0.0056
Likelihood Ratio	0.0907	0.0660	0.0468	0.0324	0.0218	0.0143	0.0092	0.0057
A&S	0.0878	0.0637	0.0450	0.0311	0.0209	0.0137	0.0088	0.0055
Proposed	**0.0820**	**0.0591**	**0.0416**	**0.0285**	**0.0191**	**0.0124**	**0.0079**	**0.0049**

Figure 3.4: p-value function for One Sample Sharpe Ratio under IID Normal Return of Fund and Market





$$\boldsymbol{\varphi}_{\boldsymbol{\theta}'}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{\sigma_X^2} & -\frac{\mu_X}{\sigma_X^4} & 0 & 0 \\ 0 & -\frac{1}{\sigma_X^4} & 0 & 0 \\ 0 & 0 & \frac{1}{\sigma_Y^2} & -\frac{\mu_Y}{\sigma_Y^4} \\ 0 & 0 & 0 & -\frac{1}{\sigma_Y^4} \end{pmatrix} . \tag{3.1.23}$$

Later on, we will also use determinant of first order derivatives $|\boldsymbol{\varphi}_{\boldsymbol{\theta}'}(\boldsymbol{\theta})| = \frac{1}{\sigma_X^6 \sigma_Y^6}$, as well as its inverse matrix $\boldsymbol{\varphi}_{\boldsymbol{\theta}'}^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} \sigma_X^2 & -\mu_X\sigma_X^2 & 0 & 0 \\ 0 & -\sigma_X^4 & 0 & 0 \\ 0 & 0 & \sigma_Y^2 & -\mu_Y\sigma_Y^2 \\ 0 & 0 & 0 & -\sigma_Y^4 \end{pmatrix}$.

The mean return for the risk-free asset is represented by r_f again. Since our interest is to test the difference in Sharpe ratio, the parameter of interest ψ and its first order derivative are then defined as follows:

$$\psi(\boldsymbol{\theta}) = \frac{\mu_X - r_f}{\sigma_X} - \frac{\mu_Y - r_f}{\sigma_Y} , \tag{3.1.24}$$

$$\psi_{\boldsymbol{\theta}'}(\boldsymbol{\theta}) = \left(\frac{1}{\sigma_X}, -\frac{\mu_X - r_f}{2\sigma_X^3}, -\frac{1}{\sigma_Y}, \frac{\mu_Y - r_f}{2\sigma_Y^3} \right) . \tag{3.1.25}$$

For fund X, its log likelihood function is $l_X = a - \frac{m}{2}\log\sigma_X^2 - \frac{\sum(X_i-\mu_X)^2}{2\sigma_X^2}$; similarly log likelihood function for fund Y will be $l_Y = b - \frac{n}{2}\log\sigma_Y^2 - \frac{\sum(Y_j-\mu_Y)^2}{2\sigma_Y^2}$. The unrestricted maximum likelihood estimation maximizes the likelihood of both funds' sample or the joint log likelihood function $l(\boldsymbol{\theta})$:

$$l(\boldsymbol{\theta}) = l_X + l_Y = a + b - \frac{m}{2}\log\sigma_X^2 - \frac{\sum(X_i-\mu_X)^2}{2\sigma_X^2} - \frac{n}{2}\log\sigma_Y^2 - \frac{\sum(Y_j-\mu_Y)^2}{2\sigma_Y^2} . \tag{3.1.26}$$

The corresponding first order and second order derivatives of $l(\boldsymbol{\theta})$ are:

$$l_{\mu_X}(\boldsymbol{\theta}) = \frac{\sum(X_i - \mu_X)}{\sigma_X^2} \qquad l_{\mu_Y}(\boldsymbol{\theta}) = \frac{\sum(Y_j - \mu_Y)}{\sigma_Y^2}$$

$$l_{\sigma_X^2}(\boldsymbol{\theta}) = -\frac{m}{2\sigma_X^2} + \frac{\sum(X_i - \mu_X)^2}{2\sigma_X^4} \quad l_{\sigma_Y^2}(\boldsymbol{\theta}) = -\frac{n}{2\sigma_Y^2} + \frac{\sum(Y_j - \mu_Y)^2}{2\sigma_Y^4}$$

$$l_{\mu_X\mu_X}(\boldsymbol{\theta}) = -\frac{m}{\sigma_X^2} \qquad l_{\mu_Y\mu_Y}(\boldsymbol{\theta}) = -\frac{n}{\sigma_Y^2}$$

$$l_{\mu_X\sigma_X^2}(\boldsymbol{\theta}) = l_{\sigma_X^2\mu_X}(\boldsymbol{\theta}) = -\frac{\sum(X_i - \mu_X)}{\sigma_X^4} \quad l_{\mu_Y\sigma_Y^2}(\boldsymbol{\theta}) = l_{\sigma_Y^2\mu_Y}(\boldsymbol{\theta}) = -\frac{\sum(Y_j - \mu_Y)}{\sigma_Y^4}$$

$$l_{\sigma_X^2\sigma_X^2}(\boldsymbol{\theta}) = \frac{m}{2\sigma_X^4} - \frac{\sum(X_i - \mu_X)^2}{\sigma_X^6} \quad l_{\sigma_Y^2\sigma_Y^2}(\boldsymbol{\theta}) = \frac{n}{2\sigma_Y^4} - \frac{\sum(Y_j - \mu_Y)^2}{\sigma_Y^6}$$

Note that all the interactive second order derivatives are equal to zero because of the independence of samples. By setting first order derivatives equal to zero, we get the unconstrained maximum likelihood estimator $\hat{\boldsymbol{\theta}}$:

$$\hat{\boldsymbol{\theta}}' = (\hat{\mu}_X, \hat{\sigma}_X^2, \hat{\mu}_Y, \hat{\sigma}_Y^2) = \left(\frac{\sum X_i}{m}, \frac{\sum(X_i - \hat{\mu}_X)^2}{m}, \frac{\sum Y_j}{n}, \frac{\sum(Y_j - \hat{\mu}_Y)^2}{n} \right). \tag{3.1.27}$$

Knowing this unconstrained MLE, we can achieve other important variable for late use, such as the estimated parameter of interest $\hat{\psi} = \psi(\hat{\boldsymbol{\theta}})$, the estimated unrestricted likelihood function $l(\hat{\boldsymbol{\theta}})$, the observed information at $\hat{\boldsymbol{\theta}}$, $\mathbf{j}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} -l_{\mu_X\mu_X}(\hat{\boldsymbol{\theta}}) & -l_{\mu_X\sigma_X^2}(\hat{\boldsymbol{\theta}}) & 0 & 0 \\ -l_{\sigma_X^2\mu_X}(\hat{\boldsymbol{\theta}}) & -l_{\sigma_X^2\sigma_X^2}(\hat{\boldsymbol{\theta}}) & 0 & 0 \\ 0 & 0 & -l_{\mu_Y\mu_Y}(\hat{\boldsymbol{\theta}}) & -l_{\mu_Y\sigma_Y^2}(\hat{\boldsymbol{\theta}}) \\ 0 & 0 & -l_{\sigma_Y^2\mu_Y}(\hat{\boldsymbol{\theta}}) & -l_{\sigma_Y^2\sigma_Y^2}(\hat{\boldsymbol{\theta}}) \end{pmatrix} = \begin{pmatrix} \frac{m}{\hat{\sigma}_X^2} & 0 & 0 & 0 \\ 0 & \frac{m}{2\hat{\sigma}_X^4} & 0 & 0 \\ 0 & 0 & \frac{n}{\hat{\sigma}_Y^2} & 0 \\ 0 & 0 & 0 & \frac{n}{2\hat{\sigma}_Y^4} \end{pmatrix}$, as well as its determinant $\frac{m^2n^2}{4\hat{\sigma}_X^6\hat{\sigma}_Y^6}$. We can see from the derivation process that this two sample unconstrained maximization is quite similar with that in one sample introduced in subsection (3.1.1). This similarity actually results from the independence assumption of two funds. However, as we will see in the next steps, the constrained likelihood maximization is far more complicated compared with the previous results.

The constrained maximum likelihood estimators are solved by the La-

grange multiplier method (see (2.3.7)). The Lagrangian function is given at (3.1.28) and its first order derivatives are listed from (3.1.29) to (3.1.33). The tilted log-likelihood function can be obtained by replacing α by $\hat{\alpha}$ on the Lagrangian function, and its second order derivatives are given by (3.1.34) to (3.1.39).

$$H(\boldsymbol{\theta},\alpha) = l(\boldsymbol{\theta}) + \alpha(\psi(\boldsymbol{\theta}) - \psi) = l(\boldsymbol{\theta}) + \alpha\left(\frac{\mu_X - r_f}{\sigma_X} - \frac{\mu_Y - r_f}{\sigma_Y} - \psi\right) , \quad (3.1.28)$$

$$H_{\mu_X}(\boldsymbol{\theta},\alpha) = l_{\mu_X}(\boldsymbol{\theta}) + \frac{\alpha}{\sigma_X} , \quad (3.1.29)$$

$$H_{\sigma_X^2}(\boldsymbol{\theta},\alpha) = l_{\sigma_X^2}(\boldsymbol{\theta}) - \frac{\alpha(\mu_X - r_f)}{2\sigma_X^3} , \quad (3.1.30)$$

$$H_{\mu_Y}(\boldsymbol{\theta},\alpha) = l_{\mu_Y}(\boldsymbol{\theta}) - \frac{\alpha}{\sigma_Y} , \quad (3.1.31)$$

$$H_{\sigma_Y^2}(\boldsymbol{\theta},\alpha) = l_{\sigma_Y^2}(\boldsymbol{\theta}) + \frac{\alpha(\mu_Y - r_f)}{2\sigma_Y^3} , \quad (3.1.32)$$

$$H_{\alpha}(\boldsymbol{\theta},\alpha) = \frac{\mu_X - r_f}{\sigma_X} - \frac{\mu_Y - r_f}{\sigma_Y} - \psi , \quad (3.1.33)$$

$$\tilde{l}_{\mu_X\mu_X}(\boldsymbol{\theta}) = H_{\mu_X\mu_X}(\boldsymbol{\theta},\hat{\alpha}) = l_{\mu_X\mu_X}(\boldsymbol{\theta}) , \quad (3.1.34)$$

$$\tilde{l}_{\mu_X\sigma_X^2}(\boldsymbol{\theta}) = \tilde{l}_{\sigma_X^2\mu_X}(\boldsymbol{\theta}) = H_{\mu_X\sigma_X^2}(\boldsymbol{\theta},\hat{\alpha}) = l_{\mu_X\sigma_X^2}(\boldsymbol{\theta}) - \frac{\hat{\alpha}}{2\sigma_X^3} , \quad (3.1.35)$$

$$\tilde{l}_{\sigma_X^2\sigma_X^2}(\boldsymbol{\theta}) = H_{\sigma_X^2\sigma_X^2}(\boldsymbol{\theta},\hat{\alpha}) = l_{\sigma_X^2\sigma_X^2}(\boldsymbol{\theta}) + \frac{3}{4}\frac{\hat{\alpha}(\mu_X - r_f)}{\sigma_X^5} , \quad (3.1.36)$$

$$\tilde{l}_{\mu_Y\mu_Y}(\boldsymbol{\theta}) = H_{\mu_Y\mu_Y}(\boldsymbol{\theta},\hat{\alpha}) = l_{\mu_Y\mu_Y}(\boldsymbol{\theta}) , \quad (3.1.37)$$

$$\tilde{l}_{\mu_Y\sigma_Y^2}(\boldsymbol{\theta}) = \tilde{l}_{\sigma_Y^2\mu_Y}(\boldsymbol{\theta}) = H_{\mu_Y\sigma_Y^2}(\boldsymbol{\theta},\hat{\alpha}) = l_{\mu_Y\sigma_Y^2}(\boldsymbol{\theta}) + \frac{\hat{\alpha}}{2\sigma_Y^3} , \quad (3.1.38)$$

$$\tilde{l}_{\sigma_Y^2\sigma_Y^2}(\boldsymbol{\theta}) = H_{\sigma_Y^2\sigma_Y^2}(\boldsymbol{\theta},\hat{\alpha}) = l_{\sigma_Y^2\sigma_Y^2}(\boldsymbol{\theta}) - \frac{3}{4}\frac{\hat{\alpha}(\mu_Y - r_f)}{\sigma_Y^5} . \quad (3.1.39)$$

Setting first order derivatives, (3.1.29) to (3.1.33), equal to zero and solving the resulting system produces the constrained MLE. Five unknown variables- four constrained MLE $\hat{\boldsymbol{\theta}}_{\psi}^{'} = (\tilde{\mu}_X, \tilde{\sigma}_X^2, \tilde{\mu}_Y, \tilde{\sigma}_Y^2)$ plus the estimated Lagrange estimator $\hat{\alpha}$ -can be obtained by solving a nonlinear system of five

equations. Unfortunately, we cannot obtain closed form analytical solutions like one sample case but numerical solutions by implementing Newton's Method (Richard L. Burden and J. Douglas Faires, 2012) with the starting values of iteration being unconstrained MLE $\hat{\boldsymbol{\theta}}^{'} = (\hat{\mu}_X, \hat{\sigma}_X^2, \hat{\mu}_Y, \hat{\sigma}_Y^2)$. By using constrained MLE, we can thus obtain the estimated restricted likelihood function: $l(\hat{\boldsymbol{\theta}}_\psi)$, the tilted observed information matrix evaluated at $\hat{\boldsymbol{\theta}}_\psi$, $\tilde{\mathbf{j}}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_\psi) = -\frac{\partial^2 \tilde{\mathbf{j}}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_\psi} = j_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_\psi) - \hat{\alpha}\psi_{\boldsymbol{\theta}\boldsymbol{\theta}'}\left(\hat{\boldsymbol{\theta}}_\psi\right) = j_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_\psi) +$

$$\begin{pmatrix} 0 & \frac{\hat{\alpha}}{2\tilde{\sigma}_X^3} & 0 & 0 \\ \frac{\hat{\alpha}}{2\tilde{\sigma}_X^3} & -\frac{3\hat{\alpha}(\tilde{\mu}_X - r_f)}{4\tilde{\sigma}_X^5} & 0 & 0 \\ 0 & 0 & 0 & -\frac{\hat{\alpha}}{2\tilde{\sigma}_Y^3} \\ 0 & 0 & -\frac{\hat{\alpha}}{2\tilde{\sigma}_Y^3} & \frac{3\hat{\alpha}(\tilde{\mu}_Y - r_f)}{4\tilde{\sigma}_Y^5} \end{pmatrix}$$

and its inverse $\tilde{\mathbf{j}}^{-1}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_\psi)$.

Given the above information, R can be constructed from (2.3.15), χ can be calculated from (2.6.17), Q can be obtained from (2.6.18) and (2.6.19), and finally $R^\star$ can be obtained from (2.6.13).

3.1.5 Simulations for Two Independent Sample Comparison on Sharpe Ratio

3.1.5.1 Reference Group of Existing Methodology

In order to illustrate the exceptional accuracy of our proposed method, we construct the following as our reference group of methodology. These existing methodology correspond to the methods in part (3.1.2.1).

1. Jobson and Korkie (1981) and Lo (2002): Suppose we have two independent samples X and Y. Each sample will have its own Sharpe ratio's asymptotic distribution, $\widehat{SR}_X \xrightarrow{d} N\left(SR_X, \frac{1}{m}(1 + \frac{1}{2}\hat{SR}_X^2)\right)$ and $\widehat{SR}_Y \xrightarrow{d} N\left(SR_Y, \frac{1}{n}(1 + \frac{1}{2}\hat{SR}_Y^2)\right)$, and then distribution of the difference can be written as

$$\widehat{SR}_X - \widehat{SR}_Y \xrightarrow{d} N\left(SR_X - SR_Y, \frac{1}{m}(1 + \frac{1}{2}\widehat{SR}_X^2) + \frac{1}{n}(1 + \frac{1}{2}\widehat{SR}_Y^2)\right) ;$$

2. Mertens (2002): Suppose we have two independent samples X and Y. Each sample will have its own Sharpe ratio's asymptotic distribution, $\widehat{SR} \xrightarrow{d} N\left(SR, \hat{\sigma}_{SR}^2\right)$ with $\hat{\sigma}_{SR}^2 = \frac{1}{n}\left(1 + \frac{\widehat{SR}^2}{2} - \hat{\alpha}_3\widehat{SR} + \frac{\hat{\alpha}_4 - 3}{4}\widehat{SR}^2\right)$, and then

distribution of the difference of Sharpe ratio can be written as

$$\widehat{SR}_X - \widehat{SR}_Y \xrightarrow{d} N\Bigg(SR_X - SR_Y, \\ \frac{1}{m}\left(1 + \frac{\widehat{SR}_X^2}{2} - \hat{\alpha}_{3,X}\widehat{SR}_X + \frac{\hat{\alpha}_{4,X}-3}{4}\widehat{SR}_X^2\right) \\ + \frac{1}{n}\left(1 + \frac{\widehat{SR}_Y^2}{2} - \hat{\alpha}_{3,Y}\widehat{SR}_Y + \frac{\hat{\alpha}_{4,Y}-3}{4}\widehat{SR}_Y^2\right)\Bigg) ;$$

3. The signed log-likelihood ratio statistic in (2.3.15):

$$R(\psi) = \text{sgn}(\hat{\psi} - \psi)\sqrt{2(l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_\psi))} ;$$

4. Abramowitz and Stegun (1964 p949):

$$\widehat{SR}_X\left(1 - \frac{1}{4\,(m-1)}\right) - \widehat{SR}_Y\left(1 - \frac{1}{4\,(n-1)}\right) \\ = \text{sgn}(\hat{\psi} - \psi)\sqrt{-T\log(\hat{\sigma}^2) + T\log(\tilde{\sigma}^2) + \hat{\alpha}\psi} ;$$

Note that results 1, 2 and 3 are all first order approximation $O(n^{-\frac{1}{2}})$, result 4 is second order approximation $O(n^{-1})$, while our proposed method is the third order approximation $O(n^{-\frac{3}{2}})$, meaning that theoretically proposed method is more valid and accurate than the above members in reference group.

3.1.5.2 Numerical Study

In this part we provide a simulation study to assess the performance of our third order likelihood method relative to the existing methodology in reference group. For each combinations of $m, n = 6, 12$, $\mu = -1, 0, 1$, $\sigma^2 = 1$ and $r_f = 0$, ten thousand Monte Carlo replications are performed. And for each generated sample, the 95% confidence interval for the difference of Sharpe ratio is calculated. The performance of a method is judged using the same criteria 1-6 in (3.1.2.2).

Results are recorded in Table (3.10). We can conclude from the simulation that the proposed modified signed log likelihood ratio method gives excellent results and outperforms the other four methods based on the criteria we examined. The analysis is similar with the analysis in (3.1.2.2). We also have record of all the other simulation results besides our setting on the parameter and they all show the same excellent results of our proposed methods.

Table 3.10: Simulation Result for Difference of Sharpe Ratio under IID Normal Return

Setting	Method	CP	LE	UE	AB	AB/ER	SY
$m = 12$, $SR_X = 1$; $n = 12$, $SR_Y = 1$	Lo	0.9352	0.0325	0.0323	0.0074	4.63	1.01
	Mertens	0.9792	0.0099	0.0109	0.0146	9.13	1.10
	Likelihood Ratio	0.9309	0.0349	0.0342	0.0096	5.97	1.02
	Abramowitz&Stegun	0.9465	0.0268	0.0267	0.0018	1.09	1.00
	Proposed	**0.9486**	**0.0256**	**0.0258**	**0.0007**	**0.44**	**1.01**
$m = 12$, $SR_X = 1$; $n = 12$, $SR_Y = 0$	Lo	0.9285	0.0467	0.0248	0.0110	6.84	1.88
	Mertens	0.9656	0.0154	0.0190	0.0078	4.88	1.23
	Likelihood Ratio	0.9242	0.0510	0.0248	0.0131	8.19	2.06
	Abramowitz&Stegun	0.9410	0.0344	0.0246	0.0049	3.06	1.40
	Proposed	**0.9497**	**0.0243**	**0.0260**	**0.0009**	**0.53**	**1.07**
$m = 12$, $SR_X = 1$; $n = 12$, $SR_Y = -1$	Lo	0.9244	0.0600	0.0156	0.0222	13.88	3.85
	Mertens	0.9757	0.0120	0.0123	0.0129	8.03	1.03
	Likelihood Ratio	0.9192	0.0657	0.0151	0.0253	15.81	4.35
	Abramowitz&Stegun	0.9416	0.0415	0.0169	0.0123	7.69	2.46
	Proposed	**0.9514**	**0.0239**	**0.0247**	**0.0007**	**0.44**	**1.03**
$m = 6$, $SR_X = 1$; $n = 6$, $SR_Y = 1$	Lo	0.9263	0.0351	0.0386	0.0119	7.41	1.10
	Mertens	0.9829	0.0087	0.0084	0.0165	10.28	1.04
	Likelihood Ratio	0.9124	0.0403	0.0473	0.0188	11.75	1.17
	Abramowitz&Stegun	0.9570	0.0205	0.0225	0.0035	2.19	1.10
	Proposed	**0.9505**	**0.0234**	**0.0261**	**0.0014**	**0.84**	**1.12**
$m = 6$, $SR_X = 1$; $n = 6$, $SR_Y = 0$	Lo	0.9067	0.0649	0.0284	0.0217	13.53	2.29
	Mertens	0.9610	0.0181	0.0209	0.0055	3.44	1.15
	Likelihood Ratio	0.8965	0.0746	0.0289	0.0268	16.72	2.58
	Abramowitz&Stegun	0.9404	0.0349	0.0247	0.0051	3.19	1.41
	Proposed	**0.9470**	**0.0273**	**0.0257**	**0.0015**	**0.94**	**1.06**
$m = 6$, $SR_X = 1$; $n = 6$, $SR_Y = -1$	Lo	0.8912	0.0962	0.0126	0.0418	26.13	7.63
	Mertens	0.9699	0.0188	0.0113	0.0100	6.22	1.66
	Likelihood Ratio	0.8738	0.1137	0.0125	0.0506	31.63	9.10
	Abramowitz&Stegun	0.9421	0.0436	0.0143	0.0147	9.16	3.05
	Proposed	**0.9502**	**0.0269**	**0.0229**	**0.0020**	**1.25**	**1.17**
$m = 6$, $SR_X = 1$; $n = 12$, $SR_Y = 1$	Lo	0.9278	0.0307	0.0415	0.0111	6.94	1.35
	Mertens	0.9809	0.0134	0.0057	0.0155	9.66	2.35
	Likelihood Ratio	0.9145	0.0319	0.0536	0.0178	11.09	1.68
	Abramowitz&Stegun	0.9539	0.0260	0.0201	0.0030	1.84	1.29
	Proposed	**0.9468**	**0.0258**	**0.0274**	**0.0016**	**1.00**	**1.06**
$m = 6$, $SR_X = 1$; $n = 12$, $SR_Y = -1$	Lo	0.9136	0.0723	0.0141	0.0291	18.19	5.13
	Mertens	0.9741	0.0139	0.0120	0.0121	7.53	1.16
	Likelihood Ratio	0.9010	0.0850	0.0140	0.0355	22.19	6.07
	Abramowitz&Stegun	0.9484	0.0356	0.0160	0.0098	6.13	2.23
	Proposed	**0.9480**	**0.0272**	**0.0248**	**0.0012**	**0.75**	**1.10**
$m = 6$, $SR_X = 1$; $n = 12$, $SR_Y = 0$	Lo	0.9129	0.0547	0.0324	0.0186	11.59	1.69
	Mertens	0.9597	0.0224	0.0179	0.0049	3.03	1.25
	Likelihood Ratio	0.9089	0.0578	0.0333	0.0206	12.84	1.74
	Abramowitz&Stegun	0.9345	0.0402	0.0253	0.0078	4.84	1.59
	Proposed	**0.9478**	**0.0261**	**0.0261**	**0.0011**	**0.69**	**1.00**
$m = 12$, $SR_X = 1$; $n = 6$, $SR_Y = 0$	Lo	0.9213	0.0562	0.0225	0.0169	10.53	2.50
	Mertens	0.9704	0.0097	0.0199	0.0102	6.38	2.05
	Likelihood Ratio	0.9078	0.0698	0.0224	0.0237	14.81	3.12
	Abramowitz&Stegun	0.9511	0.0264	0.0225	0.0020	1.22	1.17
	Proposed	**0.9470**	**0.0281**	**0.0249**	**0.0016**	**1.00**	**1.13**

3.1.6 Examples for Two Independent Sample Comparison on Sharpe Ratio

In this subsection we provide an empirical example and a simulation study for inference on the difference of Sharpe ratio from two series of IID normal return. The data set used here is the same as Table (3.3) and we may, for instance, be interested in testing whether the risk-adjusted return as captured by the Sharpe ratio of mutual fund is significantly better than that of average stock market.

The likelihood methodology constructed in this part is based on the IID normality of return. Since in subsection (3.1.3), we have proved the normality of Fund return and Market return, here we will focus on the property of independence. We calculate their correlation coefficient and test the null $H_0 : \rho = 0$. The results shows that the estimated correlation coefficient is 0.6870 and the p-value of test is 0.01358. If we take a stricter manner and decide our significance at 99%, we can conclude that there is not enough evidence to reject the null hypothesis indicating the independence between Fund return and Market return.

Although a general simulation has been performed in last subsection, here we do another simulation under the setting of our example in order to validate our statistical inference. The parameter values were chosen to mimic the example data with Fund: $N(0.0009050833, 0.00001932047)$ and Market: $N(0.002995667, 0.0001244846)$, and $r_f = 0.000242$. 90%, 95%, and 99% confidence intervals for the difference of Sharpe ratio were obtained for each sample. Table (3.11) report the results from these simulations. From the result table it is clear that, again, our proposed method outperforms the other methods based on the criteria we examined.

Our first application focuses on 95% confidence intervals for difference of Sharpe ratio from the independent samples of Fund and Market, and the results are recorded at Table (3.12). We can find that the confidence intervals obtained from the five methods produce different results. In particular, our proposed method gives more accurate confidence interval compared with the reference methodology, which can be seen from: (1) theoretically, the proposed method has third-order accuracy whereas the remaining four methods do not. This result has also been borne out in the simulations as well; (2) our proposed method produces narrower confidence interval compared with the other method on average.

The p-value functions calculated from each methods are plotted in Figures (3.5). These significance functions can be used to obtain p-values for specific hypothesized values of the difference of Sharpe ratio, which are shown in Table (3.13). One of the most important hypothesized values is zero. For example, we want to test:

$$H_0 : SR_{Fund} - SR_{Mkt} \geq 0 \text{ versus } H_1 : SR_{Fund} - SR_{Mkt} < 0$$

The corresponding p-values for our proposed method is 0.4119, and we do

Table 3.11: Simulation Result for Difference of Sharpe Ratio under IID Normal Return of Fund and Market

CI	Method	CP	LE	UE	AB	AB/SE	SY
95%	Lo	0.9265	0.0325	0.0410	0.0118	7.34	1.26
	Mertens	0.9457	0.0251	0.0292	0.0022	1.34	1.16
	Likelihood Ratio	0.9253	0.0329	0.0418	0.0124	7.72	1.27
	Abramowitz&Stegun	0.9348	0.0303	0.0349	0.0076	4.75	1.15
	Proposed	**0.9437**	**0.0267**	**0.0296**	**0.0032**	**1.97**	**1.11**
90%	Lo	0.8695	0.0598	0.0707	0.0153	6.93	1.18
	Mertens	0.8881	0.0537	0.0582	0.0060	2.70	1.08
	Likelihood Ratio	0.8684	0.0602	0.0714	0.0158	7.18	1.19
	Abramowitz&Stegun	0.8789	0.0561	0.0650	0.0106	4.80	1.16
	Proposed	**0.8930**	**0.0518**	**0.0552**	**0.0035**	**1.59**	**1.07**
99%	Lo	0.9829	0.0076	0.0095	0.0036	5.07	1.25
	Mertens	0.9909	0.0048	0.0043	0.0005	0.64	1.12
	Likelihood Ratio	0.9815	0.0081	0.0104	0.0043	6.07	1.28
	Abramowitz&Stegun	0.9878	0.0056	0.0066	0.0011	1.57	1.18
	Proposed	**0.9873**	**0.0059**	**0.0068**	**0.0014**	**1.93**	**1.15**

Table 3.12: 95% Confidence Intervals for Difference of Sharpe Ratio

Method	CI for Difference in SR	CI Length
Lo	(-0.9094 0.7090)	1.6185
Mertens	(-0.9027 0.7022)	1.6049
Likelihood Ratio	(-0.9095 0.7089)	1.6185
Abramowitz&Stegun	(-0.9080 0.7121)	1.6201
Proposed	**(-0.9011 0.7170)**	**1.6181**

not reject the null at 95% significance level, and all the other methods have the same conclusion. The indifference on Sharpe ratio between investing on Fund and market is also the same conclusion from Liu, Y., Rekkas, M., & Wong, A. (2012), though they arrive this conclusion from a difference set of data. However, in general, we can see from Table (3.13) that the p-values vary across the methods and people can result in different conclusions from the application of difference methods.

At this subsection, the third order likelihood method is constructed under the assumption of independence between two samples. For example, if we take 99% as our significance level, we can make inference that there are no significant correlation between the return of Fund and Market. However, more general and realistic condition is that two investment options are not independent but correlated, and we are going to investigate this set of theory at next part.

3.1.7 Likelihood Methodology for Two Correlated Sample Comparison of Sharpe Ratio

Consider two funds with sample log-returns $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$. Assume that these returns are from a bivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$and ρ is the correlation coefficient, and thus $\boldsymbol{\theta}' = (\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$. The canonical parameter will be:

$$\boldsymbol{\varphi}'(\boldsymbol{\theta}) = \left(\frac{1}{(1-\rho^2)\,\sigma_X^2}, \frac{1}{(1-\rho^2)\,\sigma_Y^2}, \frac{\mu_X\sigma_Y - \mu_Y\sigma_X\rho}{(1-\rho^2)\,\sigma_X^2\sigma_Y}, \frac{\mu_Y\sigma_X - \mu_X\sigma_Y\rho}{(1-\rho^2)\,\sigma_Y^2\sigma_X}, \frac{\rho}{(1-\rho^2)\,\sigma_X\sigma_Y} \right). \tag{3.1.40}$$

We will also need its first order derivative $\boldsymbol{\varphi}_{\boldsymbol{\theta}'}(\boldsymbol{\theta})$, the determinant of first order derivatives $|\varphi_{\boldsymbol{\theta}'}(\boldsymbol{\theta})|$, as well as its inverse matrix $\varphi_{\boldsymbol{\theta}'}^{-1}(\boldsymbol{\theta})$ in our later steps.

The mean return for the risk-free asset is represented by r_f. Since our aim is to test the difference in Sharpe ratio under under correlated samples, the parameter of interest ψ and its first order derivative are then defined as follows:

$$\psi(\boldsymbol{\theta}) = \frac{\mu_X - r_f}{\sigma_X} - \frac{\mu_Y - r_f}{\sigma_Y}\,, \tag{3.1.41}$$

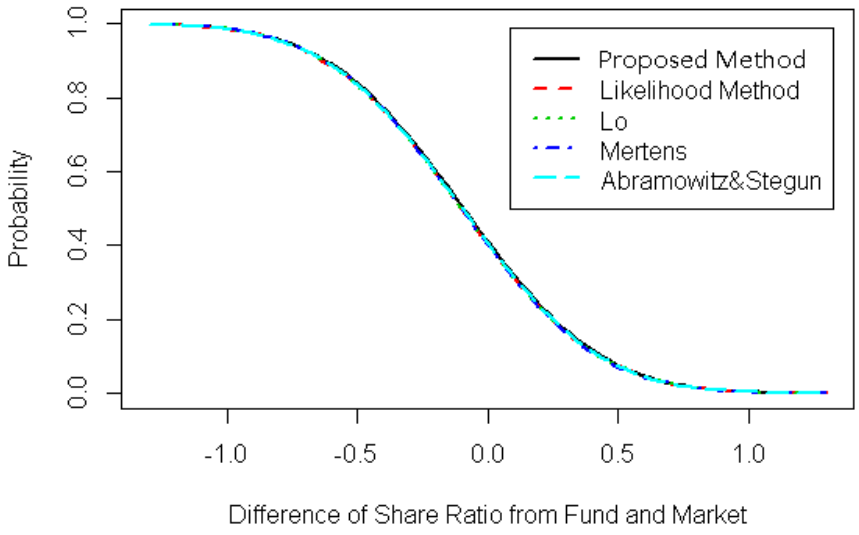
$$\psi_{\boldsymbol{\theta}'}(\boldsymbol{\theta}) = \left(\frac{1}{\sigma_X}, -\frac{1}{\sigma_Y}, -\frac{\mu_X - r_f}{2\sigma_X^3}, \frac{\mu_Y - r_f}{2\sigma_Y^3}, 0 \right)\,. \tag{3.1.42}$$

The density of $\mathbf{Z}_i = \begin{pmatrix} X_i \\ Y_i \end{pmatrix}$ is $f(\mathbf{z}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}}\, e^{-\frac{1}{2}(\mathbf{z}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{z}_i - \boldsymbol{\mu})}$.

Table 3.13: p-values for Difference of Sharpe Ratio under IID Normal Return of Fund and Market

Method	-1.1	-0.9	-0.8	0	0.5	0.6	0.8	0.9
Lo	0.9923	0.9736	0.9550	0.4041	0.0730	0.0449	0.0146	0.0077
Mertens	0.9927	0.9746	0.9563	0.4033	0.0713	0.0436	0.0139	0.0073
Likelihood Ratio	0.9923	0.9736	0.9549	0.4041	0.0730	0.0449	0.0146	0.0077
Abramowitz&Stegun	0.9923	0.9738	0.9553	0.4063	0.0740	0.0456	0.0149	0.0079
Proposed	**0.9927**	**0.9748**	**0.9568**	**0.4119**	**0.0758**	**0.0468**	**0.0153**	**0.0081**

Figure 3.5: p-value function for Difference of Sharpe Ratio under IID Normal Return of Fund and Market



Thus the joint likelihood is

$$\mathcal{L}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{z}\right) = c \cdot \prod_{i=1}^{n} f\left(\mathbf{z}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = c\left|2\pi\boldsymbol{\Sigma}\right|^{-\frac{n}{2}} e^{-\frac{1}{2}\left(\sum_{i=1}^{n}\left(\mathbf{z}_i-\boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}_i-\boldsymbol{\mu}\right)\right)}$$

The unrestricted maximum likelihood estimation maximizes the joint log likelihood function $l\left(\boldsymbol{\theta}\right)$:

$$\begin{aligned}
& l(\boldsymbol{\theta}) \qquad (3.1.43)\\
= \;& \log\left(\mathcal{L}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{z}\right)\right) = a - \frac{n}{2}\log\left|\boldsymbol{\Sigma}\right| - \frac{1}{2}\left(\sum_{i=1}^{n}\left(\mathbf{z}_i - \boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{z}_i - \boldsymbol{\mu}\right)\right)\\
= \;& a - \frac{n}{2}\log\left(\left(1-\rho^2\right)\sigma_X^2\sigma_Y^2\right) -\\
& \frac{1}{2\left(1-\rho^2\right)}\left(\sum_{i=1}^{n}\left(\mathbf{z}_i - \boldsymbol{\mu}\right)'\begin{pmatrix} \frac{1}{\sigma_X^2} & \frac{-\rho}{\sigma_X\sigma_Y} \\ \frac{-\rho}{\sigma_X\sigma_Y} & \frac{1}{\sigma_Y^2} \end{pmatrix}\left(\mathbf{z}_i - \boldsymbol{\mu}\right)\right)\\
= \;& a - n\log\sigma_X - n\log\sigma_Y - \frac{n}{2}\log\left(1-\rho^2\right) -\\
& \frac{\sum_{i=1}^{n}\left(\left(\frac{x_i-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x_i-\mu_X}{\sigma_X}\right)\left(\frac{y_i-\mu_Y}{\sigma_Y}\right) + \left(\frac{y_i-\mu_Y}{\sigma_Y}\right)^2\right)}{2\left(1-\rho^2\right)}.
\end{aligned}$$

Hence, the unconstrained maximum likelihood estimators are:

$$\begin{aligned}
\hat{\boldsymbol{\theta}}' &= \left(\hat{\mu}_X, \hat{\mu}_Y, \hat{\sigma}_X^2, \hat{\sigma}_Y^2, \hat{\rho}\right).\\
&= \left(\frac{\sum X_i}{n}, \frac{\sum Y_j}{n}, \frac{\sum(X_i - \hat{\mu}_X)^2}{n}, \frac{\sum(Y_j - \hat{\mu}_Y)^2}{n}, \frac{\sum\left(X_i - \hat{\mu}_X\right)\left(Y_j - \hat{\mu}_Y\right)}{n\hat{\sigma}_X\hat{\sigma}_Y}\right)
\end{aligned}$$

Knowing this unconstrained MLE, we can get other important variable for late use, such as the estimated parameter of interest $\hat{\psi} = \psi(\hat{\boldsymbol{\theta}}) = \frac{\hat{\mu}_X - r_f}{\hat{\sigma}_X} - \frac{\hat{\mu}_Y - r_f}{\hat{\sigma}_Y}$, the estimated unrestricted likelihood function $l(\hat{\boldsymbol{\theta}})$, the observed in-

formation matrix evaluated at $\hat{\boldsymbol{\theta}}$:

$$
\begin{aligned}
&\mathbf{j}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}) = \\
&\begin{pmatrix}
-l_{\mu_X\mu_X}(\hat{\boldsymbol{\theta}}) & -l_{\mu_X\mu_Y}(\hat{\boldsymbol{\theta}}) & -l_{\mu_X\sigma_X^2}(\hat{\boldsymbol{\theta}}) & -l_{\mu_X\sigma_Y^2}(\hat{\boldsymbol{\theta}}) & -l_{\mu_X\rho}(\hat{\boldsymbol{\theta}}) \\
-l_{\mu_Y\mu_X}(\hat{\boldsymbol{\theta}}) & -l_{\mu_Y\mu_Y}(\hat{\boldsymbol{\theta}}) & -l_{\mu_Y\sigma_X^2}(\hat{\boldsymbol{\theta}}) & -l_{\mu_Y\sigma_Y^2}(\hat{\boldsymbol{\theta}}) & -l_{\mu_Y\rho}(\hat{\boldsymbol{\theta}}) \\
-l_{\sigma_X^2\mu_X}(\hat{\boldsymbol{\theta}}) & -l_{\sigma_X^2\mu_Y}(\hat{\boldsymbol{\theta}}) & -l_{\sigma_X^2\sigma_X^2}(\hat{\boldsymbol{\theta}}) & -l_{\sigma_X^2\sigma_Y^2}(\hat{\boldsymbol{\theta}}) & -l_{\sigma_X^2\rho}(\hat{\boldsymbol{\theta}}) \\
-l_{\sigma_Y^2\mu_X}(\hat{\boldsymbol{\theta}}) & -l_{\sigma_Y^2\mu_Y}(\hat{\boldsymbol{\theta}}) & -l_{\sigma_Y^2\sigma_X^2}(\hat{\boldsymbol{\theta}}) & -l_{\sigma_Y^2\sigma_Y^2}(\hat{\boldsymbol{\theta}}) & -l_{\sigma_Y^2\rho}(\hat{\boldsymbol{\theta}}) \\
-l_{\rho\mu_X}(\hat{\boldsymbol{\theta}}) & -l_{\rho\mu_Y}(\hat{\boldsymbol{\theta}}) & -l_{\rho\sigma_X^2}(\hat{\boldsymbol{\theta}}) & -l_{\rho\sigma_Y^2}(\hat{\boldsymbol{\theta}}) & -l_{\rho\rho}(\hat{\boldsymbol{\theta}})
\end{pmatrix} \\
&= \begin{pmatrix}
\frac{n}{(1-\hat{\rho}^2)\hat{\sigma}_X^2} & \frac{-\hat{\rho}n}{(1-\hat{\rho}^2)\hat{\sigma}_X\hat{\sigma}_Y} & 0 & 0 & 0 \\
\frac{-\hat{\rho}n}{(1-\hat{\rho}^2)\hat{\sigma}_X\hat{\sigma}_Y} & \frac{n}{(1-\hat{\rho}^2)\hat{\sigma}_Y^2} & 0 & 0 & 0 \\
0 & 0 & \frac{(2-\hat{\rho}^2)n}{4\hat{\sigma}_X^4(1-\hat{\rho}^2)} & \frac{-n\hat{\rho}^2}{4\hat{\sigma}_X^2\hat{\sigma}_Y^2(1-\hat{\rho}^2)} & \frac{-n\hat{\rho}}{2\hat{\sigma}_X^2(1-\hat{\rho}^2)} \\
0 & 0 & \frac{-n\hat{\rho}^2}{4\hat{\sigma}_X^2\hat{\sigma}_Y^2(1-\hat{\rho}^2)} & \frac{(2-\hat{\rho}^2)n}{4\hat{\sigma}_Y^4(1-\hat{\rho}^2)} & \frac{-n\hat{\rho}}{2\hat{\sigma}_Y^2(1-\hat{\rho}^2)} \\
0 & 0 & \frac{-n\hat{\rho}}{2\hat{\sigma}_X^2(1-\hat{\rho}^2)} & \frac{-n\hat{\rho}}{2\hat{\sigma}_Y^2(1-\hat{\rho}^2)} & \frac{n(1+\hat{\rho}^2)}{(1-\hat{\rho}^2)^2}
\end{pmatrix}.
\end{aligned}
$$

We can see from the derivation process that, when we set $\rho = 0,$ the results would be pretty much the same as the results of independent case.

The constrained maximum likelihood estimators are solved by the Lagrange multiplier method (see (2.3.7)). The Lagrangian function is given at (3.1.44) and its first order derivatives are listed from (3.1.45) to (3.1.50).

$$H(\boldsymbol{\theta}, \alpha) = l(\boldsymbol{\theta}) + \alpha(\psi(\boldsymbol{\theta}) - \psi) = l(\boldsymbol{\theta}) + \alpha(\frac{\mu_X - r_f}{\sigma_X} - \frac{\mu_Y - r_f}{\sigma_Y} - \psi)\ , \quad (3.1.44)$$

$$H_{\mu_X}(\boldsymbol{\theta}, \alpha) = l_{\mu_X}(\boldsymbol{\theta}) + \frac{\alpha}{\sigma_X}\ , \quad (3.1.45)$$

$$H_{\mu_Y}(\boldsymbol{\theta}, \alpha) = l_{\mu_Y}(\boldsymbol{\theta}) - \frac{\alpha}{\sigma_Y}\ , \quad (3.1.46)$$

$$H_{\sigma_X^2}(\boldsymbol{\theta}, \alpha) = l_{\sigma_X^2}(\boldsymbol{\theta}) - \frac{\alpha(\mu_X - r_f)}{2\sigma_X^3}\ , \quad (3.1.47)$$

$$H_{\sigma_Y^2}(\boldsymbol{\theta}, \alpha) = l_{\sigma_Y^2}(\boldsymbol{\theta}) + \frac{\alpha(\mu_Y - r_f)}{2\sigma_Y^3}\ , \quad (3.1.48)$$

$$H_{\rho}(\boldsymbol{\theta}, \alpha) = l_{\rho}(\boldsymbol{\theta})\ , \quad (3.1.49)$$

$$H_{\alpha}(\boldsymbol{\theta}, \alpha) = \frac{\mu_X - r_f}{\sigma_X} - \frac{\mu_Y - r_f}{\sigma_Y} - \psi\ . \quad (3.1.50)$$

Setting first order derivatives, (3.1.45) to (3.1.50), equal to zero and solving the resulting system produces the constrained MLE. Thus, six unknown variables- five constrained MLE $\hat{\boldsymbol{\theta}}'_{\psi} = (\tilde{\mu}_X, \tilde{\mu}_Y, \tilde{\sigma}_X^2, \tilde{\sigma}_Y^2, \tilde{\rho})$ plus the estimated Lagrange estimator $\hat{\alpha}$ -are solved from a nonlinear system of six equa-

tions. Note that only numerical solutions can be obtained from this system with the starting values of iteration being unconstrained MLE $\hat{\boldsymbol{\theta}}' = (\hat{\mu}_X, \hat{\mu}_Y, \hat{\sigma}_X^2, \hat{\sigma}_Y^2, \hat{\rho})$. By using constrained MLE, we can thus obtain the estimated restricted likelihood function: $l(\hat{\boldsymbol{\theta}}_\psi)$, the tilted observed information matrix evaluated at $\hat{\boldsymbol{\theta}}_\psi$, $\tilde{\mathbf{j}}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_\psi) = -\frac{\partial^2 \tilde{\mathbf{j}}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_\psi} = j_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_\psi) - \hat{\alpha}\psi_{\boldsymbol{\theta}\boldsymbol{\theta}'}\left(\hat{\boldsymbol{\theta}}_\psi\right)$ and its inverse $\tilde{\mathbf{j}}^{-1}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_\psi)$.

Given the above information, R can be constructed from (2.3.15), χ can be calculated from (2.6.17), Q can be obtained from (2.6.18) and (2.6.19), and finally $R^\star$ can be obtained from (2.6.13).

3.1.8 Simulations for Two Correlated Sample Comparison on Sharpe Ratio

3.1.8.1 Reference Group of Existing Methodology

In order to illustrate the exceptional accuracy of our proposed method, we construct the followings as our reference group of methodology. These existing methodology correspond to the methods in section (3.1.2.1) and (3.1.5.1).

1. Memmel (2003): Suppose we have two samples X and Y that are correlated with correlation coefficient $\hat{\rho}$. According to Jobson and Korkie (1981) and Lo (2002), each sample will have Sharpe ratio's asymptotic distribution, $\widehat{SR}_X \xrightarrow{d} N\left(SR_X, \frac{1}{n}(1+\frac{1}{2}\widehat{SR}_X^2)\right)$ and $\widehat{SR}_Y \xrightarrow{d} N\left(SR_Y, \frac{1}{n}(1+\frac{1}{2}\widehat{SR}_Y^2)\right)$. Later Memmel (2003, eq13) obtained the distribution of the difference of estimated Sharpe ratio:

$$Cov\left(\widehat{SR}_X, \widehat{SR}_Y\right) = \frac{1}{n}\left(\hat{\rho} + \frac{\hat{\rho}^2\widehat{SR}_X\widehat{SR}_Y}{2}\right), \tag{3.1.51}$$

$$\widehat{SR}_X - \widehat{SR}_Y \xrightarrow{d} N\Bigg(SR_X - SR_Y, \frac{1}{n}\left(2 - 2\hat{\rho} + \frac{1}{2}(\widehat{SR}_X^2 + \widehat{SR}_Y^2 - 2\hat{\rho}^2\widehat{SR}_X\widehat{SR}_Y)\right)\Bigg);$$

2. Wright et al (2014): Suppose we have two samples X and Y that are correlated with correlation coefficient $\hat{\rho}$. According to Mertens (2002), each sample will have its own Sharpe ratio's asymptotic distribution, $\widehat{SR} \xrightarrow{d} N\left(SR, \hat{\sigma}^2_{SR}\right)$ with $\hat{\sigma}^2_{SR} = \frac{1}{n}\left(1 + \frac{\widehat{SR}^2}{2} - \hat{\alpha}_3\widehat{SR} + \frac{\hat{\alpha}_4 - 3}{4}\widehat{SR}^2\right)$. Later Wright et al (2014 A.11) derived the distribution of the difference of estimated Sharpe

ratio:

$$Cov\left(\widehat{SR}_X, \widehat{SR}_Y\right) =$$
$$\frac{1}{4n\sigma_X^3\sigma_Y^3}\Bigg\{4\left(\sigma_X^2+\mu_X^2\right)\left(\sigma_Y^2+\mu_Y^2\right)Cov\left(X,Y\right)+\mu_X\mu_Y Cov\left(X^2,Y^2\right)$$
$$-2\mu_Y(\sigma_X^2+\mu_X^2)Cov\left(X,Y^2\right)-2\mu_X(\sigma_Y^2+\mu_Y^2)Cov\left(X^2,Y\right)\Bigg\},$$

$$\widehat{SR}_X-\widehat{SR}_Y \xrightarrow{d} N\left(SR_X-SR_Y, \hat{\sigma}^2_{SR_X}+\hat{\sigma}^2_{SR_Y}-2\widehat{Cov}\left(\widehat{SR}_X,\widehat{SR}_Y\right)\right);$$

3. The signed log-likelihood ratio statistic in (2.3.15):

$$R(\psi)=\text{sgn}(\hat{\psi}-\psi)\sqrt{2(l(\hat{\boldsymbol{\theta}})-l(\hat{\boldsymbol{\theta}}_\psi))};$$

Results 1, 2 and 3 are all first order approximation $O(n^{-\frac{1}{2}})$, while our proposed likelihood method is third order approximation $O(n^{-\frac{3}{2}})$, indicating that theoretically our proposed likelihood method is more valid and accurate than the above members in reference group.

3.1.8.2 Another Proposed Method

Another method is being proposed by extending the results from Abramowitz and Stegun(1964 p949) as well as the results from Memmel (2003).

Lemma. *Extending the results from Abramowitz and Stegun(1964 p949) as well as the results from Memmel (2003), we can obtain the following distribution for the difference of estimated Sharpe ratio when their underlying are correlated:*

$$\frac{4n-5}{4n-4}\left(\widehat{SR}_X-\widehat{SR}_Y\right)\xrightarrow{d}$$
$$N\left(SR_X-SR_Y, \frac{2}{n}+\frac{\widehat{SR}_X^2+\widehat{SR}_Y^2}{2\left(n-1\right)}-\frac{2\left(4n-5\right)^2}{n\left(4n-4\right)^2}\left(\hat{\rho}+\frac{\hat{\rho}^2\widehat{SR}_X\widehat{SR}_Y}{2}\right)\right)$$

Proof. Given two normal random variables X and Y: $X\rightarrow N\left(\mu_X,\sigma_X^2\right)$ and $Y\rightarrow N\left(\mu_X,\sigma_Y^2\right)$. If they are correlated such that $\rho=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}$, then

$$X-Y\rightarrow N\left(\mu_X-\mu_Y,\sigma_X^2+\sigma_Y^2-2\rho\sigma_X\sigma_Y\right). \tag{3.1.52}$$

First, from the results of Abramowitz and Stegun(1964 p949), we have the individual distribution of estimated Sharpe ratio $\widehat{SR}_X\left(1-\frac{1}{4(n-1)}\right)\xrightarrow{d}$

$N\left(SR_X, \frac{1}{n} + \frac{\widehat{SR}_X^2}{2(n-1)}\right)$ and $\widehat{SR}_Y\left(1 - \frac{1}{4(n-1)}\right) \xrightarrow{d} N\left(SR_Y, \frac{1}{n} + \frac{\widehat{SR}_Y^2}{2(n-1)}\right)$. Thus by (3.1.52), we have

$$\frac{4n-5}{4n-4}\left(\widehat{SR}_X - \widehat{SR}_Y\right) \xrightarrow{d}$$
$$N\left(SR_X - SR_Y, \frac{2}{n} + \frac{\widehat{SR}_X^2 + \widehat{SR}_Y^2}{2(n-1)} - 2\left(\frac{4n-5}{4n-4}\right)^2 Cov\left(\widehat{SR}_X, \widehat{SR}_Y\right)\right).$$

Then, we can take Memmel's expression of $Cov\left(\widehat{SR}_X, \widehat{SR}_Y\right)$ at (3.1.51) into the above expression to obtain another proposed method. □

3.1.8.3 Numerical Study

In this part we provide a simulation study to assess the performance of our third order likelihood method relative to the existing methodology in reference group. For some combinations of $n = 6, 12$, $\mu = 0, 0.5, 1$, $\sigma^2 = 1$, $\rho = -0.5, 0.5$ and $r_f = 0$, ten thousand Monte Carlo replications are performed from bivariate normal distribution with parameters $\boldsymbol{\theta}' = (\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$. And for each generated sample, the 95% confidence interval for the difference of Sharpe ratio is calculated. The performance of a method is evaluated by the same criteria 1-6 in (3.1.2.2).

The simulated coverage probabilities, coverage error, upper and lower error probabilities and average biases and degree of symmetry are recorded in Table (3.14). We can conclude from the simulation that proposed modified signed log likelihood ratio method and proposed modified Abramowitz and Stegun's method give excellent results and outperform the other three methods based on average bias and central coverage.

- The performance of the methodology in the reference group are not satisfactory. In particular, Memmel's method and the likelihood ratio statistic have very similar simulation result. Method of Wright et al does not always function under the whole sample space, and for some samples of simulation (especially $SR_X - SR_Y$ is small and $\rho = 0.5$) this method may even output negative estimated variance. However, our proposed methods generally performed extremely well in the criteria considered in this section.

- Sample Size Effect on Average Bias: The performance of each methods are supposed to improve as sample size rises. In order to make this size effect visible, a second round of simulation is being conducted, setting $\mu_X = 1$, $\sigma_X = 1$, $\mu_Y = 1$, $\sigma_Y = 1$, $\rho = -0.5$ and $\mu_X = 1$, $\sigma_X = 1$, $\mu_Y = 1$, $\sigma_Y = 1$, $\rho = 0.5$, respectively. Each sample size rises from $n = 4$ to $n = 100$ and the results are recorded in Figure (3.6). Suppose we take 3 units of standard deviation as an acceptance level

Table 3.14: Simulation Result for Difference of Sharpe Ratio under Bivariate Normal Return

Setting	Method	CP	LE	UE	AB	AB/ER	SY
$n = 12$, $SR_X = 1$, $SR_Y = 1$, $\rho = 0.5$	Memmel	0.9209	0.0396	0.0395	0.0146	9.09	1.00
	Wright et al	0.9863	0.0066	0.0071	0.0182	11.34	1.08
	Likelihood Ratio	0.9193	0.0407	0.0400	0.0154	9.59	1.02
	Distribution Proposed	**0.9417**	**0.0296**	**0.0287**	**0.0042**	**2.59**	**1.03**
	Likelihood Proposed	**0.9429**	**0.0290**	**0.0281**	**0.0036**	**2.22**	**1.03**
$n = 12$, $SR_X = 1$, $SR_Y = 1$, $\rho = -0.5$	Memmel	0.9309	0.0322	0.0369	0.0096	5.97	1.15
	Wright et al	0.9753	0.0115	0.0132	0.0127	7.91	1.15
	Likelihood Ratio	0.9305	0.0325	0.0370	0.0098	6.09	1.14
	Distribution Proposed	**0.9408**	**0.0275**	**0.0317**	**0.0046**	**2.88**	**1.15**
	Likelihood Proposed	**0.9547**	**0.0211**	**0.0242**	**0.0024**	**1.47**	**1.15**
$n = 6$, $SR_X = 1$, $SR_Y = 1$, $\rho = 0.5$	Memmel	0.8869	0.0561	0.0570	0.0316	19.72	1.02
	Wright et al	0.9729	0.0143	0.0128	0.0114	7.16	1.11
	Likelihood Ratio	0.8781	0.0616	0.0603	0.0360	22.47	1.02
	Distribution Proposed	**0.9541**	**0.0212**	**0.0247**	**0.0021**	**1.28**	**1.17**
	Likelihood Proposed	**0.9442**	**0.0274**	**0.0284**	**0.0029**	**1.81**	**1.04**
$n = 6$, $SR_X = 1$, $SR_Y = 1$, $\rho = -0.5$	Memmel	0.9037	0.0442	0.0521	0.0232	14.47	1.18
	Wright et al	0.9695	0.0146	0.0159	0.0098	6.09	1.09
	Likelihood Ratio	0.8963	0.0486	0.0551	0.0269	16.78	1.13
	Distribution Proposed	**0.9401**	**0.0284**	**0.0315**	**0.0050**	**3.09**	**1.11**
	Likelihood Proposed	**0.9555**	**0.0209**	**0.0236**	**0.0028**	**1.72**	**1.13**
$n = 12$, $SR_X = 1$, $SR_Y = 0.5$, $\rho = 0.5$	Memmel	0.9199	0.0401	0.0400	0.0151	9.41	1.00
	Wright et al	0.9751	0.0088	0.0161	0.0126	7.84	1.83
	Likelihood Ratio	0.9205	0.0475	0.0320	0.0148	9.22	1.48
	Distribution Proposed	**0.9406**	**0.0277**	**0.0317**	**0.0047**	**2.94**	**1.14**
	Likelihood Proposed	**0.9424**	**0.0234**	**0.0342**	**0.0054**	**3.38**	**1.46**
$n = 12$, $SR_X = 1$, $SR_Y = 0.5$, $\rho = -0.5$	Memmel	0.9281	0.0411	0.0308	0.0110	6.84	1.33
	Wright et al	0.9643	0.0180	0.0177	0.0072	4.47	1.02
	Likelihood Ratio	0.9301	0.0403	0.0296	0.0100	6.22	1.36
	Distribution Proposed	**0.9368**	**0.0345**	**0.0287**	**0.0066**	**4.13**	**1.20**
	Likelihood Proposed	**0.9530**	**0.0221**	**0.0249**	**0.0015**	**0.94**	**1.13**
$n = 6$, $SR_X = 1$, $SR_Y = 0.5$, $\rho = 0.5$	Memmel	0.8885	0.0548	0.0567	0.0308	19.22	1.03
	Wright et al	0.9540	0.0134	0.0325	0.0095	5.96	2.42
	Likelihood Ratio	0.8804	0.0737	0.0459	0.0348	21.75	1.61
	Distribution Proposed	**0.9469**	**0.0217**	**0.0314**	**0.0049**	**3.03**	**1.45**
	Likelihood Proposed	**0.9417**	**0.0196**	**0.0387**	**0.0096**	**5.97**	**1.97**
$n = 6$, $SR_X = 1$, $SR_Y = 0.5$, $\rho = -0.5$	Memmel	0.8972	0.0580	0.0448	0.0264	16.50	1.29
	Wright et al	0.9511	0.0236	0.0253	0.0009	0.53	1.07
	Likelihood Ratio	0.8962	0.0604	0.0434	0.0269	16.81	1.39
	Distribution Proposed	**0.9280**	**0.0376**	**0.0344**	**0.0110**	**6.88**	**1.09**
	Likelihood Proposed	**0.9537**	**0.0207**	**0.0256**	**0.0025**	**1.53**	**1.24**
$n = 12$, $SR_X = 1$, $SR_Y = 0$, $\rho = 0.5$	Memmel	0.9262	0.0420	0.0318	0.0119	7.44	1.32
	Wright et al	0.9675	0.0109	0.0216	0.0088	5.47	1.98
	Likelihood Ratio	0.9239	0.0556	0.0205	0.0176	10.97	2.71
	Distribution Proposed	**0.9428**	**0.0294**	**0.0278**	**0.0036**	**2.25**	**1.06**
	Likelihood Proposed	**0.9427**	**0.0162**	**0.0411**	**0.0125**	**7.78**	**2.54**
$n = 12$, $SR_X = 1$, $SR_Y = 0$, $\rho = -0.5$	Memmel	0.9250	0.0511	0.0239	0.0136	8.50	2.14
	Wright et al	0.9555	0.0249	0.0196	0.0028	1.72	1.27
	Likelihood Ratio	0.9275	0.0491	0.0234	0.0129	8.03	2.10
	Distribution Proposed	**0.9340**	**0.0421**	**0.0239**	**0.0091**	**5.69**	**1.76**
	Likelihood Proposed	**0.9522**	**0.0206**	**0.0272**	**0.0033**	**2.06**	**1.32**

on AB, we can find that our proposed likelihood method result can achieve this level even for extreme sample size $n = 5$, while the reference group may need a sample size of 50 to achieve the same accuracy. In addition, it is interesting to see Wright et al method show no sign of decaying trend as sample size rises. This may be caused by the fact that, theoretically given normally distributed returns, the skewness, historical kurtosis, and other joint moments of bivariate returns are known, and Wright et al' form reduces to Memmel's method by substitution of those known variables, however, those statistics are unknown in practice without distributional assumption and have to be estimated from the data sample, which results in mis-estimation.

- The Effect of ρ and ψ: We conduct another round of simulation to reveal the effect of ρ and ψ on the accuracy of our methods, and the results are recorded in Figure(3.7) and Figure (3.8). We can conclude that:

1. The proposed likelihood method performs better when $\psi = SR_X - SR_Y$ is around zero while the proposed modified Abramowitz and Stegun's method is doing better when ψ is away from the zero. Very interesting, they compensate with each other and can do well at the whole domain. In particular, Figure (3.7) shows, when $\rho = 0.5$, proposed likelihood method is doing extremely well at $\psi \in (-0.5, 0.5)$ while the proposed distribution method is doing well on the rest part. On the other hand, when $\rho = -0.5$, proposed likelihood method performs better even at $\psi \in (-2, 2)$ and proposed modified Abramowitz and Stegun's method is doing good on the rest domain.

2. Figure (3.7) shows the results of proposed likelihood method are relatively worse when $\rho = 0.5$ than $\rho = -0.5$, and this can be illustrated clearer at Figure (3.8). Figure (3.8) recorded the results of simulation where 10,000 random samples are generated, each having a sample size of 12 from a bivariate normal distribution with parameters $(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2)$ being $(1, 1, 1, 1)$ and $(0.5, 0, 1, 1)$ respectively. ρ is taking a list of values from -0.95 to +0.95. We can conclude that, when $\psi = 0$, our proposed likelihood method outperforms the other method at the whole range, even at the extreme value of ± 0.95. On the other hand, when $\psi = 0.5$, the proposed likelihood method performs better when ρ is less than 0.4. At the same time, our proposed modified Abramowitz and Stegun's method shows its extremely accuracy at $\rho \in (0.5, 0.9)$. Thus, we can always find one of the method at their advantage area to give accurate results.

3. Given the fact that most Sharpe ratio are between -1 to 1 and people are more willing to compare when Sharpe ratio are close, our proposed likelihood method could be more applicable in real case.

Figure 3.6: The Effect of Sample Size on AB/ER (Up: $\rho = -0.5$ Down: $\rho = 0.5$)

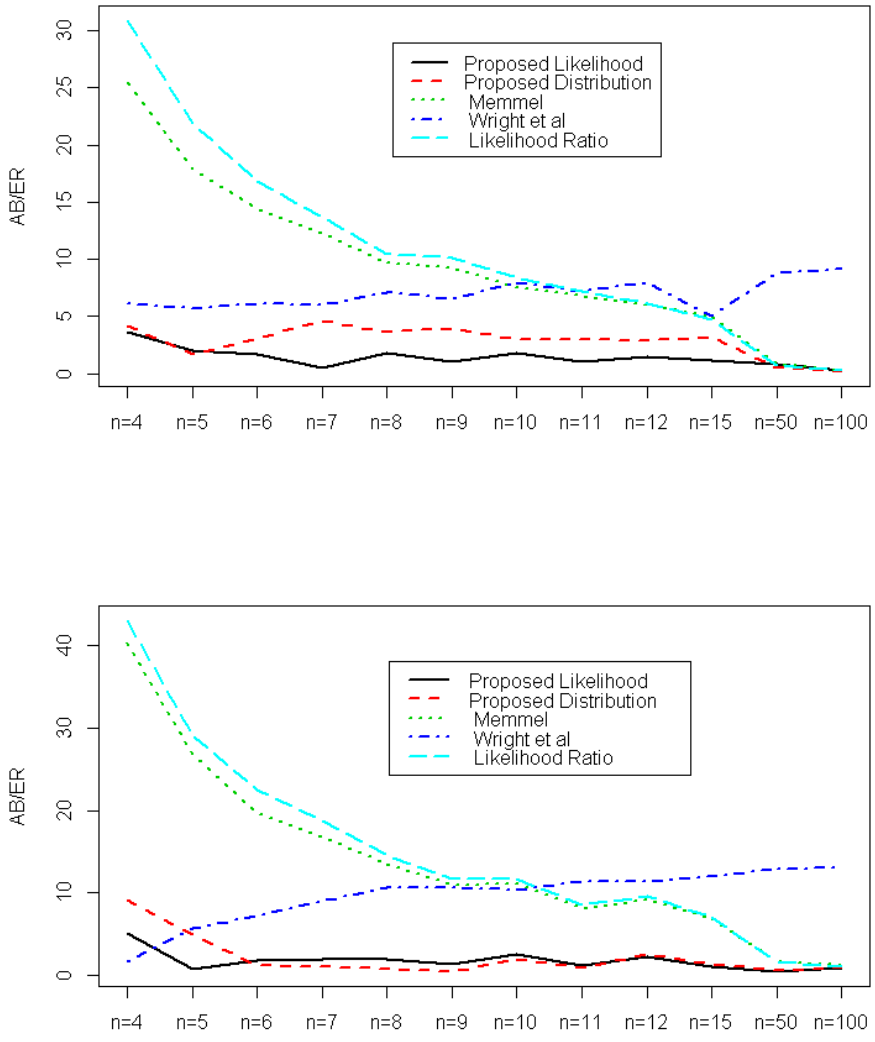


Figure 3.7: The Central Effect on AB/ER (Up: $\rho = -0.5$ Down: $\rho = 0.5$)

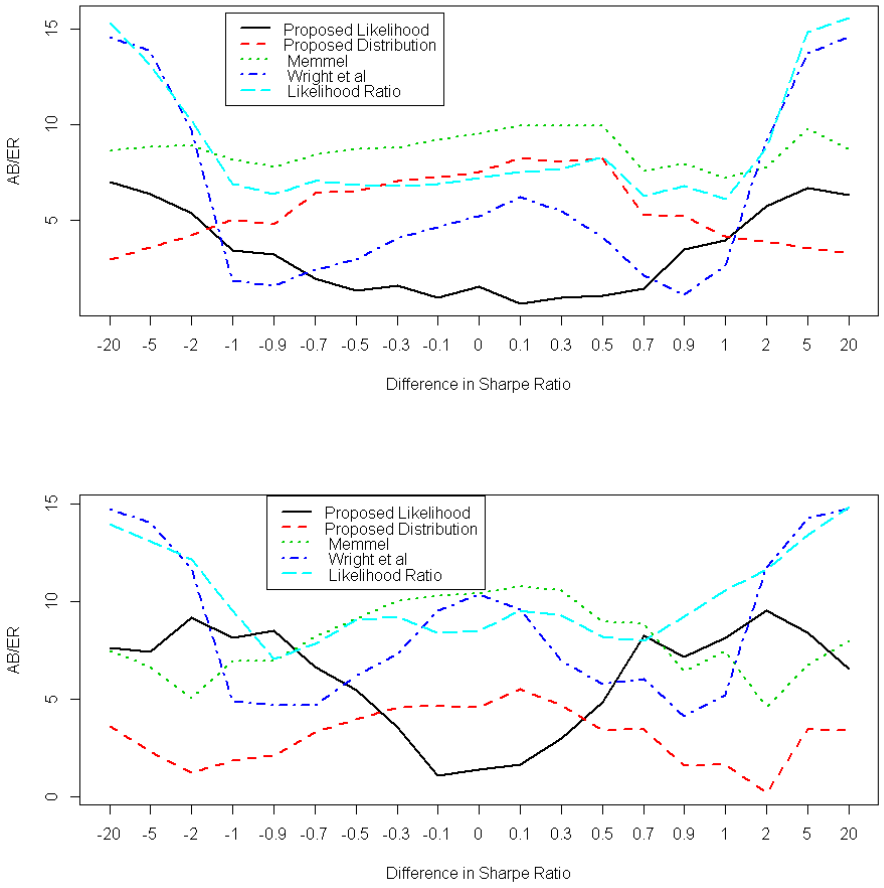
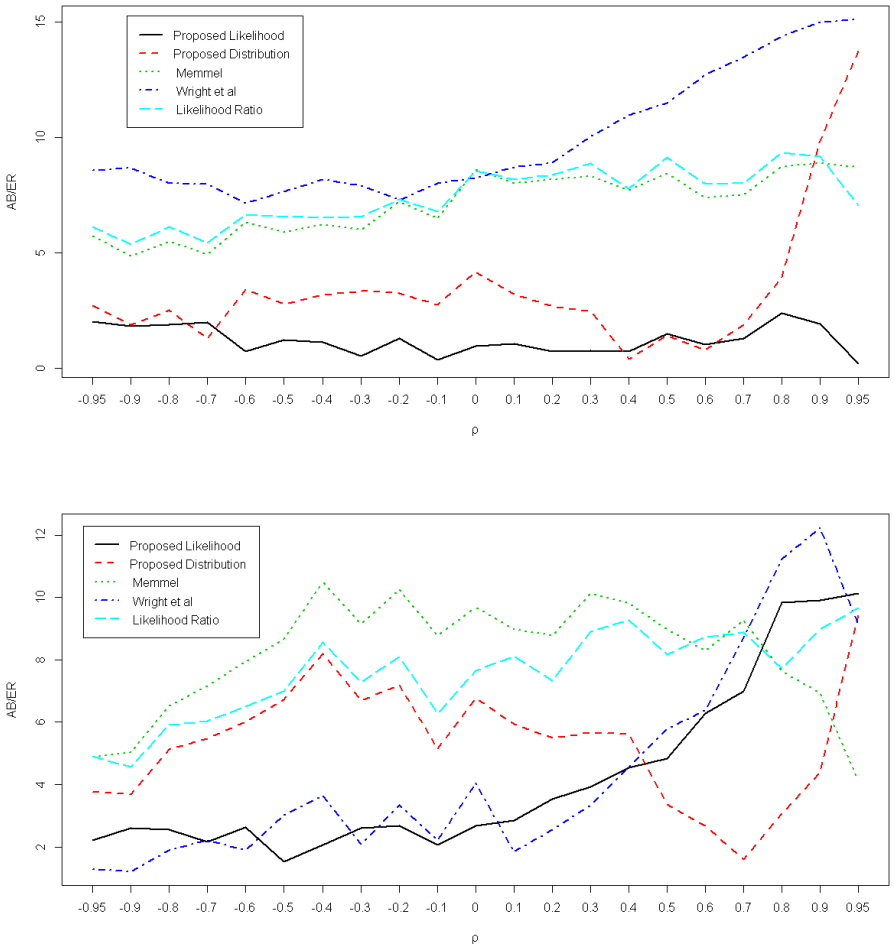


Figure 3.8: The Effect of ρ on AB/ER (Up: $\psi = 0$ Down: $\psi = 0.5$)



3.1.9 Examples for Two Correlated Sample Comparison on Sharpe Ratio

In this subsection we provide an empirical example and a simulation study for inference on the difference of Sharpe ratio from a correlated bivariate normal return. The data set used here is the same as Table (3.3) and we may, for instance, be interested in testing whether the risk-adjusted return as captured by the Sharpe ratio of mutual fund is significantly better than that of average stock market.

In subsection (3.1.3), we proved the normality of Fund return and Market return; and in subsection (3.1.6), we calculated their correlation coefficient and test the null $H_0 : \rho = 0$. The results shows that the estimated correlation coefficient is 0.6870 and the p-value of test is 0.01358. If we take significance at 95%, then we can conclude to reject the null hypothesis, indicating correlation exist between Fund return and Market return.

Although a general simulation has been performed in last subsection, here we do another simulation under the setting of our example in order to validate our statistical inference. The parameter values were chosen to mimic the example data with Fund: $N(0.0009050833, 0.00001932047)$ and Market: $N(0.002995667, 0.0001244846)$, $\rho = 0.6870$, and $r_f = 0.000242$. 90%, 95%, and 99% confidence intervals for the difference of Sharpe ratio were obtained for each sample. Table (3.15) report the results from these simulations. From the result table it is clear that, our proposed methods outperform the other methods based on the criteria we examined.

Our first application focuses on 95% confidence intervals for the difference of Sharpe ratio from the correlated samples of Fund and Market, and the results are recorded at Table (3.16).We can find that the confidence intervals obtained from the five methods produce different results. In particular, the simulations indicate that our proposed methods should give more accurate confidence interval compared with the reference methodology.

The p-value functions calculated from each methods are plotted in Figures (3.9). These significance functions can be used to obtain p-values for specific hypothesized values of the difference of Sharpe ratio, which are shown in Table (3.17). One of the most important hypothesized values is zero. For example, we want to test:

$$H_0 : SR_{Fund} - SR_{Mkt} \geq 0 \text{ versus } H_1 : SR_{Fund} - SR_{Mkt} < 0$$

The corresponding p-values for our proposed likelihood method is 0.3456 and for proposed modified Abramowitz and Stegun's method is 0.3446. They are quite close and we can not reject the null at 95% significance level, and actually all the other methods reach the same conclusion. However, in general we can see from Table (3.17) that the p-values vary across the methods and people can result in different conclusions by using difference methods.

Table 3.15: Simulation Result for Difference of Sharpe Ratio under Bivariate Correlated Normal Return of Fund and Market

CI	Method	CP	LE	UE	AB	AB/SE	SY
95%	Memmel	0.9219	0.0390	0.0391	0.0141	8.78	1.00
	Wright et al	0.9171	0.0436	0.0393	0.0164	10.27	1.11
	Likelihood Ratio	0.9230	0.0361	0.0409	0.0135	8.44	1.13
	Distribution Proposed	**0.9480**	**0.0263**	**0.0257**	**0.0010**	**0.62**	**1.02**
	Likelihood Proposed	**0.9460**	**0.0280**	**0.0260**	**0.0020**	**1.25**	**1.08**
90%	Memmel	0.8583	0.0709	0.0708	0.0209	9.48	1.00
	Wright et al	0.8599	0.0708	0.0693	0.0201	9.11	1.02
	Likelihood Ratio	0.8615	0.0671	0.0714	0.0193	8.75	1.06
	Distribution Proposed	**0.8955**	**0.0519**	**0.0526**	**0.0023**	**1.02**	**1.01**
	Likelihood Proposed	**0.8921**	**0.0566**	**0.0513**	**0.0040**	**1.80**	**1.10**
99%	Memmel	0.9784	0.0114	0.0102	0.0058	8.29	1.12
	Wright et al	0.9728	0.0156	0.0117	0.0086	12.30	1.34
	Likelihood Ratio	0.9798	0.0093	0.0109	0.0051	7.29	1.17
	Distribution Proposed	**0.9894**	**0.0050**	**0.0056**	**0.0003**	**0.43**	**1.12**
	Likelihood Proposed	**0.9884**	**0.0053**	**0.0063**	**0.0008**	**1.14**	**1.19**

Table 3.16: 95% Confidence Intervals for Difference of Sharpe Ratio under Bivariate Correlated Normal Return of Fund and Market

Method	CI for Difference in SR	CI Length
Memmel	(-0.5572 0.3568)	0.9141
Wright et al	(-0.5729 0.3725)	0.9454
Likelihood Ratio	(-0.6085 0.3964)	1.0049
Distribution Proposed	**(-0.5778 0.3819)**	**0.9597**
Likelihood Proposed	**(-0.5535 0.3737)**	**0.9272**

Table 3.17: p-values for Difference of Sharpe Ratio under Bivariate Correlated Normal Return of Fund and Market

Method	-0.7	-0.6	-0.5	0	0.2	0.3	0.4	0.5
Memmel	0.9949	0.9840	0.9568	0.3337	0.0990	0.0431	0.0160	0.0050
Wright et al	0.9936	0.9809	0.9513	0.3389	0.1066	0.0485	0.0190	0.0064
Likelihood Ratio	0.9879	0.9733	0.9439	0.3338	0.1061	0.0524	0.0243	0.0108
Distribution Proposed	0.9930	0.9799	0.9497	0.3446	0.1118	0.0520	0.0210	0.0073
Likelihood Proposed	0.9929	0.9830	0.9615	0.3456	0.0985	0.0457	0.0200	0.0084

3.1.10 Sensitivity Test for Proposed Likelihood Methodology under IID Normal Return

In order to know how sensitive our proposed method is to the independent assumption, we conduct the following sensitivity tests.

1. We generate data with AR(1) structure with Gaussian errors and then analyze the data set as they are independent data. In particular, we set the autocorrelated series with zero mean and unit error variance and the autocorrelation coefficients are 0, ± 0.1, ± 0.5, ± 0.9 respectively. In this way, we want to see if the proposed method still works for autocorrelated data despite of being designed for IID data. The results are shown in Figure (3.10) . From the figure, we can see that our proposed method work well for phi=± 0.1. It does not work for large absolute value of coefficients and the results show a huge dependency, though our proposed third order method is still better than first order methods. So, it makes sense to develop a third order likelihood method specifically with AR(1) structure and Chapter 4 will resolve the issue.

2. In order to check the empirical coverage of our proposed method, we generate simulations from independent central t distribution with degree of freedom being $1, 2, 3, 4, 6$ and use the independent normality assumption to analyze it. The results are shown in Figure (3.11).

3. Repeat above about generating from some skewed distribution such as gamma distribution as if we didn't know that simulated data is not IID. We set rate parameter to be 1 and the shape parameter to be 100, 50, 10, 5 and 1. The results are shown in Figure (3.12).

Before we precede to next section, it is worth mentioning that very often we need to deal with general IID lognormal return in real market operation, and people can just take a log transformation on that lognormal return, and then transform it to be a normal structure, and then apply our proposed third-order likelihood method introduced in this section to make reference of Sharpe ratio.

3.2 Inference for J.S. Sharpe Ratio under IID Lognormal Gross Return

In this section, we will look into a special type of Sharpe ratio-J.S. Sharpe ratio-under IID lognormal return. At the end, we will show that our proposed third order likelihood method gives good coverage regardless of which Sharpe ratio we are using.

Figure 3.9: p-value function for Difference of Sharpe Ratio under Bivariate Correlated Normal Return of Fund and Market

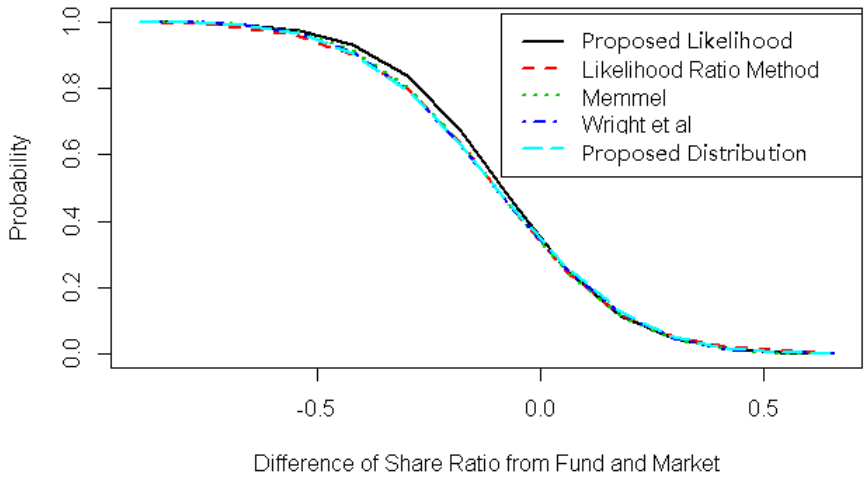


Figure 3.10: Sensitivity Test: AR(1) Structure to IID Normal

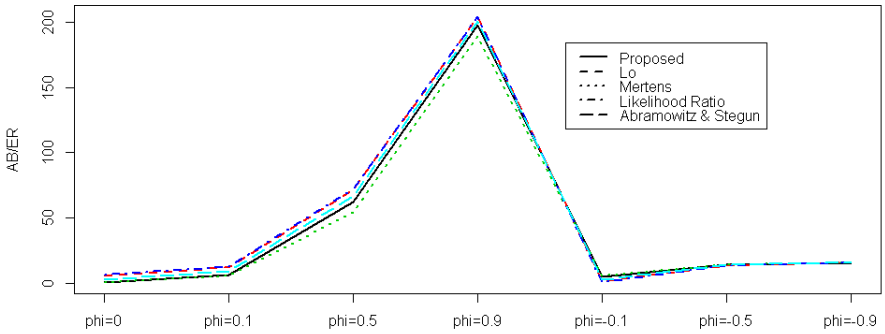


Figure 3.11: Sensitivity Test: Student t distribution Structure to IID Normal

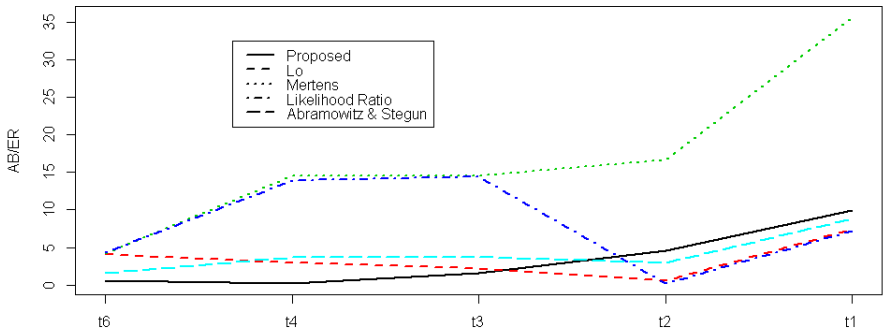
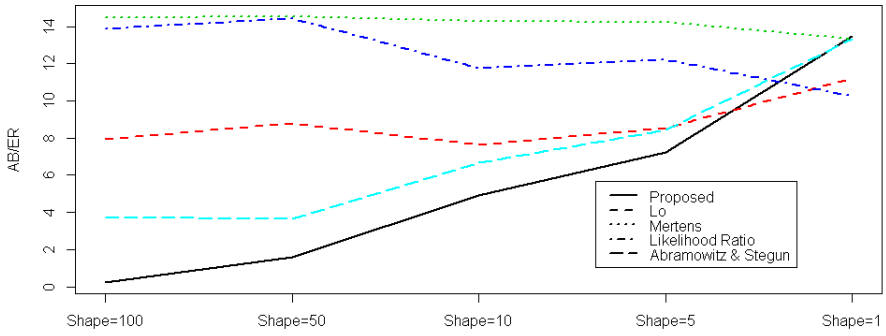


Figure 3.12: Sensitivity Test: Gamma distribution Structure to IID Normal



3.2.1 J.S. Sharpe Ratio

We first discuss about risk-free asset. There are two basic conditions that have to be met for risk-free asset. The first is that there can be no default risk. Essentially, this rules out any security issued by a private firm, since even the largest and safest firms have some measure of default risk. The only securities that have a chance of being risk free are government securities, because they control the printing of currency. Even this assumption, straightforward though it might seem, there are many emerging market economies where this assumption might not be viewed as reasonable. Governments in these markets are perceived as capable of defaulting even on local borrowing, or they refuse to honor claims made by previous regimes. There is a second condition that riskless securities need to fulfill, being often forgotten, is there can be no reinvestment risk. To illustrate this point, assume that you are trying to estimate the expected return over a five-year period, and that you want a risk free rate. A six-month Treasury bill rate, while default free, will not be risk free, because there is the reinvestment risk of not knowing what the Treasury bill rate will be in six months. Even a 5-year treasury bond is not risk free, since the coupons on the bond will be reinvested at rates that cannot be predicted today. The risk free rate for a five-year time horizon has to be the expected return on a default-free five-year zero coupon bond. This clearly has painful implications for anyone doing corporate finance or valuation, where expected returns often have to be estimated for periods ranging from one to ten years. So it does exist the condition that risk-free asset will not be available and hence the conventional type of Sharpe ratio can be modified thereafter.

J.Knight & S.Satchell(2005) advocate a new version of Sharpe ratio, referred to as J.S. Sharpe ratio , denoted by SR_l . That is, under the special condition without risk-free asset, this special form of Sharpe ratio is defined as:

$$SR_l = \frac{E(\frac{P_t}{P_{t-1}})}{Sd(\frac{P_t}{P_{t-1}})} = \frac{E(R_t)+1}{Sd(R_t)} \ . \tag{3.2.1}$$

On the other hand, under lognormal assumption, $g_t = (1+R_t) \sim LN(\mu, \sigma^2)$, we can obtain the expression of expectation and variance of net return as $E(R_t) = e^{\mu+\frac{\sigma^2}{2}} - 1$ and $var(R_t) = e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$. By using these expressions, this type Sharpe ratio can be simplified to an expression solely depending on σ,

$$SR_l = \frac{E(\frac{P_t}{P_{t-1}})}{Sd(\frac{P_t}{P_{t-1}})} = \frac{E(R_t)+1}{Sd(R_t)} = \frac{e^{\mu+\frac{\sigma^2}{2}}}{\sqrt{e^{2\mu+\sigma^2}(e^{\sigma^2}-1)}} = \frac{1}{\sqrt{e^{\sigma^2}-1}} \ . \tag{3.2.2}$$

To compare the J.S. Sharpe ratio with the conventional type Sharpe ratio, J.Knight & S.Satchell (2005) explain that SR and SR_l may rank port-

folios differently but these differences arise from different utility functions. For examples, SR may require more simplifications of quadratic utility to be applicable. In addition, they discuss the uniformly minimum variance unbiased estimator for both types of Sharpe ratio under lognormal distribution.

3.2.2 Likelihood Methodology for One Sample J.S. Sharpe Ratio

In this subsection, we implement the likelihood method introduced in Chapter 2 to get our proposed method of making reference on J.S Sharpe ratio under the assumption of lognormal gross return. Consider a fund with gross return , $g_t = (1 + R_t) \sim LN(\mu, \sigma^2)$, $t = 1, 2, \ldots, T$. then, define the parameter vector $\boldsymbol{\theta}' = (\mu, \sigma^2)$. For lognormal distribution, the canonical parameter is the same as normal case (3.1.1).

Our parameter of interest is J.S. Sharpe ratio in the expression as

$$\psi(\boldsymbol{\theta}) = SR = \frac{1}{\sqrt{e^{\sigma^2} - 1}} , \tag{3.2.3}$$

and its first order derivative is

$$\psi_{\boldsymbol{\theta}'}(\boldsymbol{\theta}) = \left(0, -\frac{1}{2}(e^{\sigma^2} - 1)^{-\frac{3}{2}} e^{\sigma^2}\right) . \tag{3.2.4}$$

The log likelihood function for lognormal gross return is

$$l(\boldsymbol{\theta}) = l(\mu, \sigma^2) = \log\left(\prod f(g_t; \boldsymbol{\theta})\right) = -\frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum(\log(g_t) - \mu)^2 . \tag{3.2.5}$$

We observe that the only difference between normal distribution and lognormal distribution is replacing r_t for $\log(g_t)$. Thus the unconstrained maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is:

$$\hat{\boldsymbol{\theta}}' = (\hat{\mu}, \hat{\sigma}^2) = (\frac{\sum \log(g_t)}{T}, \frac{\sum(\log(g_t) - \hat{\mu})^2}{T}) . \tag{3.2.6}$$

As for the constrained maximum likelihood estimation, the derivation are much easier than the case of normal, since the J.S. Sharpe ratio only involves one parameter σ. The Lagrangian function and its first order derivatives, and the second order derivatives of tilted log-likelihood function are listed below:

$$l(\boldsymbol{\theta}, \alpha) = l(\boldsymbol{\theta}) + \alpha(\psi(\boldsymbol{\theta}) - \psi) = l(\boldsymbol{\theta}) + \alpha(\frac{1}{\sqrt{e^{\sigma^2} - 1}} - \psi) , \tag{3.2.7}$$

$$l_{\mu}(\boldsymbol{\theta},\alpha) = l_{\mu}(\boldsymbol{\theta}) \ , \tag{3.2.8}$$

$$l_{\sigma^2}(\boldsymbol{\theta},\alpha) = l_{\sigma^2}(\boldsymbol{\theta}) - \frac{\alpha}{2} e^{\sigma^2}(e^{\sigma^2}-1)^{-\frac{3}{2}} \ , \tag{3.2.9}$$

$$l_{\alpha}(\boldsymbol{\theta},\alpha) = \frac{1}{\sqrt{e^{\sigma^2}-1}} - \psi \ , \tag{3.2.10}$$

$$\tilde{l}_{\mu\mu}(\boldsymbol{\theta}) = l_{\mu\mu}(\boldsymbol{\theta}) \ , \tag{3.2.11}$$

$$\tilde{l}_{\mu\sigma^2}(\boldsymbol{\theta}) = l_{\sigma^2\mu}(\boldsymbol{\theta}) \ , \tag{3.2.12}$$

$$\tilde{l}_{\sigma^2\sigma^2}(\boldsymbol{\theta}) = l_{\sigma^2\sigma^2}(\boldsymbol{\theta}) - \frac{\hat{\alpha}}{2}((e^{\sigma^2}-1)^{-\frac{3}{2}} e^{\sigma^2} - \frac{3}{2}(e^{\sigma^2}-1)^{-\frac{5}{2}} e^{2\sigma^2}) \ . \tag{3.2.13}$$

Solving first order condition (3.2.8) to (3.2.10) to zero, we will have the restricted maximum likelihood estimator.

$$\hat{\boldsymbol{\theta}}'_{\psi} = (\tilde{\mu},\tilde{\sigma}^2) = \left(\frac{\sum \log(g_t)}{T}, \log(1+\frac{1}{\psi^2}) \right) \ , \tag{3.2.14}$$

$$\hat{\alpha} = \frac{2 l_{\sigma^2}(\hat{\boldsymbol{\theta}}_{\psi})}{\psi + \psi^3} \ . \tag{3.2.15}$$

Thus the restricted likelihood function is $l(\hat{\boldsymbol{\theta}}_{\psi})$, tilted observed information matrix is $\tilde{j}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_{\psi}) = \begin{pmatrix} \frac{T}{\tilde{\sigma}^2} & 0 \\ 0 & -\tilde{l}_{\sigma^2\sigma^2}(\hat{\boldsymbol{\theta}}_{\psi}) \end{pmatrix}$ and the inverse matrix of observed information would be $\tilde{j}^{-1}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_{\psi}) = \left|\tilde{j}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_{\psi})\right|^{-1} \begin{pmatrix} -\tilde{l}_{\sigma^2\sigma^2}(\hat{\boldsymbol{\theta}}_{\psi}) & 0 \\ 0 & \frac{T}{\tilde{\sigma}^2} \end{pmatrix}$.[6]

Combining constrained MLE and unconstrained MLE, we can get : $R(\psi) = sgn(\hat{\psi}-\psi)\sqrt{-T\log(\hat{\sigma}^2) - T + T\log(\tilde{\sigma}^2) + \frac{\sum(\log(g_t)-\tilde{\mu})^2}{\tilde{\sigma}^2}}$ and the newly calibrated parameter $\chi(\boldsymbol{\theta}) = \frac{\tilde{\sigma}^4}{2\sigma^2} e^{\tilde{\sigma}^2}(e^{\tilde{\sigma}^2}-1)^{-\frac{3}{2}}$, and finally $v\hat{a}r(\chi(\hat{\boldsymbol{\theta}}) - \chi(\hat{\boldsymbol{\theta}}_{\psi})) = \frac{\frac{T}{4} e^{2\tilde{\sigma}^2}(e^{\tilde{\sigma}^2}-1)^{-3}\tilde{\sigma}^{10}}{\frac{T^2\tilde{\sigma}^6}{2}}$. Now we can obtain the Barndorff-Nielsen approximation with all given results. One merit of this application is that our proposed method here could finally obtain an analytical expression result rather than numerical as others.

[6] Note that $-\tilde{l}_{\sigma^2\sigma^2}(\hat{\theta}_{\psi})$ will be canceled later and it will not affect the final results.

3.2.3 Simulations and Examples for One Sample Sharpe Ratio

3.2.3.1 Reference Group of Existing Methodology

Although no literature can be referred on the distribution of J.Knight & S.Satchell's type Sharpe ratio, we can simulate Jobson and Korkie's process to derive the first order asymptotical distribution of SR_l. Actually the result is the same as the Wald Statistic in (2.3.13).

Lemma. *Given the definition of J.S. Sharpe ratio and lognormal assumption of the return, then we have the following first order asymptotic results:*

$$\sqrt{T}(\widehat{SR_l} - SR_l) \sim N(0, \frac{1}{2}\hat{\sigma}^4 e^{2\hat{\sigma}^2}(e^{\hat{\sigma}^2} - 1)^{-3}) \,. \tag{3.2.16}$$

Proof. Gross return follows lognormal distribution, $g_t = (1+R_t) \sim LN(\mu, \sigma^2)$, $t = 1, 2, \ldots, T$, so that the expression of maximum likelihood estimator will be

$$\hat{\theta}' = (\hat{\mu}, \hat{\sigma}^2) = \left(\frac{\sum \log(g_t)}{T}, \frac{\sum (\log(g_t) - \hat{\mu}_l)^2}{T} \right) \,.$$

Hence, the asymptotic normality for MLE is

$$\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \sim N(0, \boldsymbol{V}) \,,$$

where

$$\boldsymbol{V} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix} \,.$$

In this case, $SR_l = \frac{1}{\sqrt{e^{\sigma^2}-1}} = f(\theta)$ and thus the Jacobian term would be:

$$\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial f(\boldsymbol{\theta})}{\partial \mu} \\ \frac{\partial f(\boldsymbol{\theta})}{\partial \sigma^2} \end{pmatrix} \begin{pmatrix} 0 \\ -\frac{1}{2}e^{\sigma^2}(e^{\sigma^2} - 1)^{-\frac{3}{2}} \end{pmatrix} \,.$$

Taking this Jacobian term into the Delta method, that is $\widehat{SR_l} = f(\hat{\boldsymbol{\theta}}) \sim N(f(\boldsymbol{\theta}), \frac{1}{T}(\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}})' \boldsymbol{V} (\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}))$, we finally prove the asymptotical distribution of SR_l. □

From this distribution, the confidence interval for J.S. Sharpe ratio can be constructed in the usual fashion:

$$\left(\widehat{SR} - z_{\frac{\alpha}{2}} \sqrt{\frac{1}{2T}\hat{\sigma^4} e^{2\hat{\sigma}^2}(e^{\hat{\sigma}^2} - 1)^{-3}}, \widehat{SR} + z_{\frac{\alpha}{2}} \sqrt{\frac{1}{2T}\hat{\sigma^4} e^{2\hat{\sigma}^2}(e^{\hat{\sigma}^2} - 1)^{-3}} \right) \,. \tag{3.2.17}$$

3.2.3.2 Examples and Simulations

This subsection provides examples and simulations for inference on the J.S. Sharpe ratio. More specifically, we compute confidence intervals and p-values using our proposed third-order method given in Barndorff-Nielsen's approximation. We label this method "proposed" in the tables below. To make our results more persuasive, our third order method will compare to asymptotic distribution in (3.2.16) labeled as "Jobson and Korkie". Results from the signed log-likelihood ratio statistic in (3.1) are additionally provided and labeled "likelihood ratio." At last, 10000 times simulation will provided to prove our conclusion solid.

The data set for our examples consists of monthly returns for two time series from Jan 2013 to Dec2013. The data is listed at Table (3.18). The first series represents gross return for a large-cap mutual fund (Fund), it comes from Barclay Hedge Fund Index[7], Which is a measure of the average return of all hedge funds (excepting Funds of Funds) in the Barclay database. The second for a market index (Market), we will see to monthly return of the S&P 500 index, and the raw data is coming from[8].

We first look at estimated confidence intervals. Table (3.19) reports 95% confidence intervals for J.S. Sharpe ratio separately for the large cap mutual fund and the market index for the three methods discussed previously. We can find that the confidence intervals obtained from the three methods produce different results. Theoretically, the proposed method has third-order accuracy whereas the remaining three methods do not, thus the Wald method and likelihood ratio statistic produce relatively small variation between themselves, however a noticeable difference compared with the proposed method. This result will be borne out in the simulations as well. In addition, we notice that the J.S. Sharpe ratio are numerically different from the conventional Sharpe ratio because of distinct definitions.

The p-value functions calculated from the methods for J.S. Sharpe are plotted in Figures (3.13) and (3.14), respectively. These significance functions can be used to obtain p-values for specific hypothesized values of the Sharpe ratio. A few values with their corresponding p-values are provided in Tables (3.20) for the market fund and market index, respectively. From these tables we can see that the p-values vary across the methods. Focusing on the market index and using a 5% level of significance, the J.S Sharpe ratio that is 50, may or may not fall into rejection region depending on the method chosen for the hypothesis test.

Two simulation studies of 10,000 replications were performed to compare the two existing methods to the proposed third-order method. The simulation can be constructed by mimic Fund: $LN(0.008803383, 0.0001252401)$ and Market: $LN(0.02160774, 0.0005975605)$. From standard errors, it can be seen that the proposed method produces results that are uniformly within

[7]Refer towebsite http://www.barclayhedge.com/research/indices/ghs/Hedge_Fund_Index.html

[8]Refer to website http://finance.yahoo.com/q/hp?s=%5EGSPC&a=11&b=1&c=2012&d=00&e=1&f=2014&g=m

Table 3.18: Monthly return for Fund and Market

Month	Gross return Fund	Gross return Market
Jan 2013	1.0248	1.050428
Feb 2013	1.0028	1.011061
Mar 2013	1.0137	1.035988
Apr 2013	1.0054	1.018086
May 2013	1.0087	1.020763
Jun 2013	0.9848	0.985001
Jul 2013	1.0162	1.049462
Aug 2013	0.9939	0.968702
Sep 2013	1.0203	1.029749
Oct 2013	1.0171	1.044596
Nov 2013	1.0075	1.028049
Dec 2013	1.0116	1.023563

Table 3.19: 95% Confidence Intervals for J.S. Sharpe Ratio

Method	95% CI for SR of Fund	95% CI for SR of Market
J & K	55.98732 130.6679	25.62482 59.81766
Likelihood Ratio	58.83526 132.8542	26.92776 60.81810
Proposed	**52.77824 126.3001**	**24.15398 57.81744**

Table 3.20: Fund: p-values for J.S. SR (Up: Fund; Down:Market)

Method	20	40	60	80	100	120	140
J & K	0.9999	0.9974	0.9599	0.7579	0.3631	0.0808	0.0071
Likelihood	1.0000	0.9994	0.9704	0.7636	0.3646	0.0895	0.0108
Proposed	**1.0000**	**0.9977**	**0.9343**	**0.6421**	**0.2482**	**0.0484**	**0.0047**

Method	10	20	30	40	50	60	70
J & K	0.9999	0.9954	0.9276	0.6225	0.2020	0.0238	0.0009
Likelihood	1.0000	0.99852	0.9393	0.6238	0.2081	0.0304	0.0019
Proposed	**1.0000**	**0.9949**	**0.8806**	**0.4877**	**0.1274**	**0.0145**	**0.0007**

Figure 3.13: p-value function for Fund on J.S. Sharpe Ratio

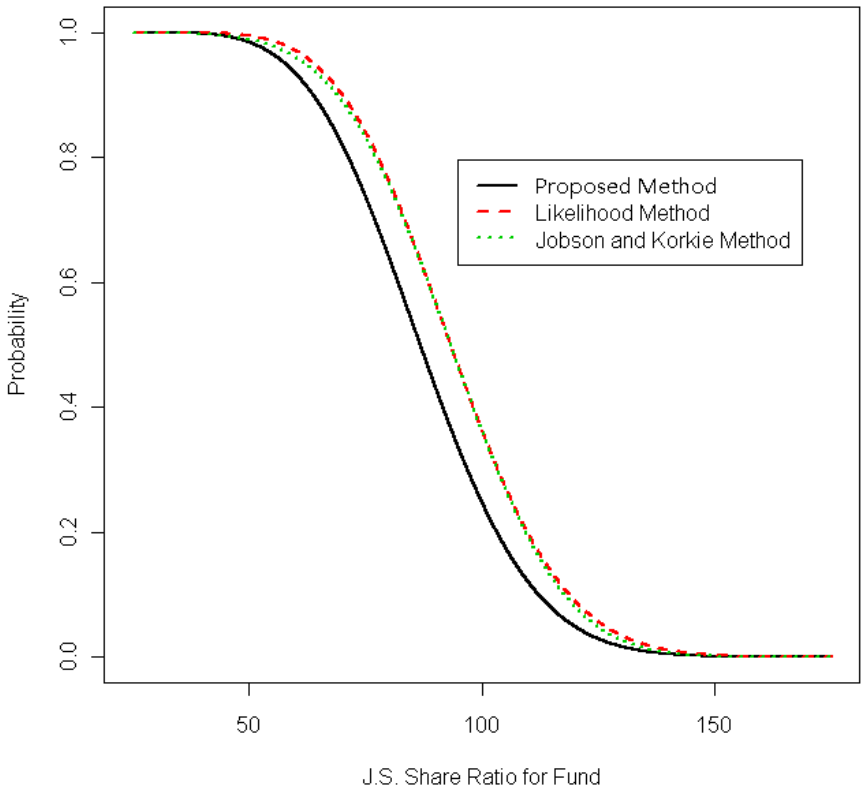


Figure 3.14: p-value function for Market on J.S. Sharpe Ratio

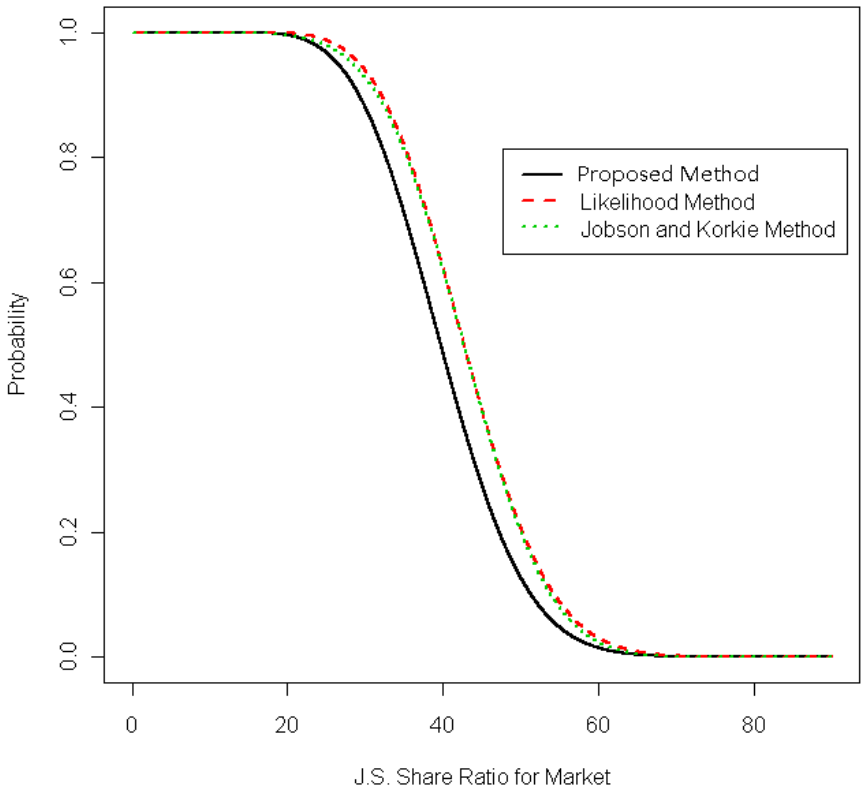


Table 3.21: Results for simulation study on J.S. Sharpe Ratio n=12 (Up: Fund; Down:Market)

CI	Method	Lower Error	Upper Error	Central Coverage
90%	Jobson and Korkie	0.0849	0.0265	0.8886
	Likelihood Ratio	0.1048	0.0218	0.8734
	Proposed	**0.0505**	**0.0479**	**0.9016**
	Nominal	0.0500	0.0500	0.9000
	Standard Error	0.0022	0.0022	0.0030
95%	Jobson and Korkie	0.0389	0.0125	0.9486
	Likelihood Ratio	0.0572	0.0095	0.9333
	Proposed	**0.0250**	**0.0216**	**0.9534**
	Nominal	0.0250	0.0250	0.9500
	Standard Error	0.0016	0.0016	0.0022
99%	Jobson and Korkie	0.0069	0.0026	0.9905
	Likelihood Ratio	0.0148	0.0018	0.9834
	Proposed	**0.0060**	**0.0037**	**0.9903**
	Nominal	0.0050	0.0050	0.9900
	Standard Error	0.0007	0.0007	0.0010

CI	Method	Lower Error	Upper Error	Central Coverage
90%	Jobson and Korkie	0.0793	0.0289	0.8918
	Likelihood Ratio	0.0988	0.0235	0.8777
	Proposed	**0.0478**	**0.0497**	**0.9025**
	Nominal	0.0500	0.0500	0.9000
	Standard Error	0.0022	0.0022	0.0030
95%	Jobson and Korkie	0.0361	0.0142	0.9497
	Likelihood Ratio	0.0548	0.0112	0.934
	Proposed	**0.0226**	**0.0232**	**0.9542**
	Nominal	0.0250	0.0250	0.9500
	Standard Error	0.0016	0.0016	0.0022
99%	Jobson and Korkie	0.0053	0.0022	0.9925
	Likelihood Ratio	0.0137	0.0015	0.9848
	Proposed	**0.0045**	**0.0033**	**0.9922**
	Nominal	0.0050	0.0050	0.9900
	Standard Error	0.0007	0.0007	0.0010

three standard deviations of the nominal value while other two methods produce less satisfactory results.

3.2.4 Likelihood Methodology for J.S. Sharpe Ratio at Two Independent Sample Comparison

Consider two independent funds with sample gross returns: $(g_{X_1}, \ldots, g_{X_n})$ and $(g_{Y_1}, \ldots, g_{Y_m})$. Further assume these returns are identically and independently distributed as $LN(\mu_X, \sigma_X^2)$ and $LN(\mu_Y, \sigma_Y^2)$, respectively.

Our parameter of interest is the subtraction of two sample J.S. Sharpe ratios:

$$\psi(\boldsymbol{\theta}) = \psi \begin{pmatrix} \mu_X \\ \sigma_X^2 \\ \mu_Y \\ \sigma_Y^2 \end{pmatrix} = \frac{1}{\sqrt{e^{\sigma_X^2} - 1}} - \frac{1}{\sqrt{e^{\sigma_Y^2} - 1}} \, . \tag{3.2.18}$$

And its first order derivative is

$$\psi_{\boldsymbol{\theta}'}(\boldsymbol{\theta}) = \left(0, -\frac{1}{2}(e^{\sigma_X^2} - 1)^{-\frac{3}{2}} e^{\sigma_X^2}, 0, \frac{1}{2}(e^{\sigma_Y^2} - 1)^{-\frac{3}{2}} e^{\sigma_Y^2} \right) \, . \tag{3.2.19}$$

The joint log likelihood function is the addition of two sample log likelihood function.

$$\begin{aligned} l(\theta) = l(\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2) = \\ -\frac{n}{2}\log\sigma_X^2 - \frac{\sum(\log(g_{X_i}) - \mu_X)^2}{2\sigma_X^2} - \frac{m}{2}\log\sigma_Y^2 - \frac{\sum(\log(g_{Y_j}) - \mu_Y)^2}{2\sigma_Y^2} \, . \end{aligned} \tag{3.2.20}$$

To obtain unconstrained MLE, we maximize this joint log-likelihood function. Thus we have

$$\begin{aligned} \hat{\boldsymbol{\theta}}' &= (\hat{\mu}, \hat{\sigma}^2) \qquad (3.2.21) \\ &= \left(\frac{\sum \log(g_{X_i})}{n}, \frac{\sum(\log(g_{X_i}) - \hat{\mu}_X)^2}{n}, \frac{\sum \log(g_{Y_j})}{m}, \frac{\sum(\log(g_{Y_j}) - \hat{\mu}_Y)^2}{m}\right) \end{aligned}$$

When we consider the constrained maximum likelihood estimation, actually we do not need to take μ_X or μ_Y into consideration, since the J.S. Sharpe ratio only relates to variance. Thus, $\tilde{\mu}_X = \hat{\mu}_X = \frac{\sum \log(g_{X_i})}{n}$ and $\tilde{\mu}_Y = \hat{\mu}_Y = \frac{\sum \log(g_{Y_j})}{m}$. Now we focus critically on the variance related parts:

$$l(\boldsymbol{\theta},\alpha) = l(\boldsymbol{\theta}) + \alpha(\psi(\theta) - \psi) = l(\boldsymbol{\theta}) + \alpha(\frac{1}{\sqrt{e^{\sigma_X^2}-1}} - \frac{1}{\sqrt{e^{\sigma_Y^2}-1}} - \psi)\,, \quad (3.2.22)$$

$$l_{\sigma_X^2}(\boldsymbol{\theta},\alpha) = l_{\sigma_X^2}(\boldsymbol{\theta}) - \frac{\alpha}{2} e^{\sigma_X^2}(e^{\sigma_X^2}-1)^{-\frac{3}{2}}\,, \quad (3.2.23)$$

$$l_{\sigma_Y^2}(\boldsymbol{\theta},\alpha) = l_{\sigma_Y^2}(\boldsymbol{\theta}) + \frac{\alpha}{2} e^{\sigma_Y^2}(e^{\sigma_Y^2}-1)^{-\frac{3}{2}}\,, \quad (3.2.24)$$

$$l_{\alpha}(\boldsymbol{\theta},\alpha) = \frac{1}{\sqrt{e^{\sigma_X^2}-1}} - \frac{1}{\sqrt{e^{\sigma_Y^2}-1}} - \psi\,, \quad (3.2.25)$$

$$\tilde{l}_{\sigma_X^2\sigma_X^2}(\boldsymbol{\theta}) = l_{\sigma_X^2\sigma_X^2}(\boldsymbol{\theta}) - \frac{\hat{\alpha}}{2}((e^{\sigma_X^2}-1)^{-\frac{3}{2}} e^{\sigma_X^2} - \frac{3}{2}(e^{\sigma_X^2}-1)^{-\frac{5}{2}} e^{2\sigma_X^2})\,, \quad (3.2.26)$$

$$\tilde{l}_{\sigma_Y^2\sigma_Y^2}(\boldsymbol{\theta}) = l_{\sigma_Y^2\sigma_Y^2}(\boldsymbol{\theta}) + \frac{\hat{\alpha}}{2}((e^{\sigma_Y^2}-1)^{-\frac{3}{2}} e^{\sigma_Y^2} - \frac{3}{2}(e^{\sigma_Y^2}-1)^{-\frac{5}{2}} e^{2\sigma_Y^2})\,. \quad (3.2.27)$$

The calculation for unconstrained MLE here only involves three first order conditions (3.2.23) to (3.2.25) . After getting $\hat{\boldsymbol{\theta}}_{\psi}^{'} = (\tilde{\mu}_X, \tilde{\sigma}_X^2, \tilde{\mu}_Y, \tilde{\sigma}_Y^2)$, we can then solve the $\hat{\alpha}$ from

$$\hat{\alpha} = -\frac{-\frac{m}{\tilde{\sigma}_Y^2} + \frac{\sum(\log(g_{Y_j})-\tilde{\mu}_Y)^2}{\tilde{\sigma}_Y^4}}{e^{\sigma_Y^2}(e^{\sigma_Y^2}-1)^{-\frac{3}{2}}} = \frac{-\frac{n}{\tilde{\sigma}_X^2} + \frac{\sum(\log(g_{X_i})-\tilde{\mu}_X)^2}{\tilde{\sigma}_X^4}}{e^{\tilde{\sigma}_X^2}(e^{\tilde{\sigma}_X^2}-1)^{-\frac{3}{2}}}\,. \quad (3.2.28)$$

Then we get all the values needed by our proposed method.

3.2.5 Examples and Simulations for Two Independent Sample Comparison on J.S. Sharpe ratio

In this part, we implement the third order likelihood methodology derived from last subsection into the data in section (3.2.3). And we will compare the third order likelihood based inference method to the classical methods used for testing, namely, Jobson and Korkie and the likelihood ratio statistic.

We continue to use the data presented in Table (3.18) to compare Sharpe ratios. We may, for instance, be interested in whether the mutual fund's

risk-adjusted return as captured by the Sharpe ratio is significantly better than the market's return. In Table (3.22), we present the 95% confidence interval for the difference between the Sharpe ratios for the mutual fund and market index. As this is an example, we cannot comment on which interval is more accurate, but it is relevant to note the differences between the intervals which may be important in real world settings.

For Table (3.23), people can check p-values for testing a null hypothesis of a zero difference between the Sharpe ratios of the mutual fund and market index. As tail probabilities tend to be small probabilities, it is important to approximate these as accurately as we can. In particular, we find that when we measure in J.S. Sharpe ratio, the mutual fund's risk-adjusted return is better than market average return at 99% significant level. It can be used for comparison of the performance levels last year. At last, the p-value functions calculated from the methods discussed in this paper for both types of Sharpe ratio are plotted in Figures (3.15).

In this section, we provide a simulation study to assess the performance of the third order method relative to the Jobson and Korkie method and likelihood ratio. Table (3.24) records the results from bundles of simulation. The size of each simulation is 10,000 and the parameter are chosen as Fund X: $LN(0.008803383, 0.0001252401)$ and Fund Y: $LN(0.02160774, 0.0005975605)$. As in the one sample case, these simulation results generally indicate that the proposed method outperforms the other methods based on the criteria we examined.

In financial market, the assumption to return based on time series structure is much popular and close to data observation than IID structure. Thus, in the next chapter, we will precede to the inference of Sharpe ratio given the assumption of autoregressive return. For our proposed third-order likelihood method, the most significant difference from IID structure is that the canonical parameters is no longer directly available and we need to obtain a locally defined canonical parameter instead.

Table 3.22: 95% Confidence intervals for Sharpe ratio difference

Method	95% for Conventional Sharpe Ratio	95% for J.S. Sharpe Ratio
Jobson and Korkie	-1.060531 0.7980580	9.538317 91.67439
Likelihood Ratio	-1.061872 0.7967091	11.54955 93.36287
Proposed	**-1.050943 0.8052257**	**8.636630 90.70401**

Table 3.23: p-values for testing a null hypothesis of a zero difference between the Sharpe ratios

Method	For J.S. Sharpe Ratio
Jobson and Korkie	0.9921362
Likelihood Ratio	0.9951321
Proposed	**0.9923669**

Figure 3.15: p-value function for two Sample Comparison

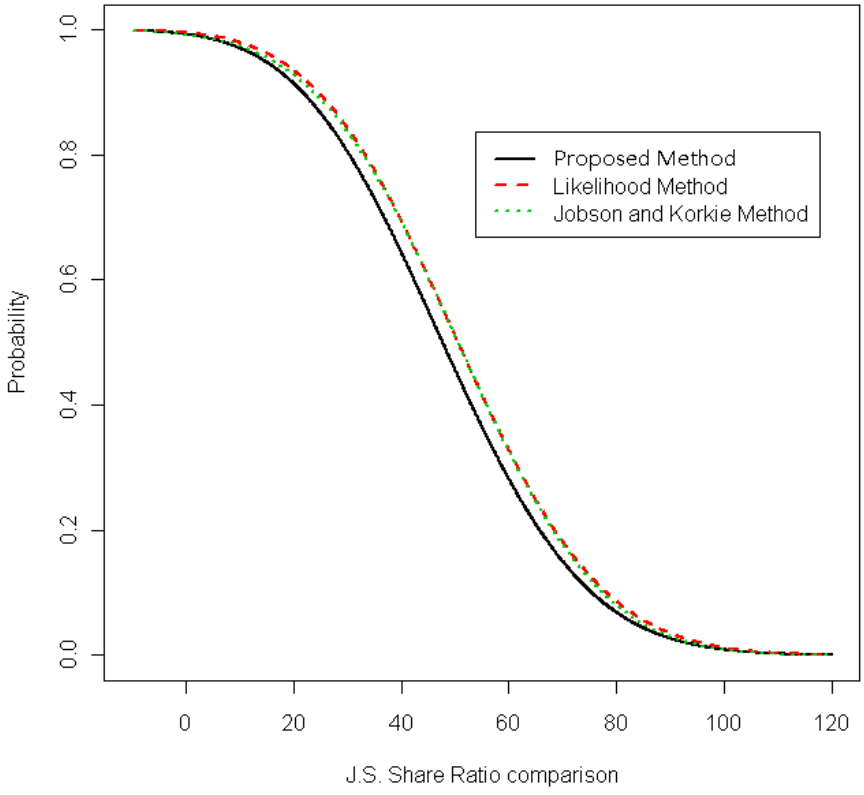


Table 3.24: Simulation Studies for the difference of two J.S. Sharpe Ratios

CI	n,m	Method	Lower Error	Upper Error	Central Coverage
90%	n=12,m=12	Jobson and Korkie	0.0608	0.0492	0.8900
		Likelihood Ratio	0.0726	0.0468	0.8806
		Proposed	**0.0471**	**0.0500**	**0.9029**
	n=12,m=24	Jobson and Korkie	0.0750	0.0392	0.8858
		Likelihood Ratio	0.0908	0.0362	0.8730
		Proposed	**0.0534**	**0.0474**	**0.8992**
	n=24,m=12	Jobson and Korkie	0.0436	0.0661	0.8903
		Likelihood Ratio	0.0464	0.0697	0.8839
		Proposed	**0.0470**	**0.0529**	**0.9001**
	Reference	*Nominal*	*0.0500*	*0.0500*	*0.9000*
		Standard Error	*0.0022*	*0.0022*	*0.0030*
95%	n=12,m=12	Jobson and Korkie	0.0291	0.0256	0.9453
		Likelihood Ratio	0.0396	0.0244	0.9360
		Proposed	**0.0248**	**0.0245**	**0.9507**
	n=12,m=24	Jobson and Korkie	0.0349	0.0178	0.9473
		Likelihood Ratio	0.0482	0.0160	0.9358
		Proposed	**0.0273**	**0.0225**	**0.9502**
	n=24,m=12	Jobson and Korkie	0.0214	0.0334	0.9452
		Likelihood Ratio	0.0243	0.0380	0.9377
		Proposed	**0.0229**	**0.0257**	**0.9514**
	Reference	*Nominal*	*0.0250*	*0.0250*	*0.9500*
		Standard Error	*0.0016*	*0.0016*	*0.0022*
99%	n=12,m=12	Jobson and Korkie	0.0038	0.0058	0.9904
		Likelihood Ratio	0.0106	0.0059	0.9835
		Proposed	**0.0048**	**0.0056**	**0.9896**
	n=12,m=24	Jobson and Korkie	0.0060	0.0046	0.9894
		Likelihood Ratio	0.0134	0.0040	0.9826
		Proposed	**0.0061**	**0.0053**	**0.9886**
	n=24,m=12	Jobson and Korkie	0.0034	0.0042	0.9924
		Likelihood Ratio	0.0045	0.0080	0.9875
		Proposed	**0.0041**	**0.0035**	**0.9924**
	Reference	*Nominal*	*0.0050*	*0.0050*	*0.9900*
		Standard Error	*0.0007*	*0.0007*	*0.0010*

Chapter 4

Asymptotic Likelihood Inference for Sharpe Ratio under Gaussian Autocorrelated Return

4.1 Likelihood Methodology for One Sample Sharpe Ratio under AR(1) Return

The third-order methodology discussed in Chapter 2 is applicable under any parametric distributional assumptions, and in this chapter we will first demonstrate the use of the method under the Gaussian AR(1) returns.

Consider a fund with log-return at time t denoted by r_t, where $t = 1, 2, \ldots, T$. Under Gaussian AR(1) assumption on this return series, we have the following basic setting: [1]

$$\begin{cases} r_t = \mu + \epsilon'_t & t \geqslant 1; \\ \epsilon'_t = \rho\epsilon'_{t-1} + \sigma v_t & t \geq 2; \\ v_t \sim N(0,1) & t \geqslant 2. \end{cases}$$

Additionally, to make this AR(1) process stationary, we assume

$$|\rho| < 1 \, .$$

Stationary process is a stochastic process whose joint probability distribution does not change when shifted in time. Consequently, parameters such

[1] Another form of expression can be $\begin{cases} r_t - \mu = \rho(r_{t-1} - \mu) + \sigma v_t & t \geq 2 \\ v_t \sim N(0,1) & t \geqslant 2 \end{cases}$

as the mean and variance, if they are present, also do not change over time and do not follow any trends. Thus, at stationary AR(1) setting, we have

$$r_t \sim N\left(\mu, \tfrac{\sigma^2}{1-\rho^2}\right) \quad \text{for } t \geqslant 1 \quad , \tag{4.1.1}$$

$$Cov\left(r_i, r_j\right) = \frac{\sigma^2 \rho^{|i-j|}}{1-\rho^2} \quad \text{for any } i,j\text{in (1,2,..., n)} \, . \tag{4.1.2}$$

Both expressions in (4.1.1) and (4.1.2) are independent of t, which reinsure stationary process. (4.1.1) tells that the expectation of return series is μ and the variance is $\frac{\sigma^2}{1-\rho^2}$, and we can use this result to obtain the expression of parameter of interest and its derivative:

$$\psi\left(\boldsymbol{\theta} = \begin{pmatrix} \rho \\ \mu \\ \sigma^2 \end{pmatrix}\right) = \frac{\mu_{r_t} - \mu_f}{\sqrt{var\left(r_t\right)}} = \frac{\mu - \mu_f}{\sqrt{\frac{\sigma^2}{1-\rho^2}}} \, ,$$

$$\psi_{\boldsymbol{\theta}'}\left(\boldsymbol{\theta}\right) = \left(\psi_\rho\left(\boldsymbol{\theta}\right), \psi_\mu\left(\boldsymbol{\theta}\right), \psi_{\sigma^2}\left(\boldsymbol{\theta}\right)\right) \, .$$

For a clearer understanding of our testing procedure, it is useful to rewrite our model in reduced matrix formulations:

$$\mathbf{r} = \mu \cdot \mathbf{1} + \sigma \cdot \boldsymbol{\epsilon} = \mu \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \sigma \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_T \end{pmatrix} \, ,$$

$$\epsilon \sim N\left(\mathbf{0}, \boldsymbol{\Omega} = \left(\frac{\rho^{|i-j|}}{1-\rho^2}\right)_{ij}\right) \text{ or } \sigma\boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma} = \sigma^2 \cdot \boldsymbol{\Omega} = \left(\frac{\sigma^2 \rho^{|i-j|}}{1-\rho^2}\right)_{ij}\right) \, .$$

In addition, what we will make great use of later is the inverse matrix of $\boldsymbol{\Omega}$, its Cholesky decomposition and its derivative matrix. Specifically,

$$\mathbf{A} = \boldsymbol{\Omega}^{-1} = \begin{pmatrix} 1 & -\rho & 0 & 0 & \cdots & 0 & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & \cdots & 0 & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -\rho & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & 0 & \cdots & 0 & -\rho & 1 \end{pmatrix} = \mathbf{L}'\mathbf{L} \, ,$$

$$\mathbf{A}_\rho = \frac{\partial \mathbf{A}}{\partial \rho} = \begin{pmatrix} 0 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2\rho & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 2\rho & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 2\rho & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 0 \end{pmatrix} \, ,$$

$$\mathbf{A}_{\rho\rho} = \frac{\partial^2 \mathbf{A}}{\partial \rho^2} = \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 2 & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix},$$

$$\mathbf{L} = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -\rho & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\rho & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & -\rho & 1 \end{pmatrix}, \tag{4.1.3}$$

$$\mathbf{L}_\rho = \frac{\partial \mathbf{L}}{\partial \rho} = \begin{pmatrix} \frac{-\rho}{\sqrt{1-\rho^2}} & 0 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & -1 & 0 \end{pmatrix}.$$

Note that the dependence of matrices $\mathbf{A}$ and $\mathbf{L}$ are only on the parameter ρ. [2,3]

---

[2] Another way to express $\mathbf{A}$ is $a_{ii} = \begin{cases} 1 & i = 1 \text{ or } n \\ 1+\rho^2 & i = 2, 3, \ldots n-1 \end{cases}$ and $a_{ij} = \begin{cases} -\rho & |i-j| = 1 \\ 0 & \text{other cases when i} \neq \text{j} \end{cases}$

[3] Except Cholesky decomposition, we have another more statistical way to obtain $\mathbf{A} = \mathbf{\Omega}^{-1} = \mathbf{L}'\mathbf{L}$.

At footnote (1), we know our model can be expressed as

$$\begin{cases} r_t - \mu = \rho\left(r_{t-1} - \mu\right) + \sigma v_t & t \geq 2; \\ v_t \sim N\left(0, 1\right) & t \geqslant 2; \\ r_1 \sim N\left(\mu, \frac{\sigma^2}{1-\rho^2}\right) & t = 1. \end{cases}$$

This allows us to construct a matrix $\mathbf{L}$ such that

$$\mathbf{L}\left(\mathbf{r} - \mu \cdot \mathbf{1}\right) = \sigma \cdot \mathbf{v}\,,$$

The property for $\mathbf{L}$ is that for all $t \geq 2$, the t th row of $\mathbf{L}$ has 1 in the t th position, $-\rho$ in the $(t-1)$ st position and 0s everywhere else. To account for the first observation, we set the first row of $\mathbf{L}$ to have $\sqrt{1-\rho^2}$ in the first position, and 0s everywhere else. Therefore, by this method the $\mathbf{L}$ here is exactly the same as Cholesky decomposition of $\mathbf{A}$ in (4.1.3).

We have $\mathbf{r} - \mu \cdot \mathbf{1} = \mathbf{L}^{-1}\sigma \cdot \mathbf{v} = \sigma \cdot \boldsymbol{\epsilon}$ which implies that

$$\mathbf{L}^{-1}\mathbf{v} = \boldsymbol{\epsilon} \text{ or } \mathbf{v} = \mathbf{L} \cdot \boldsymbol{\epsilon}\,,$$

Since the variance of $\mathbf{v}$ is identity matrix, then

$$var\left(\mathbf{L}\boldsymbol{\epsilon}\right) = E\left(\mathbf{L}\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{L}'\right) = \mathbf{L}E\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right)\mathbf{L}' = \mathbf{L}\mathbf{\Omega}\mathbf{L}' = \mathbf{I}\,.$$

For the parameter vector, $\boldsymbol{\theta} = \begin{pmatrix} \rho \\ \mu \\ \sigma^2 \end{pmatrix}$, the probability density function of $\mathbf{r}_t$ is given by

$$f(\mathbf{r};\boldsymbol{\theta}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{r}-\mu\cdot\mathbf{1})'\boldsymbol{\Sigma}^{-1}(\mathbf{r}-\mu\cdot\mathbf{1})} ,$$

Or

$$\begin{aligned} f(\mathbf{r};\boldsymbol{\theta}) &= f(r_2,\dots r_n|r_1;\boldsymbol{\theta}) \cdot f(r_1;\boldsymbol{\theta}) \\ &= \{f(r_n|r_{n-1};\boldsymbol{\theta}) \cdot f(r_{n-1}|r_{n-2};\boldsymbol{\theta}) \cdots f(r_2|r_1;\boldsymbol{\theta})\} \cdot f(r_1;\boldsymbol{\theta}) \\ &= \prod_{i=2}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(r_i-\mu-\rho(r_{i-1}-\mu))^2} \right) \cdot \frac{1}{\sqrt{2\pi\frac{\sigma^2}{1-\rho^2}}} e^{-\frac{1}{2\frac{\sigma^2}{1-\rho^2}}(r_1-\mu)^2} . \end{aligned}$$

Since $|\boldsymbol{\Omega}| = \frac{1}{1-\rho^2}$, $|\boldsymbol{\Sigma}| = \frac{\sigma^{2n}}{1-\rho^2}$, and $\boldsymbol{\Sigma}^{-1} = \frac{\boldsymbol{\Omega}^{-1}}{\sigma^2}$, the log-likelihood function can be written as

$$\ell(\boldsymbol{\theta}) = a - \frac{n}{2}\log\sigma^2 + \frac{1}{2}\log\left(1-\rho^2\right) - \frac{1}{2\sigma^2}(\mathbf{r}-\mu\cdot\mathbf{1})'\mathbf{A}(\mathbf{r}-\mu\cdot\mathbf{1}) ,$$

Or

$$\begin{aligned} \ell(\boldsymbol{\theta}) = a &- \frac{n-1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=2}^{T}(r_i-\mu-\rho(r_{i-1}-\mu))^2 \\ &- \frac{1}{2}\log\frac{\sigma^2}{1-\rho^2} - \frac{1}{2\frac{\sigma^2}{1-\rho^2}}(r_1-\mu)^2 \\ = a &- \frac{n}{2}\log\sigma^2 + \frac{1}{2}\log\left(1-\rho^2\right) - \frac{1}{2\sigma^2}\sum_{i=2}^{T}(r_i-\mu-\rho(r_{i-1}-\mu))^2 \\ &- \frac{1-\rho^2}{2\sigma^2}(r_1-\mu)^2 . \end{aligned}$$

Next, we can start with unconstrained maximum likelihood estimation, which maximize the log-likelihood function $l(\boldsymbol{\theta})$. First order and second order derivatives of $\ell(\boldsymbol{\theta})$ are calculated as follows:

$$l_\rho(\boldsymbol{\theta}) = \frac{-\rho}{1-\rho^2} - \frac{(\mathbf{r}-\mu\cdot\mathbf{1})'\mathbf{A}_\rho(\mathbf{r}-\mu\cdot\mathbf{1})}{2\sigma^2} , \tag{4.1.4}$$

$$l_\mu(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\mathbf{1}'\mathbf{A}(\mathbf{r}-\mu\cdot\mathbf{1}) , \tag{4.1.5}$$

$$l_{\sigma^2}(\boldsymbol{\theta}) = -\frac{n}{2\sigma^2} + \frac{(\mathbf{r}-\mu\cdot\mathbf{1})'\mathbf{A}(\mathbf{r}-\mu\cdot\mathbf{1})}{2\sigma^4} , \tag{4.1.6}$$

$$l_{\rho\rho}(\boldsymbol{\theta}) = -\frac{1+\rho^2}{\left(1-\rho^2\right)^2} - \frac{\left(\mathbf{r}-\mu\cdot\mathbf{1}\right)'\mathbf{A}_{\rho\rho}\left(\mathbf{r}-\mu\cdot\mathbf{1}\right)}{2\sigma^2}\ ,$$

$$l_{\rho\mu}(\boldsymbol{\theta}) = \frac{\mathbf{1}'\mathbf{A}_{\rho}\left(\mathbf{r}-\mu\cdot\mathbf{1}\right)}{\sigma^2}\ ,$$

$$l_{\rho\sigma^2}(\boldsymbol{\theta}) = \frac{\left(\mathbf{r}-\mu\cdot\mathbf{1}\right)'\mathbf{A}_{\rho}\left(\mathbf{r}-\mu\cdot\mathbf{1}\right)}{2\sigma^4}\ ,$$

$$l_{\mu\mu}(\boldsymbol{\theta}) = \frac{-1}{\sigma^2}\mathbf{1}'\mathbf{A}\cdot\mathbf{1}\ ,$$

$$l_{\mu\sigma^2}(\boldsymbol{\theta}) = -\frac{1}{\sigma^4}\mathbf{1}'\mathbf{A}\left(\mathbf{r}-\mu\cdot\mathbf{1}\right)\ ,$$

$$l_{\sigma^2\sigma^2}(\boldsymbol{\theta}) = \frac{T}{2\sigma^4} - \frac{\left(\mathbf{r}-\mu\cdot\mathbf{1}\right)'\mathbf{A}\left(\mathbf{r}-\mu\cdot\mathbf{1}\right)}{\sigma^6}\ .$$

To obtain the unrestricted maximum likelihood estimator $\hat{\boldsymbol{\theta}}$, we solve simultaneously the first order conditions, from (4.1.4) and (4.1.6) equal to zero. Unfortunately, only numerical solutions can be obtained and some iterative procedure is needed. Given this information about the overall maximum likelihood estimate, we can obtain other important variable for future use, such as the estimated Sharpe ratio $\hat{\psi} = \frac{\hat{\mu}-\mu_f}{\sqrt{\frac{\hat{\sigma^2}}{1-\hat{\rho}^2}}}$, the estimated unrestricted likelihood function $\ell\left(\hat{\boldsymbol{\theta}}\right)$, the observed information matrix evaluated at $\hat{\boldsymbol{\theta}}$, $\mathbf{j}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} -\ell_{\rho\rho}\left(\hat{\boldsymbol{\theta}}\right) & -\ell_{\rho\mu}\left(\hat{\boldsymbol{\theta}}\right) & -\ell_{\rho\sigma^2}\left(\hat{\boldsymbol{\theta}}\right) \\ -\ell_{\mu\rho}\left(\hat{\boldsymbol{\theta}}\right) & -l_{\mu\mu}(\hat{\boldsymbol{\theta}}) & -l_{\mu\sigma^2}(\hat{\boldsymbol{\theta}}) \\ -\ell_{\sigma^2\rho}\left(\hat{\boldsymbol{\theta}}\right) & -l_{\sigma^2\mu}(\hat{\boldsymbol{\theta}}) & -l_{\sigma^2\sigma^2}(\hat{\boldsymbol{\theta}}) \end{pmatrix}$, [4] and its determinant.

To derive the constrained MLE, the log-likelihood function $\ell(\boldsymbol{\theta})$ must be maximized with respect to ρ, μ, σ^2 while holding ψ fixed and this process can be managed by the Lagrange multiplier method (see (2.3.7)). The La-

[4] We can actually obtain the expected Fisher full information matrix as

$$\begin{aligned} \mathbf{I}(\hat{\boldsymbol{\theta}}) &= \begin{pmatrix} -E\left[\ell_{\rho\rho}\left(\hat{\boldsymbol{\theta}}\right)\right] & -E\left[\ell_{\rho\mu}\left(\hat{\boldsymbol{\theta}}\right)\right] & -E\left[\ell_{\rho\sigma^2}\left(\hat{\boldsymbol{\theta}}\right)\right] \\ -E\left[\ell_{\mu\rho}\left(\hat{\boldsymbol{\theta}}\right)\right] & -E\left[l_{\mu\mu}(\hat{\boldsymbol{\theta}})\right] & -E\left[l_{\mu\sigma^2}(\hat{\boldsymbol{\theta}})\right] \\ -E\left[\ell_{\sigma^2\rho}\left(\hat{\boldsymbol{\theta}}\right)\right] & -E\left[l_{\sigma^2\mu}(\hat{\boldsymbol{\theta}})\right] & -E\left[l_{\sigma^2\sigma^2}(\hat{\boldsymbol{\theta}})\right] \end{pmatrix} \\ &= \begin{pmatrix} \frac{1+\rho^2}{\left(1-\rho^2\right)^2} + \frac{n-2}{1-\rho^2} & 0 & \frac{\rho}{\sigma^2\left(1-\rho^2\right)} \\ 0 & \frac{(n-2)(\rho-1)^2+2(1-\rho)}{\sigma^2} & 0 \\ \frac{\rho}{\sigma^2\left(1-\rho^2\right)} & 0 & \frac{n}{2\sigma^4} \end{pmatrix} \end{aligned}$$

Theoretically, the application of expected Fisher full information matrix instead of the observed information matrix evaluated at $\hat{\boldsymbol{\theta}}$ should increase the accuracy of our method, but since the improvement is not quite significant when we put into simulation, in our paper we will still use the observed information for the proposed method.

grangian function is given at (4.1.7) and its first order derivatives are listed from (4.1.8) to (4.1.11).

$$H(\boldsymbol{\theta},\alpha)=l(\boldsymbol{\theta})+\alpha(\psi(\boldsymbol{\theta})-\psi)=l(\boldsymbol{\theta})+\alpha\left(\frac{\mu-\mu_f}{\sqrt{\frac{\sigma^2}{1-\rho^2}}}-\psi\right)\ , \tag{4.1.7}$$

$$H_\rho(\boldsymbol{\theta},\alpha)=l_\rho(\boldsymbol{\theta})+\alpha\psi_\rho\left(\boldsymbol{\theta}\right)\ , \tag{4.1.8}$$

$$H_\mu(\boldsymbol{\theta},\alpha)=l_\mu(\boldsymbol{\theta})+\alpha\psi_\mu\left(\boldsymbol{\theta}\right)\ , \tag{4.1.9}$$

$$H_{\sigma^2}(\boldsymbol{\theta},\alpha)=l_{\sigma^2}(\boldsymbol{\theta})+\alpha\psi_{\sigma^2}\left(\boldsymbol{\theta}\right)\ , \tag{4.1.10}$$

$$H_\alpha(\boldsymbol{\theta},\alpha)=\frac{\mu-\mu_f}{\sqrt{\frac{\sigma^2}{1-\rho^2}}}-\psi\ . \tag{4.1.11}$$

Solving first order derivatives, from (4.1.8) to (4.1.11), equal to zero, we can obtain the restricted maximum likelihood estimator, $\hat{\boldsymbol{\theta}}_\psi^{'}=(\tilde{\rho},\tilde{\mu},\tilde{\sigma}^2)$. The tilted log-likelihood function can be obtained by replacing α by $\hat{\alpha}$ on the Lagrangian function, and taking its second order derivatives. In addition, we can also obtain the estimated restricted likelihood function: $l(\hat{\boldsymbol{\theta}}_\psi)$, the tilted observed information matrix evaluated at $\hat{\boldsymbol{\theta}}_\psi$, $\tilde{\mathbf{j}}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_\psi)=-\frac{\partial^2\tilde{\mathbf{j}}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_\psi}=j_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_\psi)-\hat{\alpha}\psi_{\boldsymbol{\theta}\boldsymbol{\theta}'}\left(\hat{\boldsymbol{\theta}}_\psi\right)$ and its inverse $\tilde{\mathbf{j}}^{-1}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_\psi)$.

In Chapter 3, the canonical parameters for normal distribution and log-normal distribution are available explicitly from the exponential family; However, here the general canonical parameter for AR(1) is not available but we can obtain a locally defined canonical parameter from (2.6.15) and (2.6.16). To do that, we first need to find a full-dimensional pivotal quantity $\mathbf{z}$. Since $\mathbf{r}=\mu\cdot\mathbf{1}+\sigma\cdot\boldsymbol{\epsilon}$, and $var\left(\sigma\boldsymbol{\epsilon}\right)=\boldsymbol{\Sigma}=\sigma^2\cdot\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}^{-1}=\mathbf{L}'\mathbf{L}$, the pivotal quantity $\mathbf{z}$ for this problem is specified as the vector of independent standard normal deviates:

$$\begin{aligned}\mathbf{z} &= \boldsymbol{\Sigma}^{-\frac{1}{2}}\left(\mathbf{r}-\mu\cdot\mathbf{1}\right)=\frac{\boldsymbol{\Omega}^{-\frac{1}{2}}\left(\mathbf{r}-\mu\cdot\mathbf{1}\right)}{\sigma}=\frac{\mathbf{L}\left(\mathbf{r}-\mu\cdot\mathbf{1}\right)}{\sigma}\\ &= \begin{pmatrix}\sqrt{1-\rho^2}\epsilon_1\\ \epsilon_2-\rho\epsilon_1\\ \vdots\\ \epsilon_n-\rho\epsilon_{n-1}\end{pmatrix}.\end{aligned}$$

This choice of pivotal quantity coincides with the standard quantity used to estimate the parameters of an AR(1) model in the literature (see for example Hamilton (1994)). Then, the ancillary direction array $\mathbf{V}$ can be

constructed from (2.6.16)

$$\begin{aligned}
\mathbf{V} &= -\left(\frac{\partial \mathbf{z}}{\partial \mathbf{r}'}\right)^{-1} \frac{\partial \mathbf{z}}{\partial \boldsymbol{\theta}'}\bigg|_{\hat{\boldsymbol{\theta}}} \\
&= -\left(\frac{\mathbf{L}}{\sigma}\right)^{-1} \cdot \begin{pmatrix} \frac{\partial \mathbf{z}}{\partial \rho} & \frac{\partial \mathbf{z}}{\partial \mu} & \frac{\partial \mathbf{z}}{\partial \sigma^2} \end{pmatrix}\bigg|_{\hat{\boldsymbol{\theta}}} \\
&= -\left(\frac{\mathbf{L}}{\sigma}\right)^{-1} \cdot \begin{pmatrix} \frac{\mathbf{L}_\rho(\mathbf{r}-\mu\cdot\mathbf{1})}{\sigma} & -\frac{\mathbf{L}\cdot\mathbf{1}}{\sigma} & -\frac{\mathbf{L}(\mathbf{r}-\mu\cdot\mathbf{1})}{2\sigma^3} \end{pmatrix}\bigg|_{\hat{\boldsymbol{\theta}}} \\
&= \begin{pmatrix} -\hat{\mathbf{L}}^{-1}\hat{\mathbf{L}}_\rho\left(\mathbf{r}-\hat{\mu}\cdot\mathbf{1}\right) & \mathbf{1} & \frac{\mathbf{r}-\hat{\mu}\cdot\mathbf{1}}{2\hat{\sigma}^2} \end{pmatrix}.
\end{aligned}$$

Note that $\mathbf{V}$ is a matrix of sample return $\mathbf{r}$ and it is not related to the parameter $\boldsymbol{\theta}$. Finally, the new locally defined canonical parameter at the data $\mathbf{r}$ can be obtained from (2.6.15), given that the sample space gradient of the likelihood evaluated at the data is $\frac{\partial}{\partial \mathbf{r}'}\ell\left(\boldsymbol{\theta}\right) = -\frac{1}{\sigma^2}\left(\mathbf{r}-\mu\cdot\mathbf{1}\right)'\mathbf{A}$.

$$\begin{aligned}
\boldsymbol{\varphi}'\left(\boldsymbol{\theta}\right) = \begin{pmatrix} \varphi_1\left(\boldsymbol{\theta}\right) & \varphi_2\left(\boldsymbol{\theta}\right) & \varphi_3\left(\boldsymbol{\theta}\right) \end{pmatrix} &= \frac{\partial}{\partial \mathbf{r}'}\ell\left(\boldsymbol{\theta}\right)\cdot\mathbf{V} \\
&= -\frac{1}{\sigma^2}\left(\mathbf{r}-\mu\cdot\mathbf{1}\right)'\mathbf{A}\cdot\mathbf{V}.
\end{aligned}$$

Or $\boldsymbol{\varphi}\left(\boldsymbol{\theta}\right) = \begin{pmatrix} \varphi_1\left(\boldsymbol{\theta}\right) \\ \varphi_2\left(\boldsymbol{\theta}\right) \\ \varphi_3\left(\boldsymbol{\theta}\right) \end{pmatrix} = \mathbf{V}'\cdot\frac{\partial}{\partial \mathbf{r}}\ell\left(\boldsymbol{\theta}\right) = \mathbf{V}'\cdot\left(-\frac{1}{\sigma^2}\mathbf{A}\left(\mathbf{r}-\mu\cdot\mathbf{1}\right)\right)$. In addition, we also need the first order derivative of canonical parameter,

$$\boldsymbol{\varphi}_{\boldsymbol{\theta}'}\left(\boldsymbol{\theta}\right) = \mathbf{V}'\cdot\left(\frac{\partial^2}{\partial \mathbf{r}\partial \boldsymbol{\theta}'}\ell\left(\boldsymbol{\theta}\right)\right) = \mathbf{V}'\cdot\begin{pmatrix} \frac{\partial^2\ell(\boldsymbol{\theta})}{\partial \mathbf{r}\partial\rho} & \frac{\partial^2\ell(\boldsymbol{\theta})}{\partial \mathbf{r}\partial\mu} & \frac{\partial^2\ell(\boldsymbol{\theta})}{\partial \mathbf{r}\partial\sigma^2} \end{pmatrix},$$

where $\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial \mathbf{r}\partial\rho} = -\frac{1}{\sigma^2}\mathbf{A}_\rho\left(\mathbf{r}-\mu\cdot\mathbf{1}\right)$, $\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial \mathbf{r}\partial\mu} = \frac{1}{\sigma^2}\mathbf{A}\mathbf{1}$, and $\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial \mathbf{r}\partial\sigma^2} = \frac{1}{\sigma^4}\mathbf{A}\left(\mathbf{r}-\mu\cdot\mathbf{1}\right)$. Before we precede, please note the dimension reduction here: the dimension is reduced from n (the dimension of $\mathbf{r}$) to 3 (the dimension of the parameter $\boldsymbol{\theta}$ evidenced from the expression for $\boldsymbol{\varphi}\left(\boldsymbol{\theta}\right)$) [5]. To further reduce the dimension of the problem from 3 to the dimension of the parameter of interesting ψ, the calculation of newly calibrated parameter $\chi\left(\boldsymbol{\theta}\right)$ which in turn involves the parameter vector $\boldsymbol{\varphi}\left(\boldsymbol{\theta}\right)$ as well as the constrained MLE $\hat{\boldsymbol{\theta}}_\psi$ is required.

With the above information, the signed log-likelihood ratio statistic R can be constructed from (2.3.15), the newly calibrated parameter χ can be calculated from (2.6.17), the modified maximum likelihood departure measure Q can be obtained from (2.6.18) and (2.6.19), and finally, the proposed third order likelihood approximation Barndorff-Nielsen method $R^\star$ can be obtained from (2.6.13). Unfortunately, an explicit formula is not available as a closed form solution for the MLE does not exist.

A centered $(1-\alpha)\times 100\%$ confidence interval for ψ can be obtained by

[5] The dimension of variables are $\mathbf{z}_{n\times 1}$, $\mathbf{V}_{n\times 3}$, $\frac{\partial\ell(\boldsymbol{\theta})}{\partial \mathbf{y}'}_{1\times n}$, $\boldsymbol{\varphi}'\left(\boldsymbol{\theta}\right)_{1\times 3}$, $\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial \mathbf{y}\partial\boldsymbol{\theta}'}_{n\times 3}$.

solving $P\left(|R^*| < z_{\alpha/2}\right) = 1-\alpha$ where z_α is the α quartile of the standard normal distribution, or calculating

$$\left(\min\left\{p^{-1}\left(\frac{\alpha}{2}\right), p^{-1}\left(1-\frac{\alpha}{2}\right)\right\}, \max\left\{p^{-1}\left(\frac{\alpha}{2}\right), p^{-1}\left(1-\frac{\alpha}{2}\right)\right\}\right) . \quad (4.1.12)$$

4.2 Simulations for One Sample Sharpe Ratio under AR(1) Return

4.2.1 Reference Group of Existing Methodology

In order to illustrate the exceptional accuracy of our proposed method, we construct the followings as our reference group of methodology.

1. Lo (2002) proposed that, for non-IID returns, the distribution of estimated Sharpe ratio can be derived by using MLE plus Delta method:

$$\widehat{SR} = \psi\left(\hat{\boldsymbol{\theta}}\right) \xrightarrow{d} N\left(\psi\left(\boldsymbol{\theta}\right), \left(\left.\frac{\partial\psi\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}}\right)' \mathbf{I}^{-1}\left(\boldsymbol{\theta}\right)\left.\frac{\partial\psi\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}}\right) ; \quad (4.2.1)$$

2. Or we can replace the Fisher expected information matrix from the result above with the observed information matrix evaluated at $\hat{\boldsymbol{\theta}}$:

$$\widehat{SR} = \psi\left(\hat{\boldsymbol{\theta}}\right) \xrightarrow{d} N\left(\psi\left(\boldsymbol{\theta}\right), \left(\left.\frac{\partial\psi\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}}\right)' \mathbf{j}^{-1}\left(\hat{\boldsymbol{\theta}}\right)\left.\frac{\partial\psi\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}}\right) ; \quad (4.2.2)$$

3. The signed log-likelihood ratio statistic in (4.2.1):

$$R(\psi) = \text{sgn}(\hat{\psi} - \psi)\sqrt{2(l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_\psi))} ;$$

Results 1, 2 and 3 are all first order approximation $O(n^{-\frac{1}{2}})$, while our proposed likelihood method is third order approximation $O(n^{-\frac{3}{2}})$, indicating that theoretically our proposed likelihood method is more valid and accurate than the above members in reference group.

4. In addition, Van Belle (2002) noted a special rule of thumb, that is, under formulation of AR(1) with ρ being the autocorrelation of the series of returns and $\mu = r_f$, the noncentral t statistic becomes a central t statistic and

$$\sqrt{n}\widehat{SR} = t_{n-1} \xrightarrow{d} N\left(0, \frac{1+\rho}{1-\rho}\right) .$$

Since this rule of thumb does not contain any theoretical background, thus our numerical study of next round can illustrate its accuracy.

4.2.2 Numerical Study

In this part we provide a simulation study to assess the performance of our third order likelihood method relative to the existing methodology in reference group. For some combinations of $n = 26, 52$, $\mu = -1, 0, 1$, $\sigma^2 = 1$, $\rho = -0.5, 0.5$ and $r_f = 0$, ten thousand Monte Carlo replications are performed from a AR(1) process with parameters $\boldsymbol{\theta} = \begin{pmatrix} \rho \\ \mu \\ \sigma^2 \end{pmatrix}$. And for each generated sample, the 95% confidence interval for the difference of Sharpe ratio is calculated. The performance of a method is evaluated by the same criteria 1-6 in (3.1.2.2).

The simulated coverage probabilities, coverage error, upper and lower error probabilities and average biases and degree of symmetry are recorded in Table (4.1). We can conclude from the simulation that the proposed modified signed log likelihood ratio method gives excellent results and outperforms the other methods.

- The performance of the methodology in the reference group are not satisfactory. In particular, Both of the Lo's methods with expected information and with observed information share very similar simulation result. Method of Van Belle outputs worse results than Lo's method. Yet it is more than that. Van Belle's method only functions when $\mu = r_f$, therefore very limited. In general, our proposed method performed extremely well in the criteria considered in this section.

- Sample Size Effect on Average Bias: The performance of each methods are supposed to improve as sample size rises. In order to make this size effect visible, a second round of simulation is being conducted, setting $\mu = 0$, $\sigma^2 = 1$, $\rho = 0.5$ and $\mu = 0$, $\sigma^2 = 1$, $\rho = -0.5$, respectively. Each sample size rises from $n = 6$ to $n = 100$ and the results are recorded in Figure (4.1). Suppose we take 3 units of standard deviation as an acceptance level on AB, we can find that our proposed likelihood method result can achieve this level even for extreme sample size $n = 8$, while the reference group may need a sample size of over 100 to achieve the same accuracy when $\rho > 0$.

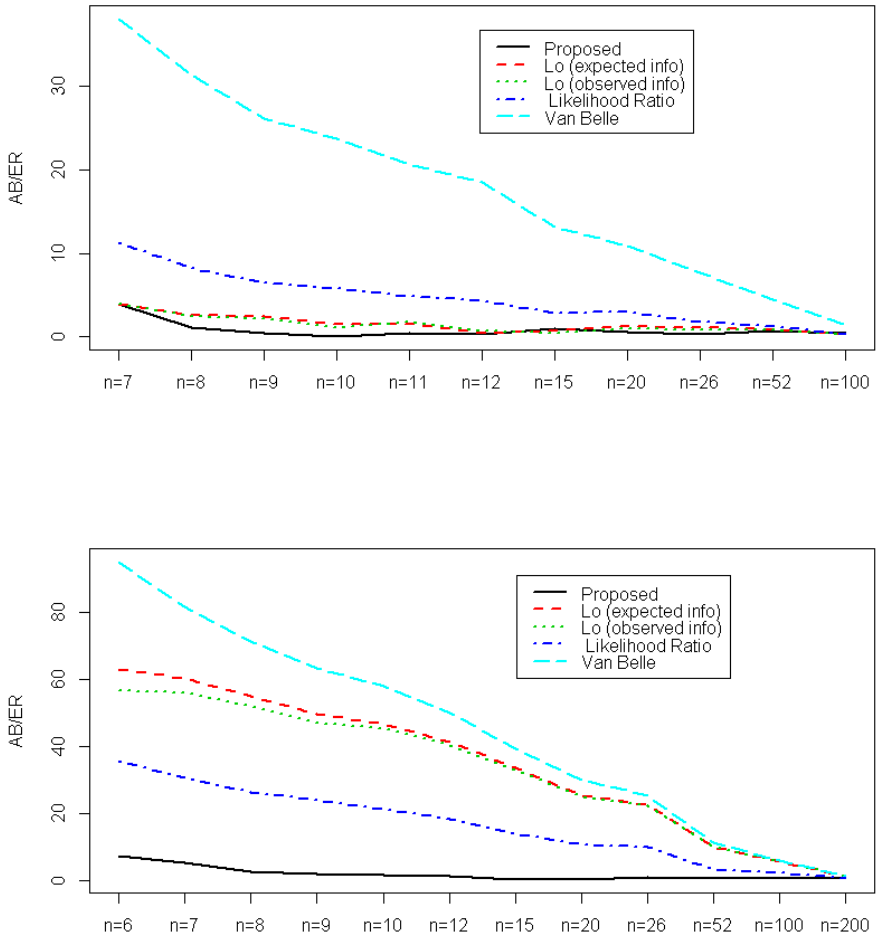
- The Effect of ρ and ψ: Figure (4.1) also shows that when $\rho = -0.5$ the simulation results of reference group improve a lot compared with $\rho = 0.5$. We conduct another round of simulation to reveal the effect of ρ and ψ on the accuracy of our methods, and the results are recorded in Figure(4.2) and Figure (4.3). We can conclude that:

1. The proposed likelihood method performs constantly well over the whole domain of $\psi = SR$. Meanwhile, the reference group does not give constant performance and it gives poorer results when the absolute value of ψ is large.

Table 4.1: Simulation Result for Difference of Sharpe Ratio under Bivariate Normal Return

Setting	Method	CP	LE	UE	AB	AB/ER	SY
$n = 52$, $\mu = 0$, $\sigma^2 = 1$, $\rho = 0.5$	Lo(exp)	0.9172	0.0414	0.0414	0.0164	10.25	1.00
	Lo(obs)	0.9172	0.0415	0.0413	0.0164	10.25	1.00
	Likelihood Ratio	0.9389	0.0301	0.0310	0.0056	3.47	1.03
	Van Belle	0.9129	0.0433	0.0438	0.0186	11.59	1.01
	Proposed	**0.9522**	**0.0238**	**0.0240**	**0.0011**	**0.69**	**1.01**
$n = 52$, $\mu = 0$, $\sigma^2 = 1$, $\rho = -0.5$	Lo(exp)	0.9471	0.0253	0.0276	0.0015	0.91	1.09
	Lo(obs)	0.9472	0.0253	0.0275	0.0014	0.87	1.09
	Likelihood Ratio	0.9458	0.0259	0.0283	0.0021	1.31	1.09
	Van Belle	0.9355	0.0316	0.0329	0.0073	4.53	1.04
	Proposed	**0.9492**	**0.0243**	**0.0265**	**0.0011**	**0.69**	**1.09**
$n = 52$, $\mu = 1$, $\sigma^2 = 1$, $\rho = 0.5$	Lo(exp)	0.9226	0.0527	0.0247	0.0140	8.75	2.13
	Lo(obs)	0.9235	0.0527	0.0238	0.0145	9.03	2.21
	Likelihood Ratio	0.9398	0.0364	0.0238	0.0063	3.94	1.53
	Proposed	**0.9481**	**0.0255**	**0.0264**	**0.0009**	**0.59**	**1.04**
$n = 52$, $\mu = 1$, $\sigma^2 = 1$, $\rho = -0.5$	Lo(exp)	0.9515	0.0274	0.0211	0.0032	1.97	1.30
	Lo(obs)	0.9518	0.0274	0.0208	0.0033	2.06	1.32
	Likelihood Ratio	0.9534	0.0282	0.0184	0.0049	3.06	1.53
	Proposed	**0.9467**	**0.0261**	**0.0272**	**0.0017**	**1.03**	**1.04**
$n = 52$, $\mu = -1$, $\sigma^2 = 1$, $\rho = 0.5$	Lo(exp)	0.9216	0.0232	0.0552	0.0160	10.00	2.38
	Lo(obs)	0.9225	0.0230	0.0545	0.0158	9.84	2.37
	Likelihood Ratio	0.9382	0.0228	0.0390	0.0081	5.06	1.71
	Proposed	**0.9497**	**0.0248**	**0.0255**	**0.0004**	**0.22**	**1.03**
$n = 52$, $\mu = -1$, $\sigma^2 = 1$, $\rho = -0.5$	Lo(exp)	0.9531	0.0202	0.0267	0.0033	2.03	1.32
	Lo(obs)	0.9528	0.0205	0.0267	0.0031	1.94	1.30
	Likelihood Ratio	0.9544	0.0178	0.0278	0.0050	3.13	1.56
	Proposed	**0.9476**	**0.0256**	**0.0268**	**0.0012**	**0.75**	**1.05**
$n = 26$, $\mu = 0$, $\sigma^2 = 1$, $\rho = 0.5$	Lo(exp)	0.8769	0.0578	0.0653	0.0366	22.84	1.13
	Lo(obs)	0.8780	0.0578	0.0642	0.0360	22.50	1.11
	Likelihood Ratio	0.9174	0.0383	0.0443	0.0163	10.19	1.16
	VB	0.8690	0.0613	0.0697	0.0405	25.31	1.14
	Proposed	**0.9462**	**0.0249**	**0.0289**	**0.0020**	**1.25**	**1.16**
$n = 26$, $\mu = 0$, $\sigma^2 = 1$, $\rho = -0.5$	Lo(exp)	0.9465	0.0278	0.0257	0.0018	1.09	1.08
	Lo(obs)	0.9474	0.0274	0.0252	0.0013	0.81	1.09
	Likelihood Ratio	0.9443	0.0287	0.0270	0.0029	1.78	1.06
	VB	0.9256	0.0383	0.0361	0.0122	7.63	1.06
	Proposed	**0.9489**	**0.0258**	**0.0253**	**0.0005**	**0.34**	**1.02**
$n = 26$, $\mu = 1$, $\sigma^2 = 1$, $\rho = 0.5$	Lo(exp)	0.8964	0.0724	0.0312	0.0268	16.75	2.32
	Lo(obs)	0.8966	0.0724	0.0310	0.0267	16.69	2.34
	Likelihood Ratio	0.9264	0.0476	0.0260	0.0118	7.38	1.83
	Proposed	**0.9497**	**0.0256**	**0.0247**	**0.0005**	**0.28**	**1.04**
$n = 26$, $\mu = 1$, $\sigma^2 = 1$, $\rho = -0.5$	Lo(exp)	0.9526	0.0260	0.0214	0.0023	1.44	1.21
	Lo(obs)	0.9533	0.0264	0.0203	0.0031	1.91	1.30
	Likelihood Ratio	0.9568	0.0260	0.0172	0.0044	2.75	1.51
	Proposed	**0.9496**	**0.0243**	**0.0261**	**0.0009**	**0.56**	**1.07**

Figure 4.1: The Effect of Sample Size on AB/ER under AR(1) Return (Up: $\rho = -0.5$; Down: $\rho = 0.5$)



2. Figure (4.2) also shows the results of reference group performs worse when $\rho = 0.5$ than the case when $\rho = -0.5$, and this can be illustrated clearer at Figure (4.3). Figure (4.3) recorded the results of simulation where 10,000 random samples are generated, each having a sample size of 26 from a AR(1) process with parameters (ρ, μ, σ^2) being $(\rho, 0, 1)$. ρ is taking a list of values from -0.9 to +0.9. We can conclude that, different from other compared method, our proposed method is doing constantly well over the whole region of $\rho \in (-1, 1)$.

4.3 Examples for One Sample Sharpe Ratio under AR(1) Return

In this subsection, we will provide empirical examples for inference on Sharpe ratio under AR(1) return. The data used are listed at Table (4.2) and are coming from David Ruppert(2004, page113). The 40 sample series represent the daily closing prices and returns for GE common stock on the January 2000 and February 2000.[6] Our proposed likelihood methodology is based on returns which follows an autoregressive process of order one, and David Ruppert(2004 page124) had tested the validity of this assumption on this data set. Finally we will use the average daily return of 3-Month Treasury Bill during the above period as the risk-free rate r_f, and their value are 0.000145712 for January and 0.000152 for February.[7]

Although a general simulation has been performed at last subsection, here we do another round of simulation under the setting of our example in order to validate our statistical inference. By maximum likelihood method, we obtain the MLE $\hat{\boldsymbol{\theta}} = \begin{pmatrix} \hat{\rho} \\ \hat{\mu} \\ \hat{\sigma^2} \end{pmatrix} = \begin{pmatrix} 0.2802319652 \\ -0.0074951041 \\ 0.0002741755 \end{pmatrix}$ for the January GE returns. Then, we can mimic the January GE return data with an simulated AR(1) process with population parameter $\rho = 0.2802319652$, $\mu = -0.0074951041$, $\sigma^2 = 0.0002741755$, $r_f = 0.000145712$ and $n = 20$. Doing this in the same way, we can also mimic the February GE return data with another AR(1) process with $\rho = 0.1741995144$, $\mu = -0.0005866871$, $\sigma^2 = 0.0002432038$, $r_f = 0.000152$ and $n = 20$. 90%, 95%, and 99% confidence intervals for the Sharpe ratio were obtained for each sample. Table (4.3) and Table (4.4) report the results from 10000 simulations for the January and February returns, respectively. From these results tables it is clear that, again, our proposed method outperforms the other methods based on the criteria we examined.

[6] It is only a part of the data set of David Ruppert (2004) who make use of a larger sample for GE from December 1999 to December 2000

[7] Source: Board of Governors of the Federal Reserve System (US), 3-Month Treasury Bill: Secondary Market Rate [DTB3], retrieved from FRED, Federal Reserve Bank of St. Louis https://research.stlouisfed.org/fred2/series/DTB3, May 3, 2016.

Figure 4.2: The Central Effect on AB/ER (Up: $\rho = -0.5$; Down: $\rho = 0.5$)

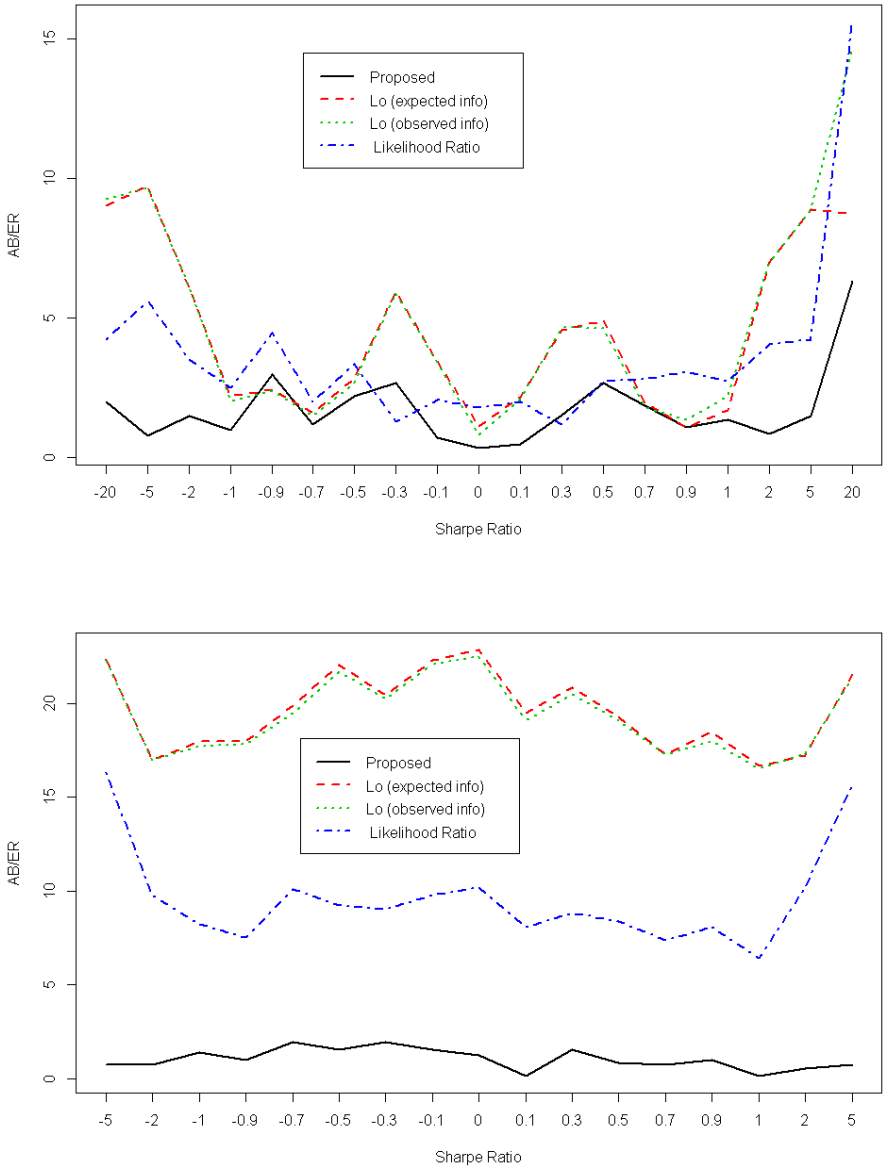


Figure 4.3: The Effect of ρ on AB/ER when $\psi = 0$

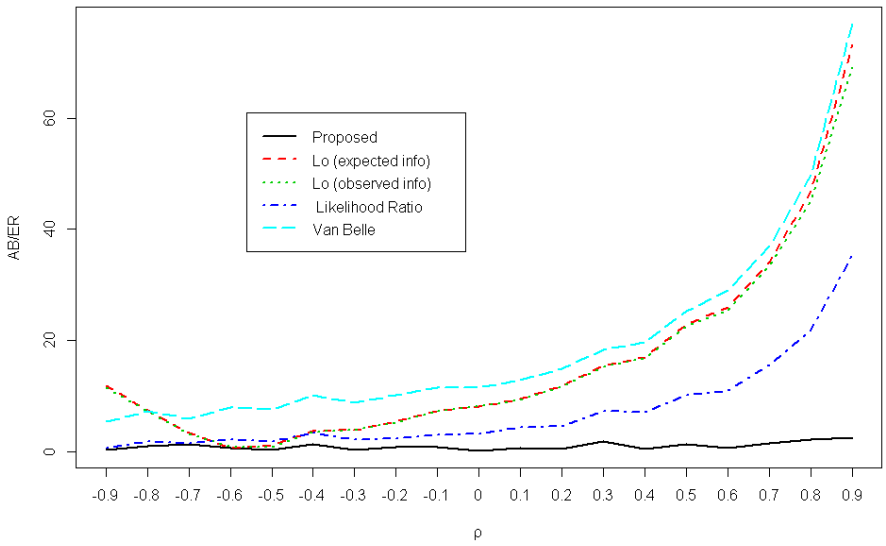


Table 4.2: GE daily closing prices and daily return

Date	Closing prices	Return	Date	Closing prices	Return
1/3/2000	50.4792	-0.02143497	2/1/2000	45.1667	0.007408923
1/4/2000	48.6771	-0.036352677	2/2/2000	45.2812	0.002531846
1/5/2000	48.2604	-0.008597345	2/3/2000	45.8438	0.012348031
1/6/2000	48.2604	0	2/4/2000	47.2708	0.030652803
1/7/2000	49.3854	0.023043485	2/7/2000	46.2708	-0.021381676
1/10/2000	50.8646	0.029512366	2/8/2000	45.8229	-0.009727126
1/11/2000	50.5521	-0.006162713	2/9/2000	45.2917	-0.011660173
1/12/2000	50.6354	0.001646449	2/10/2000	45.0104	-0.006230219
1/13/2000	51.3229	0.01348611	2/11/2000	45.1458	0.003003678
1/14/2000	50.6979	-0.012252557	2/14/2000	44.9167	-0.005087589
1/18/2000	49.3958	-0.026019089	2/15/2000	45.4896	0.012674065
1/19/2000	49.5312	0.002737374	2/16/2000	45.2188	-0.005970799
1/20/2000	48.7292	-0.016324334	2/17/2000	44.2708	-0.021187612
1/21/2000	48.6979	-0.000642532	2/18/2000	42.8125	-0.033495201
1/24/2000	47.0625	-0.034159402	2/22/2000	42.5104	-0.007081364
1/25/2000	46.2292	-0.017864872	2/23/2000	43.5521	0.024209171
1/26/2000	46.8438	0.01320703	2/24/2000	43.1667	-0.008888558
1/27/2000	46.4688	-0.008037543	2/25/2000	42.7708	-0.009213738
1/28/2000	45.6875	-0.016956382	2/28/2000	43.0417	0.006313786
1/31/2000	44.8333	-0.018873571	2/29/2000	44.0208	0.022492836

Table 4.3: Simulation Result for Sharpe Ratio under AR(1) January GE Return

CI	Method	CP	LE	UE	AB	AB/SE	SY
95%	Lo(exp)	0.9402	0.0108	0.0490	0.0191	11.94	4.54
	Lo(obs)	0.9404	0.0108	0.0488	0.0190	11.88	4.52
	Likelihood Ratio	0.9471	0.0117	0.0412	0.0148	9.22	3.52
	Proposed	**0.9494**	**0.0270**	**0.0236**	**0.0017**	**1.06**	**1.14**
90%	Lo(exp)	0.8800	0.0254	0.0946	0.0346	15.73	3.72
	Lo(obs)	0.8807	0.0244	0.0949	0.0353	16.02	3.89
	Likelihood Ratio	0.8864	0.0259	0.0877	0.0309	14.05	3.39
	Proposed	**0.8965**	**0.0508**	**0.0527**	**0.0017**	**0.80**	**1.04**
99%	Lo(exp)	0.9884	0.0020	0.0096	0.0038	5.43	4.80
	Lo(obs)	0.9888	0.0018	0.0094	0.0038	5.43	5.22
	Likelihood Ratio	0.9908	0.0021	0.0071	0.0025	3.57	3.38
	Proposed	**0.9910**	**0.0058**	**0.0032**	**0.0013**	**1.86**	**1.81**

Table 4.4: Simulation Result for Sharpe Ratio under AR(1) February GE Return

Setting	Method	CP	LE	UE	AB	AB/SE	SY
95%	Lo(exp)	0.9480	0.0142	0.0378	0.0118	7.38	2.66
	Lo(obs)	0.9488	0.0139	0.0373	0.0117	7.31	2.68
	Likelihood Ratio	0.9532	0.0135	0.0333	0.0099	6.19	2.47
	Proposed	**0.9545**	**0.0269**	**0.0186**	**0.0042**	**2.59**	**1.45**
90%	Lo(exp)	0.8941	0.0281	0.0778	0.0249	11.30	2.77
	Lo(obs)	0.8945	0.0277	0.0778	0.0251	11.39	2.81
	Likelihood Ratio	0.8996	0.0271	0.0733	0.0231	10.50	2.70
	Proposed	**0.9035**	**0.0526**	**0.0439**	**0.0044**	**1.98**	**1.20**
99%	Lo(exp)	0.9908	0.0032	0.0060	0.0014	2.00	1.88
	Lo(obs)	0.9911	0.0028	0.0061	0.0017	2.36	2.18
	Likelihood Ratio	0.9930	0.0022	0.0048	0.0015	2.14	2.18
	Proposed	**0.9907**	**0.0074**	**0.0019**	**0.0028**	**3.93**	**3.89**

Our first application focuses on confidence intervals for Sharpe ratio. Table (4.5) reports 95% confidence intervals for Sharpe ratio separately for the January GE returns and February GE return. We can find that the confidence intervals obtained from the five methods produce different results and our proposed method should give more accurate confidence interval compared with the reference methodology. This is because, theoretically, the proposed method has third-order accuracy whereas the remaining three methods do not and this fact is also borne out in the simulations.

The p-value functions calculated from each methods are plotted in Figures (4.4). These significance functions can be used to obtain p-values for specific hypothesized values of the Sharpe ratio, which are shown in Table (4.6) and (4.7). We can see that the p-values vary across the methods and people can result in different conclusions from the application of difference methods. For example, for GE January return data, we want to test:

$$H_0 : SR \leq 0 \text{ versus } H_1 : SR > 0$$

The corresponding p-values for our proposed method is 0.1028, which can not reject the null at 90% significance level, while all the other methods reject the null at 90% significance level. As we are typically interested in tail probabilities which tend to be very small, it is important to estimate such probabilities with precision, and our proposed method is believable to give more convincing value.

4.4 Likelihood Methodology for Two Independent Sample Comparison of Sharpe Ratio under AR(1) Return

Suppose one is interested in testing hypotheses concerning the comparison of Sharpe ratio of two independent funds 1 and 2. For instance, one may be interested in testing the null hypothesis $SR_1 \geq SR_2$ against $SR_1 < SR_2$; or, testing the null $SR_1 = SR_2$ against the alternative hypothesis $SR_1 \neq SR_2$. In fact, this is more useful and practical to know, because in most applicable conditions people need to identify their preferable investment from two or more available strategies at hand. In this subsection we apply the third order methodology introduced in Chapter 2 to test the difference of Sharpe ratio of independent funds 1 and 2 when the underlying returns are following autoregressive process of order one.

Consider two independent funds' log-returns, $\mathbf{r}_1$ and $\mathbf{r}_2$. They are following two separate stationary AR(1) process.

Table 4.5: 95% Confidence Intervals for Sharpe Ratio under January and February GE Return

Method	CI for SR of January GE Returns	CI for SR of February GE Returns
Lo(exp)	(-1.0352, 0.1492)	(-0.5640, 0.4707)
Lo(obs)	(-1.0326, 0.1467)	(-0.5683, 0.4750)
Likelihood Ratio	(-1.0748, 0.2044)	(-0.5868, 0.5573)
Proposed	**(-1.0690, 0.2801)**	**(-0.6043, 0.5953)**

Table 4.6: p-values for Sharpe Ratio under AR(1) GE January Return

ψ	-1.2	-1.1	-1	-0.9	-0.8	-0.5	-0.3	-0.1
Lo(exp)	0.9939	0.9852	0.9674	0.9348	0.8813	0.5749	0.3181	0.1282
Lo(obs)	0.9941	0.9855	0.9680	0.9356	0.8823	0.5752	0.3173	0.1271
Likelihood Ratio	0.9890	0.9787	0.9600	0.9279	0.8763	0.5751	0.3180	0.1335
Proposed	**0.9890**	**0.9793**	**0.9621**	**0.9327**	**0.8856**	**0.6059**	**0.3554**	**0.1631**

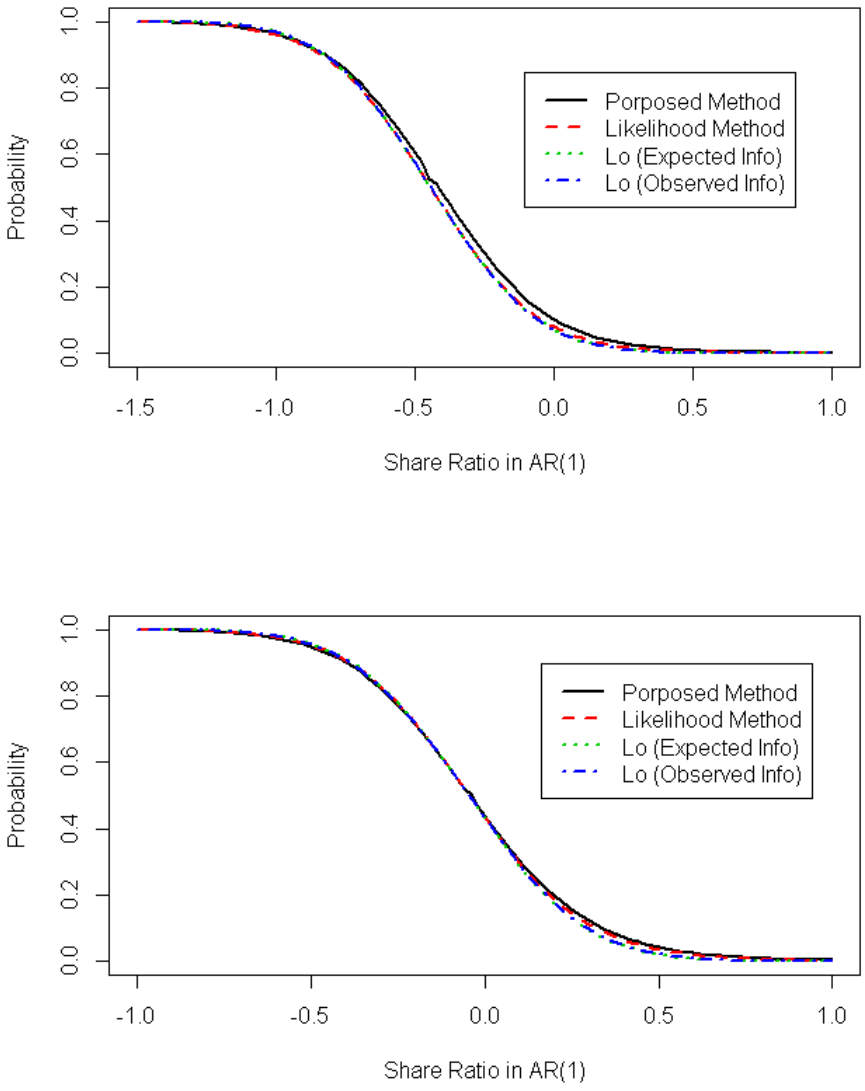
ψ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
Lo(exp)	0.0713	0.0362	0.0167	0.0070	0.0026	0.0009	0.0003	0.0001
Lo(obs)	0.0705	0.0356	0.0163	0.0068	0.0025	0.0009	0.0003	0.0001
Likelihood Ratio	0.0796	0.0457	0.0257	0.0143	0.0080	0.0045	0.0026	0.0016
Proposed	**0.1028**	**0.0628**	**0.0377**	**0.0226**	**0.0137**	**0.0084**	**0.0054**	**0.0036**

Table 4.7: p-values for Sharpe Ratio under AR(1) GE February Return

ψ	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1
Lo(exp)	0.9978	0.9933	0.9820	0.9570	0.9096	0.8314	0.7194	0.5801
Lo(obs)	0.9977	0.9929	0.9812	0.9557	0.9078	0.8294	0.7177	0.5794
Likelihood Ratio	0.9953	0.9896	0.9774	0.9527	0.9069	0.8307	0.7194	0.5798
Proposed	**0.9941**	**0.9876**	**0.9742**	**0.9481**	**0.9009**	**0.8243**	**0.7145**	**0.5790**

ψ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
Lo(exp)	0.4299	0.2893	0.1751	0.0946	0.0453	0.0192	0.0072	0.0023
Lo(obs)	0.4304	0.2908	0.1771	0.0964	0.0467	0.0200	0.0076	0.0025
Likelihood Ratio	0.4308	0.2950	0.1875	0.1121	0.0640	0.0354	0.0193	0.0104
Proposed	**0.4351**	**0.3032**	**0.1978**	**0.1223**	**0.0728**	**0.0423**	**0.0244**	**0.0141**

Figure 4.4: p-value function for Sharpe Ratio (Up: January; Down: February)



$$\mathbf{r}_{1(2)} = \mu_{1(2)} \cdot \mathbf{1}_{1(2)} + \sigma_{1(2)} \cdot \boldsymbol{\epsilon}_{1(2)} = \mu_{1(2)} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \sigma_{1(2)} \begin{pmatrix} \epsilon_{1(2),1} \\ \vdots \\ \epsilon_{1(2),T_1(T_2)} \end{pmatrix} ,$$

$$\epsilon_{1(2)} \sim N \left(\mathbf{0}, \boldsymbol{\Omega}_{1(2)} = \left(\frac{\rho_{1(2)}^{|i-j|}}{1-\rho_{1(2)}^2} \right)_{ij} \right)$$

$$\text{or } \sigma_{1(2)}\epsilon_{1(2)} \sim N \left(\mathbf{0}, \boldsymbol{\Sigma}_{1(2)} = \sigma_{1(2)}^2 \cdot \boldsymbol{\Omega}_{1(2)} = \left(\frac{\sigma_{1(2)}^2 \rho_{1(2)}^{|i-j|}}{1-\rho_{1(2)}^2} \right)_{ij} \right)$$

The inverse matrix of $\boldsymbol{\Omega}_{1(2)}$, its Cholesky decomposition and its derivative matrix are all same as the results in last subsection.

Since our interest is to test the difference in Sharpe ratio, the parameter of interest ψ and its first order derivative are then defined as follows:

$$\psi(\boldsymbol{\theta}) = \psi \begin{pmatrix} \rho_1 \\ \mu_1 \\ \sigma_1^2 \\ \rho_2 \\ \mu_2 \\ \sigma_2^2 \end{pmatrix} = \frac{\mu_1 - \mu_{f1}}{\sqrt{\frac{\sigma_1^2}{1-\rho_1^2}}} - \frac{\mu_2 - \mu_{f2}}{\sqrt{\frac{\sigma_2^2}{1-\rho_2^2}}} ,$$

$$\psi_{\boldsymbol{\theta}'}(\boldsymbol{\theta}) = \left(\psi_{\rho_1}(\boldsymbol{\theta}), \psi_{\mu_1}(\boldsymbol{\theta}), \psi_{\sigma_1^2}(\boldsymbol{\theta}), \psi_{\rho_2}(\boldsymbol{\theta}), \psi_{\mu_2}(\boldsymbol{\theta}), \psi_{\sigma_2^2}(\boldsymbol{\theta}) \right) .$$

For fund 1 (or 2), its individual log likelihood function is

$$l_{1(2)} = a - \frac{n_{1(2)}}{2} \log \sigma_{1(2)}^2 + \frac{1}{2} \log \left(1 - \rho_{1(2)}^2 \right) - \frac{1}{2\sigma_{1(2)}^2} \left(\mathbf{r}_{1(2)} - \mu_{1(2)} \cdot \mathbf{1}_{1(2)} \right)' \mathbf{A}_{1(2)} \left(\mathbf{r}_{1(2)} - \mu_{1(2)} \cdot \mathbf{1}_{1(2)} \right)$$

Then, the unrestricted maximum likelihood estimation maximizes the likelihood of both funds' sample or the joint log likelihood function $l(\boldsymbol{\theta})$:

$$\begin{aligned} l(\boldsymbol{\theta}) &= l_1 + l_2 = \\ & a + b - \frac{n_1}{2} \log \sigma_1^2 + \frac{1}{2} \log \left(1 - \rho_1^2 \right) - \frac{1}{2\sigma_1^2} \left(\mathbf{r}_1 - \mu_1 \cdot \mathbf{1}_1 \right)' \mathbf{A}_1 \left(\mathbf{r}_1 - \mu_1 \cdot \mathbf{1}_1 \right) \\ & - \frac{n_2}{2} \log \sigma_2^2 + \frac{1}{2} \log \left(1 - \rho_2^2 \right) - \frac{1}{2\sigma_2^2} \left(\mathbf{r}_2 - \mu_2 \cdot \mathbf{1}_2 \right)' \mathbf{A}_2 \left(\mathbf{r}_2 - \mu_2 \cdot \mathbf{1}_2 \right) . \end{aligned} \quad (4.4.1)$$

Since the two funds are independent, the overall unrestricted MLE actually is the combination of separate unrestricted MLE. That is, if $\hat{\rho}_1$, $\hat{\mu}_1$, $\hat{\sigma}_1^2$ is the unrestricted MLE for ℓ_1, in addition, $\hat{\rho}_2$, $\hat{\mu}_2$, $\hat{\sigma}_2^2$ is the unrestricted MLE for ℓ_2, then the unrestricted MLE for $\ell(\boldsymbol{\theta})$ is

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} \hat{\rho}_1 \\ \hat{\mu}_1 \\ \hat{\sigma}_1^2 \\ \hat{\rho}_2 \\ \hat{\mu}_2 \\ \hat{\sigma}_2^2 \end{pmatrix}.$$

Given this information about the overall maximum likelihood estimate, we can obtain other important variable for future use, such as the estimated parameter of interest $\hat{\psi} = \frac{\hat{\mu}_1 - \mu_{f1}}{\sqrt{\frac{\hat{\sigma}_1^2}{1-\hat{\rho}_1^2}}} - \frac{\hat{\mu}_2 - \mu_{f2}}{\sqrt{\frac{\hat{\sigma}_2^2}{1-\hat{\rho}_2^2}}}$, the estimated unrestricted likelihood function $\ell\left(\hat{\boldsymbol{\theta}}\right)$. Also, because of the independence of samples, the observed information matrix evaluated at $\hat{\boldsymbol{\theta}}$ can be obtained by augmenting the individual observed information matrix. In particular, if $\mathbf{j}_1$ is the observed information matrix for ℓ_1 and $\mathbf{j}_2$ is the observed information matrix for ℓ_2, then we can obtain the following [8]

$$\mathbf{j}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} \mathbf{j}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{j}_2 \end{pmatrix} =$$

$$- \begin{pmatrix} \ell_{\rho_1\rho_1}\left(\hat{\boldsymbol{\theta}}\right) & \ell_{\rho_1\mu_1}\left(\hat{\boldsymbol{\theta}}\right) & \ell_{\rho_1\sigma_1^2}\left(\hat{\boldsymbol{\theta}}\right) & & & \\ \ell_{\mu_1\rho_1}\left(\hat{\boldsymbol{\theta}}\right) & l_{\mu_1\mu_1}(\hat{\boldsymbol{\theta}}) & l_{\mu_1\sigma_1^2}(\hat{\boldsymbol{\theta}}) & & \mathbf{0} & \\ \ell_{\sigma_1^2\rho_1}\left(\hat{\boldsymbol{\theta}}\right) & l_{\sigma_1^2\mu_1}(\hat{\boldsymbol{\theta}}) & l_{\sigma_1^2\sigma_1^2}(\hat{\boldsymbol{\theta}}) & & & \\ & & & \ell_{\rho_2\rho_2}\left(\hat{\boldsymbol{\theta}}\right) & \ell_{\rho_2\mu_2}\left(\hat{\boldsymbol{\theta}}\right) & \ell_{\rho_2\sigma_2^2}\left(\hat{\boldsymbol{\theta}}\right) \\ & \mathbf{0} & & \ell_{\mu_2\rho_2}\left(\hat{\boldsymbol{\theta}}\right) & l_{\mu_2\mu_2}(\hat{\boldsymbol{\theta}}) & l_{\mu_2\sigma_2^2}(\hat{\boldsymbol{\theta}}) \\ & & & \ell_{\sigma_2^2\rho_2}\left(\hat{\boldsymbol{\theta}}\right) & l_{\sigma_2^2\mu_2}(\hat{\boldsymbol{\theta}}) & l_{\sigma_2^2\sigma_2^2}(\hat{\boldsymbol{\theta}}) \end{pmatrix}.$$

After processing unconstrained MLE, we start to derive the constrained MLE. The log-likelihood function $\ell(\boldsymbol{\theta})$ must be maximized with respect to ρ_1, μ_1, σ_1^2,ρ_2, μ_2, σ_2^2 while holding ψ fixed and this process can be managed by the Lagrange multiplier method (see (2.3.7)). The Lagrangian function is given at (4.4.2) and its first order derivatives are listed from (4.4.3) to (4.4.9).

[8] We can actually obtain the expected Fisher full information matrix. Theoretically, the application of expected Fisher full information matrix instead of the observed information matrix evaluated at $\hat{\boldsymbol{\theta}}$ should increase the accuracy of our method, but since the improvement is not quite significant when we put into simulation, in our paper we will still use the observed information for our proposed method.

$$H(\boldsymbol{\theta},\alpha) = l(\boldsymbol{\theta}) + \alpha(\psi(\boldsymbol{\theta}) - \psi) = l(\boldsymbol{\theta}) + \alpha(\frac{\mu_1 - \mu_{f1}}{\sqrt{\frac{\sigma_1^2}{1-\rho_1^2}}} - \frac{\mu_2 - \mu_{f2}}{\sqrt{\frac{\sigma_2^2}{1-\rho_2^2}}} - \psi)\ , \quad (4.4.2)$$

$$H_{\rho_1}(\boldsymbol{\theta},\alpha) = l_{\rho_1}(\boldsymbol{\theta}) + \alpha\psi_{\rho_1}(\boldsymbol{\theta})\ , \quad (4.4.3)$$

$$H_{\mu_1}(\boldsymbol{\theta},\alpha) = l_{\mu_1}(\boldsymbol{\theta}) + \alpha\psi_{\mu_1}(\boldsymbol{\theta})\ , \quad (4.4.4)$$

$$H_{\sigma_1^2}(\boldsymbol{\theta},\alpha) = l_{\sigma_1^2}(\boldsymbol{\theta}) + \alpha\psi_{\sigma_1^2}(\boldsymbol{\theta})\ , \quad (4.4.5)$$

$$H_{\rho_2}(\boldsymbol{\theta},\alpha) = l_{\rho_2}(\boldsymbol{\theta}) + \alpha\psi_{\rho_2}(\boldsymbol{\theta})\ , \quad (4.4.6)$$

$$H_{\mu_2}(\boldsymbol{\theta},\alpha) = l_{\mu_2}(\boldsymbol{\theta}) + \alpha\psi_{\mu_2}(\boldsymbol{\theta})\ , \quad (4.4.7)$$

$$H_{\sigma_2^2}(\boldsymbol{\theta},\alpha) = l_{\sigma_2^2}(\boldsymbol{\theta}) + \alpha\psi_{\sigma_2^2}(\boldsymbol{\theta})\ , \quad (4.4.8)$$

$$H_{\alpha}(\boldsymbol{\theta},\alpha) = \frac{\mu_1 - \mu_{f1}}{\sqrt{\frac{\sigma_1^2}{1-\rho_1^2}}} - \frac{\mu_2 - \mu_{f2}}{\sqrt{\frac{\sigma_2^2}{1-\rho_2^2}}} - \psi\ . \quad (4.4.9)$$

Solving first order derivatives, from (4.4.3) to (4.4.9) equal to zero, we obtain the restricted maximum likelihood estimator, $\hat{\boldsymbol{\theta}}_{\psi}^{'} = (\tilde{\rho}_1, \tilde{\mu}_1, \tilde{\sigma}_1^2, \tilde{\rho}_2, \tilde{\mu}_2, \tilde{\sigma}_2^2)$. Notice that this process involves seven unknown variables- six constrained MLE $\hat{\boldsymbol{\theta}}_{\psi}^{'} = (\tilde{\rho}_1, \tilde{\mu}_1, \tilde{\sigma}_1^2, \tilde{\rho}_2, \tilde{\mu}_2, \tilde{\sigma}_2^2)$ plus the estimated Lagrange estimator $\hat{\alpha}$ and these seven are being solved by a nonlinear system of seven equations. Unfortunately, we cannot obtain closed form analytical solutions, however, numerical solutions can be got by implementing Newton's Method (Richard L. Burden and J. Douglas Faires, 2012) with the starting values of iteration being unconstrained MLE $\hat{\boldsymbol{\theta}}^{'} = (\hat{\rho}_1, \hat{\mu}_1, \hat{\sigma}_1^2, \hat{\rho}_2, \hat{\mu}_2, \hat{\sigma}_2^2)$. After that, the tilted log-likelihood function can be obtained by replacing α by $\hat{\alpha}$ on the Lagrangian function. Thus, we can also obtain the estimated restricted likelihood function, $l(\hat{\boldsymbol{\theta}}_{\psi})$, the tilted observed information matrix evaluated at $\hat{\boldsymbol{\theta}}_{\psi}$, $\tilde{\mathbf{j}}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_{\psi}) = -\frac{\partial^2 \tilde{\mathbf{j}}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\psi}} = j_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_{\psi}) - \hat{\alpha}\psi_{\boldsymbol{\theta}\boldsymbol{\theta}'}\left(\hat{\boldsymbol{\theta}}_{\psi}\right)$ and its inverse $\tilde{\mathbf{j}}^{-1}_{\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_{\psi})$.

Like one sample case, here the general canonical parameter for AR(1) is not available and we try to obtain a locally defined canonical parameter by the method of (2.6.15) and (2.6.16). First we need to organize our two independent AR(1) processes into a unified one.

$$\begin{aligned} \mathbf{r} &= \begin{pmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{pmatrix} = \begin{pmatrix} \mu_1 \cdot \mathbf{1}_1 \\ \mu_2 \cdot \mathbf{1}_2 \end{pmatrix} + \begin{pmatrix} \sigma_1 \cdot \boldsymbol{\epsilon}_1 \\ \sigma_2 \cdot \boldsymbol{\epsilon}_2 \end{pmatrix} \\ &\sim N\left(\begin{pmatrix} \mu_1 \cdot \mathbf{1}_1 \\ \mu_2 \cdot \mathbf{1}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 \end{pmatrix}\right)\ . \end{aligned}$$

Therefore, the pivotal quantity $\mathbf{z}$ for this problem is specified as the vector

of independent standard normal deviates:

$$\begin{aligned}\mathbf{z} &= \begin{pmatrix} \mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_2 \end{pmatrix}^{-\frac{1}{2}} \cdot \begin{pmatrix} \mathbf{r}_1 - \mu_1 \cdot \mathbf{1}_1 \\ \mathbf{r}_2 - \mu_2 \cdot \mathbf{1}_2 \end{pmatrix} \\ &= \begin{pmatrix} \frac{\mathbf{L}_1}{\sigma_1} & \mathbf{0} \\ \mathbf{0} & \frac{\mathbf{L}_2}{\sigma_2} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{r}_1 - \mu_1 \cdot \mathbf{1}_1 \\ \mathbf{r}_2 - \mu_2 \cdot \mathbf{1}_2 \end{pmatrix} \\ &= \begin{pmatrix} \frac{\mathbf{L}_1(\mathbf{r}_1 - \mu_1 \cdot \mathbf{1}_1)}{\sigma} \\ \frac{\mathbf{L}_2(\mathbf{r}_2 - \mu_2 \cdot \mathbf{1}_2)}{\sigma_2} \end{pmatrix}.\end{aligned}$$

This choice of pivotal quantity coincides with the standard quantity used to estimate the parameters of an AR(1) model in the literature (see for example Hamilton (1994)). Then, the ancillary direction array $\mathbf{V}$ can be constructed from (2.6.16)

$$\begin{aligned}\mathbf{V} &= -\left(\frac{\partial \mathbf{z}}{\partial \mathbf{r}'}\right)^{-1} \frac{\partial \mathbf{z}}{\partial \boldsymbol{\theta}'}\bigg|_{\hat{\boldsymbol{\theta}}} \\ &= -\begin{pmatrix} \frac{\mathbf{L}_1}{\sigma_1} & \mathbf{0} \\ \mathbf{0} & \frac{\mathbf{L}_2}{\sigma_2} \end{pmatrix}^{-1} \cdot \begin{pmatrix} \frac{\partial \mathbf{z}}{\partial \rho_1} & \frac{\partial \mathbf{z}}{\partial \mu_1} & \frac{\partial \mathbf{z}}{\partial \sigma_1^2} & \mathbf{0} \\ \mathbf{0} & & & \frac{\partial \mathbf{z}}{\partial \rho_2} & \frac{\partial \mathbf{z}}{\partial \mu_2} & \frac{\partial \mathbf{z}}{\partial \sigma_2^2} \end{pmatrix}\bigg|_{\hat{\boldsymbol{\theta}}} \\ &= \begin{pmatrix} \mathbf{V}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2 \end{pmatrix}.\end{aligned}$$

where $\mathbf{V}_1 = \begin{pmatrix} -\hat{\mathbf{L}}_1^{-1}\hat{\mathbf{L}}_{\rho_1}(\mathbf{r}_1 - \hat{\mu}_1 \cdot \mathbf{1}_1) & \mathbf{1}_1 & \frac{\mathbf{r}_1 - \hat{\mu}_1 \cdot \mathbf{1}_1}{2\hat{\sigma}_1^2} \end{pmatrix}$ as well as matrix $\mathbf{V}_2 = \begin{pmatrix} -\hat{\mathbf{L}}_2^{-1}\hat{\mathbf{L}}_{\rho_2}(\mathbf{r}_2 - \hat{\mu}_2 \cdot \mathbf{1}_2) & \mathbf{1}_2 & \frac{\mathbf{r}_2 - \hat{\mu}_2 \cdot \mathbf{1}_2}{2\hat{\sigma}_2^2} \end{pmatrix}$. Note that $\mathbf{V}$ is a matrix of sample return $\mathbf{r}$ and it is not related to $\boldsymbol{\theta}$. Finally, the new locally defined canonical parameter at the data $\mathbf{r}$ can be obtained by (2.6.15), given that the sample space gradient of the likelihood evaluated at the data is $\frac{\partial}{\partial \mathbf{r}'}\ell(\boldsymbol{\theta}) = \begin{pmatrix} -\frac{1}{\sigma_1^2}(\mathbf{r}_1 - \mu_1 \cdot \mathbf{1}_1)' \mathbf{A}_1, & -\frac{1}{\sigma_2^2}(\mathbf{r}_2 - \mu_2 \cdot \mathbf{1}_2)' \mathbf{A}_2 \end{pmatrix}$.

$$\begin{aligned}\boldsymbol{\varphi}'(\boldsymbol{\theta}) &= \begin{pmatrix} \varphi_1(\boldsymbol{\theta}) & \varphi_2(\boldsymbol{\theta}) & \varphi_3(\boldsymbol{\theta}) & \varphi_4(\boldsymbol{\theta}) & \varphi_5(\boldsymbol{\theta}) & \varphi_6(\boldsymbol{\theta}) \end{pmatrix} = \frac{\partial}{\partial \mathbf{r}'}\ell(\boldsymbol{\theta}) \cdot \mathbf{V} \\ &= \left(-\frac{1}{\sigma_1^2}(\mathbf{r}_1 - \mu_1 \cdot \mathbf{1}_1)' \mathbf{A}_1 \cdot \mathbf{V}_1, -\frac{1}{\sigma_2^2}(\mathbf{r}_2 - \mu_2 \cdot \mathbf{1}_2)' \mathbf{A}_2 \cdot \mathbf{V}_2\right),\end{aligned}$$

Or $\boldsymbol{\varphi}(\boldsymbol{\theta}) = \begin{pmatrix} \varphi_1(\boldsymbol{\theta}) \\ \varphi_2(\boldsymbol{\theta}) \\ \varphi_3(\boldsymbol{\theta}) \\ \varphi_4(\boldsymbol{\theta}) \\ \varphi_5(\boldsymbol{\theta}) \\ \varphi_6(\boldsymbol{\theta}) \end{pmatrix} = \mathbf{V}' \cdot \frac{\partial}{\partial \mathbf{r}}\ell(\boldsymbol{\theta})$. The first order derivative of canon-

ical parameter is

$$\begin{aligned}
\boldsymbol{\varphi}_{\boldsymbol{\theta}'}(\boldsymbol{\theta}) &= \mathbf{V}' \cdot \left(\frac{\partial^2}{\partial \mathbf{r} \partial \boldsymbol{\theta}'} \ell(\boldsymbol{\theta}) \right) \\
&= \mathbf{V}' \cdot \begin{pmatrix} \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \mathbf{r} \partial \rho_1} & \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \mathbf{r} \partial \mu_1} & \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \mathbf{r} \partial \sigma_1^2} & \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \mathbf{r} \partial \rho_2} & \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \mathbf{r} \partial \mu_2} & \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \mathbf{r} \partial \sigma_2^2} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{V}_1' & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2' \end{pmatrix} \cdot \\
&\quad \begin{pmatrix} \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \mathbf{r}_1 \partial \rho_1} & \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \mathbf{r}_1 \partial \mu_1} & \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \mathbf{r}_1 \partial \sigma_1^2} & & \mathbf{0} & \\ & \mathbf{0} & & \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \mathbf{r}_2 \partial \rho_2} & \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \mathbf{r}_2 \partial \mu_2} & \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \mathbf{r}_2 \partial \sigma_2^2} \end{pmatrix},
\end{aligned}$$

where $\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \mathbf{r}_{1(2)} \partial \rho_{1(2)}} = -\frac{1}{\sigma_{1(2)}^2} \mathbf{A}_{\rho_{1(2)}} \left(\mathbf{r}_{1(2)} - \mu_{1(2)} \cdot \mathbf{1}_{1(2)} \right)$, and also $\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \mathbf{r}_{1(2)} \partial \mu_{1(2)}} = \frac{1}{\sigma_{1(2)}^2} \mathbf{A}_{1(2)} \mathbf{1}_{1(2)}$, and $\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \mathbf{r}_{1(2)} \partial \sigma_{1(2)}^2} = \frac{1}{\sigma_{1(2)}^4} \mathbf{A}_{1(2)} \left(\mathbf{r}_{1(2)} - \mu_{1(2)} \cdot \mathbf{1}_{1(2)} \right)$. Before we precede, please note the dimension reduction here: the dimension is reduced from $n_1 + n_2$ (the dimension of $\mathbf{r}$) to 6 (the dimension of the parameter $\boldsymbol{\theta}$ evidenced from the expression for $\boldsymbol{\varphi}(\boldsymbol{\theta})$) [9]. To further reduce the dimension of the problem from 6 to the dimension of the parameter of interesting ψ, the calculation of newly calibrated parameter $\chi(\boldsymbol{\theta})$ which in turn involves the parameter vector $\boldsymbol{\varphi}(\boldsymbol{\theta})$ as well as the constrained MLE $\hat{\boldsymbol{\theta}}_\psi$ is required.

With the above information, the signed log-likelihood ratio statistic R can be constructed from (2.3.15), the newly calibrated parameter χ can be calculated from (2.6.17), the modified maximum likelihood departure measure Q can be obtained from (2.6.18) and (2.6.19), and finally, the proposed third order likelihood approximation Barndorff-Nielsen method $R^\star$ can be obtained from (2.6.13). A centered $(1-\alpha) \times 100\%$ confidence interval for ψ can be obtained by solving $P\left(|R^*| < z_{\alpha/2}\right) = 1 - \alpha$ where z_α is the α quartile of the standard normal distribution, or calculating

$$\left(\min\left\{ p^{-1}\left(\frac{\alpha}{2}\right), p^{-1}\left(1-\frac{\alpha}{2}\right) \right\}, \max\left\{ p^{-1}\left(\frac{\alpha}{2}\right), p^{-1}\left(1-\frac{\alpha}{2}\right) \right\} \right). \quad (4.4.10)$$

[9] The dimension of important variables are $\mathbf{z}_{(n_1+n_2)\times 1}$, $\mathbf{V}_{(n_1+n_2)\times 6}$, $\frac{\partial \ell(\boldsymbol{\theta})}{\partial \mathbf{r}'}_{1\times(n_1+n_2)}$, $\boldsymbol{\varphi}'(\boldsymbol{\theta})_{1\times 6}$, $\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \mathbf{r} \partial \boldsymbol{\theta}'}_{(n_1+n_2)\times 6}$.

4.5 Simulations for Two Independent Sample Comparison of Sharpe Ratio under AR(1) Return

4.5.1 Reference Group of Existing Methodology

In order to illustrate the exceptional accuracy of our proposed method, we construct the followings as our reference group of methodology. These existing methodology correspond to the methods in section (3.1.2.1).

1.Suppose we have two samples 1 and 2 that are independent with each other. According to Lo (2002), each sample will have its own Sharpe ratio's asymptotic distribution as (4.2.1). Therefore, the distribution of the difference of two estimated Sharpe ratio will be

$$\widehat{SR}_1 - \widehat{SR}_2 = \psi\left(\hat{\boldsymbol{\theta}}\right) \xrightarrow{d} N\Bigg(\psi\left(\boldsymbol{\theta}\right),$$
$$\left(\frac{\partial\psi_1\left(\boldsymbol{\theta}_1\right)}{\partial\boldsymbol{\theta}_1}\bigg|_{\hat{\boldsymbol{\theta}}_1}\right)' \mathbf{I}_1^{-1}\left(\boldsymbol{\theta}_1\right)\frac{\partial\psi_1\left(\boldsymbol{\theta}_1\right)}{\partial\boldsymbol{\theta}_1}\bigg|_{\hat{\boldsymbol{\theta}}_1} + \left(\frac{\partial\psi_2\left(\boldsymbol{\theta}_2\right)}{\partial\boldsymbol{\theta}_2}\bigg|_{\hat{\boldsymbol{\theta}}_2}\right)' \mathbf{I}_2^{-1}\left(\boldsymbol{\theta}_2\right)\frac{\partial\psi_2\left(\boldsymbol{\theta}_2\right)}{\partial\boldsymbol{\theta}_2}\bigg|_{\hat{\boldsymbol{\theta}}_2}\Bigg);$$

2. Like (4.2.2), we can replace the Fisher expected information matrix from the result above with the observed information matrix evaluated at $\hat{\boldsymbol{\theta}}$:

$$\widehat{SR}_1 - \widehat{SR}_2 = \psi\left(\hat{\boldsymbol{\theta}}\right) \xrightarrow{d} N\Bigg(\psi\left(\boldsymbol{\theta}\right),$$
$$\left(\frac{\partial\psi_1\left(\boldsymbol{\theta}_1\right)}{\partial\boldsymbol{\theta}_1}\bigg|_{\hat{\boldsymbol{\theta}}_1}\right)' \mathbf{j}_1^{-1}\left(\boldsymbol{\theta}_1\right)\frac{\partial\psi_1\left(\boldsymbol{\theta}_1\right)}{\partial\boldsymbol{\theta}_1}\bigg|_{\hat{\boldsymbol{\theta}}_1} + \left(\frac{\partial\psi_2\left(\boldsymbol{\theta}_2\right)}{\partial\boldsymbol{\theta}_2}\bigg|_{\hat{\boldsymbol{\theta}}_2}\right)' \mathbf{j}_2^{-1}\left(\boldsymbol{\theta}_2\right)\frac{\partial\psi_2\left(\boldsymbol{\theta}_2\right)}{\partial\boldsymbol{\theta}_2}\bigg|_{\hat{\boldsymbol{\theta}}_2}\Bigg);$$

3. The signed log-likelihood ratio statistic in (2.3.15):

$$R(\psi) = \mathbf{sgn}(\hat{\psi} - \psi)\sqrt{2(l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_\psi))}\,\mathbf{.}$$

Results 1, 2 and 3 are all first order approximation $O(n^{-\frac{1}{2}})$, while our proposed likelihood method is third order approximation $O(n^{-\frac{3}{2}})$, indicating that theoretically our proposed likelihood method is more valid and accurate than the above members in reference group.

4. In addition, a permissive derivative of Van Belle (2002)'s special rule of thumb, under $\mu = r_f$, is

$$\widehat{SR}_1 - \widehat{SR}_2 \xrightarrow{d} N\left(0, \frac{1}{n_1}\frac{1+\rho_1}{1-\rho_1} + \frac{1}{n_2}\frac{1+\rho_2}{1-\rho_2}\right).$$

Since this rule of thumb does not contain any theoretical background, thus our numerical study of next round can illustrate its accuracy.

4.5.2 Numerical Study

In this part we provide a simulation study to assess the performance of our third order likelihood method relative to the existing methodology in reference group. For some combinations of $n_1, n_2 = 20, 30$, $\mu = -1, 0, 1$, $\sigma^2 = 1$, $\rho = -0.2, 0.7$ and $r_f = 0$, ten thousand Monte Carlo replications are performed. And for each generated sample, the 95% confidence interval for the difference of Sharpe ratio is calculated. The performance of a method is judged using the same criteria 1-6 in (3.1.2.2). The simulated coverage probabilities, coverage error, upper and lower error probabilities and average biases and degree of symmetry are recorded in Table(4.8). We can conclude from the simulation that the proposed modified signed log likelihood ratio method gives excellent results and outperforms the other four methods based on the criteria we examined. We also have record of all the other simulation results besides our setting on the parameter and they all show the same excellent results of our proposed methods.

The analysis is very much similar with the analysis in (4.2). In particular, the performance of the methodology in the reference group are not satisfactory. Both of the Lo's methods with expected information and with observed information share very similar simulation result; besides method of Van Belle outputs worse results than Lo's method. In addition, Van Belle's method only functions when $\mu = r_f$, therefore this method is very limited. In general, our proposed method performed extremely well in the criteria considered in this section.

Table 4.8: Simulation Result for Difference of Sharpe Ratio under AR(1) Return

Setting	Method	CP	LE	UE	AB	AB/ER	SY
$n_1 = 20$, $\rho_1 = -0.2$, $n_2 = 30$,$\rho_2 = 0.7$. $\mu_1 = \mu_2 = 0$; $\sigma_1^2 = \sigma_2^2 = 1$.	Lo(exp)	0.8744	0.0640	0.0616	0.0378	23.63	1.04
	Lo(obs)	0.8759	0.0630	0.0611	0.0371	23.16	1.03
	Likelihood Ratio	0.9129	0.0450	0.0421	0.0186	11.59	1.07
	VB	0.8683	0.0670	0.0647	0.0409	25.53	1.04
	Proposed	**0.9479**	**0.0278**	**0.0243**	**0.0018**	**1.09**	**1.14**
$n_1 = 20$, $\rho_1 = 0.7$, $n_2 = 30$,$\rho_2 = 0.7$. $\mu_1 = \mu_2 = 0$; $\sigma_1^2 = \sigma_2^2 = 1$.	Lo(exp)	0.8443	0.0783	0.0774	0.0529	33.03	1.01
	Lo(obs)	0.8471	0.0767	0.0762	0.0515	32.16	1.01
	Likelihood Ratio	0.8893	0.0543	0.0564	0.0304	18.97	1.04
	VB	0.8436	0.0793	0.0771	0.0532	33.25	1.03
	Proposed	**0.9494**	**0.0248**	**0.0258**	**0.0005**	**0.31**	**1.04**
$n_1 = 20$, $\rho_1 = -0.2$, $n_2 = 30$,$\rho_2 = 0.7$. $\mu_1 = -1$,$\mu_2 = 0$; $\sigma_1^2 = \sigma_2^2 = 1$.	Lo(exp)	0.8935	0.0462	0.0603	0.0283	17.66	1.31
	Lo(obs)	0.8955	0.0451	0.0594	0.0273	17.03	1.32
	Likelihood Ratio	0.9225	0.0326	0.0449	0.0138	8.59	1.38
	Proposed	**0.9530**	**0.0252**	**0.0218**	**0.0017**	**1.06**	**1.16**
$n_1 = 20$, $\rho_1 = 0.7$, $n_2 = 30$,$\rho_2 = 0.7$. $\mu_1 = -1$,$\mu_2 = 0$; $\sigma_1^2 = \sigma_2^2 = 1$.	Lo(exp)	0.8483	0.0485	0.1032	0.0509	31.78	2.13
	Lo(obs)	0.8516	0.0473	0.1011	0.0492	30.75	2.14
	Likelihood Ratio	0.8901	0.0387	0.0712	0.0300	18.72	1.84
	Proposed	**0.9473**	**0.0235**	**0.0292**	**0.0029**	**1.78**	**1.24**
$n_1 = 20$, $\rho_1 = -0.2$, $n_2 = 30$,$\rho_2 = 0.7$. $\mu_1 = 0$,$\mu_2 = 1$; $\sigma_1^2 = \sigma_2^2 = 1$.	Lo(exp)	0.8707	0.0394	0.0899	0.0397	24.78	2.28
	Lo(obs)	0.8720	0.0386	0.0894	0.0390	24.38	2.32
	Likelihood Ratio	0.9144	0.0312	0.0544	0.0178	11.13	1.74
	Proposed	**0.9499**	**0.0258**	**0.0243**	**0.0008**	**0.47**	**1.06**
$n_1 = 20$, $\rho_1 = -0.2$, $n_2 = 20$,$\rho_2 = 0.7$. $\mu_1 = \mu_2 = 0$; $\sigma_1^2 = \sigma_2^2 = 1$.	Lo(exp)	0.8377	0.0811	0.0812	0.0562	35.09	1.00
	Lo(obs)	0.8407	0.0799	0.0794	0.0547	34.16	1.01
	Likelihood Ratio	0.8937	0.0529	0.0534	0.0282	17.59	1.01
	VB	0.8264	0.0861	0.0875	0.0618	38.63	1.02
	Proposed	**0.9491**	**0.0247**	**0.0262**	**0.0008**	**0.47**	**1.06**
$n_1 = 20$, $\rho_1 = 0.7$, $n_2 = 20$,$\rho_2 = 0.7$. $\mu_1 = \mu_2 = 0$; $\sigma_1^2 = \sigma_2^2 = 1$.	Lo(exp)	0.8264	0.0858	0.0878	0.0618	38.63	1.02
	Lo(obs)	0.8330	0.0817	0.0853	0.0585	36.56	1.04
	Likelihood Ratio	0.8790	0.0594	0.0616	0.0355	22.19	1.04
	VB	0.8276	0.0854	0.0870	0.0612	38.25	1.02
	Proposed	**0.9502**	**0.0239**	**0.0259**	**0.0010**	**0.62**	**1.08**
$n_1 = 20$, $\rho_1 = -0.2$, $n_2 = 20$,$\rho_2 = 0.7$. $\mu_1 = -1$,$\mu_2 = 0$; $\sigma_1^2 = \sigma_2^2 = 1$.	Lo(exp)	0.8570	0.0651	0.0779	0.0465	29.06	1.20
	Lo(obs)	0.8597	0.0637	0.0766	0.0452	28.22	1.20
	Likelihood Ratio	0.9004	0.0429	0.0567	0.0248	15.50	1.32
	Proposed	**0.9495**	**0.0272**	**0.0233**	**0.0020**	**1.22**	**1.17**
$n_1 = 20$, $\rho_1 = 0.7$, $n_2 = 20$,$\rho_2 = 0.7$. $\mu_1 = -1$,$\mu_2 = 0$; $\sigma_1^2 = \sigma_2^2 = 1$.	Lo(exp)	0.8303	0.0616	0.1081	0.0599	37.41	1.75
	Lo(obs)	0.8352	0.0592	0.1056	0.0574	35.88	1.78
	Likelihood Ratio	0.8782	0.0474	0.0744	0.0359	22.44	1.57
	Proposed	**0.9480**	**0.0248**	**0.0272**	**0.0012**	**0.75**	**1.10**
$n_1 = 20$, $\rho_1 = -0.2$, $n_2 = 20$,$\rho_2 = 0.7$. $\mu_1 = 0$,$\mu_2 = 1$; $\sigma_1^2 = \sigma_2^2 = 1$.	Lo(exp)	0.8416	0.0487	0.1097	0.0542	33.88	2.25
	Lo(obs)	0.8445	0.0477	0.1078	0.0528	32.97	2.26
	Likelihood Ratio	0.8994	0.0355	0.0651	0.0253	15.81	1.83
	Proposed	**0.9511**	**0.0241**	**0.0248**	**0.0006**	**0.34**	**1.03**

Chapter 5

Discussion and Future Work

Throughout this dissertation, we applied third-order likelihood-based asymptotic methods to provide statistical inference for a widely applied risk-adjusted return measures Sharpe ratio. There are mainly two scenarios where our methods can make great use of:

1. Limited Sample Size: Ideally, to mitigate various data biases, the larger the data set, the better the estimation would be. However, it is not only financial costly for researches to acquire data from data vendors, the existing databases are largely overlapped. In addition, data vendors do not provide information in a uniformed way. So it would be time consuming for researches to clean up the data even when acquiring larger data set is possible. In this condition, our proposed methods is more reliable because they are shown to be extremely accurate even when the sample size is small.

2. Time Aggregation Data: As pointed by Lo(2002), in many applications it is necessary to convert Sharpe ratio estimates from one frequency to another, however, the aggregation process may create a mis-specified manner, especially for non-IID returns. Thus, for example, we take yearly data for comparison, and the available sample size may be very small and our proposed method may be very useful.

The following highlight some possible future researches for the next few years.

- We can compare the general third order likelihood approach with the non-parametric Bootstrap method.

- All the current research is done in the frequentists' context as all the parameters are assumed to be unknown but fixed. It is also rea-

sonable to assume that the parameters have certain prior distribution. However, due to the stationary condition, the prior distributions should be chosen with caution as the stationary constraints on the parameters. Therefore, Bayesian approach for both inference and prediction will be taken into consideration in my future research.

- Autoregressive models and moving average models are well-known time series models with relatively simple structure. Similar research can be applied to more complex time series model structures, such as the stationary ARMA model which captures both autoregressive and moving average trend. Even try to apply generally into ARCH or GARCH.

- Models with gaussian error structure are widely studied as normal distribution is a simple and reasonable choice for the error term. Alternative to the normal distribution, the error terms can be assumed to follow other distributions, such as student-t distribution, Cauchy distribution. Therefore, similar studies can also be performed for models with non-Gaussian errors.

Bibliography

[1] Abramowitz, M., & Stegun, I. A. (1964). Handbook of mathematical functions: with formulas, graphs, and mathematical tables (No. 55). Courier Corporation.

[2] Akahira, M., Sato, M., & Torigoe, N. (1995). On the new approximation to non-central t-distributions. Journal of the Japan Statistical Society, 25(1), 1-18.

[3] Akahira, M. (1995). A higher order approximation to a percentage point of the non–central t-distribution. Communications in Statistics-Simulation and Computation, 24(3), 595-605.

[4] Bailey, D. H., & Lopez de Prado, M. (2012). The Sharpe ratio efficient frontier. Journal of Risk, 15(2), 13.

[5] Barndorff-Nielsen, O. E., & Chamberlin, S. R. (1991). An ancillary invariant modification of the signed log likelihood ratio. Scandinavian journal of statistics, 341-352.

[6] Barndorff-Nielsen, O., & Cox, D. R. (1979). Edgeworth and saddle-point approximations with statistical applications. Journal of the Royal Statistical Society. Series B (Methodological), 279-312.

[7] Barndorff-Nielsen, O. E., & Cox, D. R. (1984). Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. Journal of the Royal Statistical Society. Series B (Methodological), 483-495.

[8] Barndorff-Nielsen, O. E., & Cox, D. R. (1989). Asymptotic techniques for use in statistics. Chapman\ & Hall.

[9] Barndorff-Nielsen, O. (1980). Conditionality resolutions. Biometrika, 67(2), 293-310.

[10] Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator. Biometrika, 70(2), 343-365.

[11] Barndorff-Nielsen, O. E. (1984). On conditionality resolution and the likelihood ratio for curved exponential models. Scandinavian journal of statistics, 157-170.

[12] Barndorff-Nielsen, O. E. (1985). Confidence Limits from in the Single-Parameter Case. Scandinavian journal of statistics, 83-87.

[13] Barndorff-Nielsen, O. E. (1986a). Inference on full or partial parameters based on the standardized signed log likelihood ratio. Biometrika, 73, 307-322. Mathematical Reviews (MathSciNet): MR855891 Zentralblatt MATH, 605.

[14] Barndorff-Nielsen, O. E. (1986b). Likelihood and observed geometries. The Annals of Statistics, 856-873.

[15] Barndorff-Nielsen, O. E. (1990). Approximate interval probabilities. Journal of the Royal Statistical Society. Series B (Methodological), 485-496.

[16] Barndorff-Nielsen, O. E. (1990b). A note on the standardized signed log likelihood ratio. Scandinavian Journal of Statistics, 157-160.

[17] Barndorff-Nielsen, O. E. (1991). Modified signed log likelihood ratio. Biometrika, 78(3), 557-563.

[18] Barndorff-Nielsen, O. E. (2012). Parametric statistical models and likelihood (Vol. 50). Springer Science & Business Media.

[19] Barndorff-Nielsen, O. (2014). Information and exponential families in statistical theory. John Wiley & Sons.

[20] Basu, D. (1977). On the elimination of nuisance parameters. Journal of the American Statistical Association, 72(358), 355-366.

[21] Bentkus, V., Jing, B. Y., Shao, Q. M., & Zhou, W. (2007). Limiting distributions of the non-central t-statistic and their applications to the power of t-tests under non-normality. Bernoulli, 13(2), 346-364.

[22] Billingsley, P. (2008). Probability and measure. John Wiley & Sons.

[23] Bleistein, N., & Handelsman, R. A. (1975). Asymptotic expansions of integrals. Courier Corporation.

[24] Blæsild, P., & Jensen, J. L. (1985). Saddlepoint formulas for reproductive exponential models. Scandinavian journal of statistics, 193-202.

[25] Bochner, S. (2012). Harmonic analysis and the theory of probability. Courier Corporation.

[26] Brent, R. P. (2013). Algorithms for minimization without derivatives. Courier Corporation.

[27] Bulmer, M. G. (2012). Principles of statistics. Courier Corporation.

[28] Butler, R. W. (2007). Saddlepoint approximations with applications (Vol. 22). Cambridge University Press.

[29] Campbell, J. Y., Lo, A. W. C., & MacKinlay, A. C. (1997). The econometrics of financial markets (Vol. 2, pp. 149-180). Princeton, NJ: princeton University press.

[30] Christie, S. (2005). Is the Sharpe ratio useful in asset allocation?. Macquarie Applied Finance Centre Research Paper.

[31] Courant, R., & Hilbert, D. (1953). Methods of Mathematical Physics (Interscience, New York, 1953). Vol. I, 63.

[32] Cox, D. R., & Barndorff-Nielsen, O. E. (1994). Inference and asymptotics (Vol. 52). CRC Press.

[33] Cox, D. R., & Hinkley, D. V. (1979). Theoretical statistics. CRC Press.

[34] Cox, D. R. (1980). Local ancillarity. Biometrika, 67(2), 279-286.

[35] Cox, D. R. (1988). Some aspects of conditional and asymptotic inference: A review. Sankhyā: The Indian Journal of Statistics, Series A, 314-337.

[36] Cramér, H. (1999). Mathematical methods of statistics (Vol. 9). Princeton university press.

[37] Cuppens, R. (2014). Decomposition of multivariate probabilities (Vol. 29). Academic Press.

[38] Daniels, H. E. (1954). Saddlepoint approximations in statistics. The Annals of Mathematical Statistics, 631-650.

[39] Daniels, H. E. (1958). Discussion of" The regression analysis of binary sequences" by DR Cox. J. Royal Statist. Soc. B, 20, 236-238.

[40] Daniels, H. E. (1987). Tail probability approximations. International Statistical Review/Revue Internationale de Statistique, 37-48.

[41] Davison, A. C., & Hinkley, D. V. (1988). Saddlepoint approximations in resampling methods. Biometrika, 75(3), 417-431.

[42] Dawid, A. P. (1991). Fisherian inference in likelihood and prequential frames of reference. Journal of the Royal Statistical Society. Series B (Methodological), 79-109.

[43] De Bruijn, N. G. (1970). Asymptotic methods in analysis (Vol. 4). Courier Corporation.

[44] DiCICCIO, T. J., Field, C. A., & Fraser, D. A. S. (1990). Approximations of marginal tail probabilities and inference for scalar parameters. Biometrika, 77(1), 77-95.

[45] DiCiccio, T. J., & Martin, M. A. (1993). Simple modifications for signed roots of likelihood ratio statistics. Journal of the Royal Statistical Society. Series B (Methodological), 305-316. Doganaksoy, N., & Schmee, J. (1993). Comparisons of approximate confidence intervals for distributions used in life-data analysis. Technometrics, 35(2), 175-184.

[46] Durbin, J. (1980). Approximations for densities of sufficient estimators. Biometrika, 67(2), 311-333.

[47] Efron, B., & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. Biometrika, 65(3), 457-483.

[48] Efron, B. (1981). Nonparametric standard errors and confidence intervals. The Canadian Journal of Statistics/La Revue Canadienne de Statistique, 139-158.

[49] Engle, R. F. (1984). Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. Handbook of econometrics, 2, 775-826.

[50] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 309-368.

[51] Fisher, R. A. (1925). Theory of statistical estimation. In Mathematical Proceedings of the Cambridge Philosophical Society (Vol. 22, No. 05, pp. 700-725). Cambridge University Press.

[52] Fisher, R. A. (1934). Two new properties of mathematical likelihood. Proceedings of the Royal Society of London. Series A, 144(852), 285-307.

[53] Fog, A. (2008). Calculation methods for Wallenius' noncentral hypergeometric distribution. Communications in Statistics—Simulation and Computation®, 37(2), 258-273.

[54] Fraser, D. A. S., & Fraser, D. A. (1968). The structure of inference (Vol. 23). New York: Wiley.

[55] Fraser, D. A. S., Reid, N., Li, R., & Wong, A. (2003). p-value formulas from likelihood asymptotics bridging the singularities. Journal of Statistical Research, 37(1), 1-15.

[56] Fraser, D. A. S., Reid, N., & Wong, A. (1991). Exponential linear models: a two-pass procedure for saddlepoint approximation. Journal of the Royal Statistical Society. Series B (Methodological), 483-492.

[57] Fraser, D. A. S., Reid, N., & Wu, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. Biometrika, 86(2), 249-264.

[58] Fraser, D. A. S., & Reid, N. (1988). On conditional inference for a real parameter: a differential approach on the sample space. Biometrika, 75(2), 251-264.

[59] Fraser, D. A. S., & Reid, N. (1993). Third order asymptotic models: Likelihood functions leading to accurate approximations for distribution functions. Statist. Sinica, 3, 67-82.

[60] Fraser, D. A. S., & Reid, N. (1995). Ancillaries and third order significance. Utilitas Mathematica, 7, 33-53.

[61] Fraser, D. A. S. (1956). Sufficient statistics with nuisance parameters. The Annals of Mathematical Statistics, 27(3), 838-842.

[62] Fraser, D. A. S. (1964). Local conditional sufficiency. Journal of the Royal Statistical Society. Series B (Methodological), 52-62.

[63] Fraser, D. A. S. (1966). Sufficiency for regular models. Sankhyā: The Indian Journal of Statistics, Series A, 137-144.

[64] Fraser, D. A. S. (1976). Probability and statistics: Theory and applications. North Scituate, Mass.: Duxbury Press.

[65] Fraser, D. A. S. (1979). Inference and linear models. New York: McGraw-Hill.

[66] Fraser, D. A. S. (1988). Normed likelihood as saddlepoint approximation. Journal of Multivariate Analysis, 27(1), 181-193.

[67] Fraser, D. A. S. (1990). Tail probabilities from observed likelihoods. Biometrika, 77(1), 65-76.

[68] Fraser, D. A. S. (1991). Statistical inference: Likelihood to significance. Journal of the American Statistical Association, 86(414), 258-265.

[69] Greene, W. H. (2003). Econometric analysis. Pearson Education India.

[70] He, H., & Leland, H. (1993). On equilibrium asset price processes. Review of Financial Studies, 6(3), 593-617.

[71] Hinkley, D. V. (1980). Likelihood as approximate pivotal distribution. Biometrika, 67(2), 287-292.

[72] Hogben, D., Pinkham, R. S., & Wilk, M. B. (1961). The moments of the non-central t-distribution. Biometrika, 465-468.

[73] Hogg, R. V., Tanis, E. A., & Rao, M. J. M. (1977). Probability and statistical inference (Vol. 993). New York: Macmillan.

[74] Huzurbazar, V. S. (1976). Sufficient statistics: selected contributions. A. M. Kshirsagar (Ed.). Dekker.

[75] Jensen, J. L. (1988). Uniform saddlepoint approximations. Advances in applied probability, 622-634.

[76] Jensen, J. L. (1992). The modified signed likelihood statistic and saddlepoint approximations. Biometrika, 79(4), 693-703.

[77] Jensen, J. L. (1995). Saddlepoint approximations (No. 16). Oxford University Press.

[78] Jobson, J. D., & Korkie, B. M. (1981). Performance hypothesis testing with the Sharpe and Treynor measures. Journal of Finance, 889-908.

[79] Johnson, N. L., & Welch, B. L. (1940). Applications of the non-central t-distribution. Biometrika, 362-389.

[80] Jørgensen, B. (1994). The rules of conditional inference: Is there a universal definition of nonformation? Journal of the Italian Statistical Society, 3(3), 355-384.

[81] Kalbfleisch, J. D. (1975). Sufficiency and conditionality. Biometrika, 62(2), 251-259.

[82] Kass, R. E. (1989). The geometry of asymptotic inference. Statistical Science, 188-219.

[83] Kendall, M. G., & Stuart, A. (1969). The Advanced Theory of Statistics (Volume 1) Griffin.

[84] Kolassa, J. E. (2006). Series approximation methods in statistics (Vol. 88). Springer Science & Business Media.

[85] Ledoit, O., & Wolf, M. (2008). Robust performance hypothesis testing with the Sharpe ratio. Journal of Empirical Finance, 15(5), 850-859.

[86] Lehmann, E. L., & Casella, G. (1998). Theory of point estimation (Vol. 31). Springer Science & Business Media.

[87] Lehmann, E. L. (1986). Testing Statistical Hypotheses Wadsworth & Brooks. Cole, Pacific Grove, California.

[88] Leung, P. L., & Wong, W. K. (2006). On testing the equality of the multiple Sharpe Ratios, with application on the evaluation of iShares. Available at SSRN 907270.

[89] Lindsey, J. K. (1996). Parametric statistical inference. Oxford University Press.

[90] Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. The review of economics and statistics, 13-37.

[91] Liu, Y., Rekkas, M., & Wong, A. (2012). Inference for the Sharpe ratio using a likelihood-based approach. Journal of Probability and Statistics, 2012.

[92] Lo, A. W. (2002). The statistics of Sharpe ratios. Financial Analysts Journal, 58(4), 36-52.

[93] Lugannani, R., & Rice, S. (1980). Saddle point approximation for the distribution of the sum of independent random variables. Advances in applied probability, 475-490.

[94] Lukacs, E., & King, E. P. (1954). A property of the normal distribution. The Annals of Mathematical Statistics, 25(2), 389-394.

[95] Lukacs, E. (1970). Characteristics functions. Griffin, London.

[96] Markowitz, H. (1952). Portfolio selection*. The journal of finance, 7(1), 77-91.

[97] McCullagh, P. (1984). Local sufficiency. Biometrika, 71(2), 233-244.

[98] McCullagh, P. (1987). Tensor methods in statistics (Vol. 161). London: Chapman and Hall.

[99] Memmel, C. (2003). Performance hypothesis testing with the Sharpe ratio. Finance Letters, 1(1).

[100] Mertens, E. (2002). Comments on variance of the IID estimator in Lo (2002). Working paper.

[101] Miller, R. E., & Gehr, A. K. (1978). Sample size bias and Sharpe's performance measure: A note. Journal of Financial and Quantitative Analysis, 13(05), 943-946.

[102] Modigliani, F., & Modigliani, L. (1997). Risk-adjusted performance. The Journal of Portfolio Management, 23(2), 45-54.

[103] Neyman, J., & Pearson, E. S. (1992). On the problem of the most efficient tests of statistical hypotheses (pp. 73-108). Springer New York.

[104] Oberhettinger, F. (2014). Fourier transforms of distributions and their inverses: a collection of tables (Vol. 16). Academic press.

[105] Olver, F. W., Lozier, D. W., Boisvert, R. F., & Clark, C. W. (2010). Digital library of mathematical functions. National Institute of Standards and Technology from http://dlmf. nist. gov/(release date 2011-07-01), Washington, DC.

[106] Pierce, D. A., & Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families. Journal of the Royal Statistical Society. Series B (Methodological), 701-737.

[107] Pinsky, M. A. (2002). Introduction to Fourier analysis and wavelets (pp. 181-194). Pacific Grove: Brooks/Cole.

[108] Pitman, E. J. G. (1936). Sufficient statistics and intrinsic accuracy. In Mathematical Proceedings of the Cambridge Philosophical Society (Vol. 32, No. 04, pp. 567-579). Cambridge University Press.

[109] Reid, N. (1988). Saddlepoint methods and statistical inference. Statistical Science, 213-227.

[110] Reid, N. (1995). The roles of conditioning in inference. Statistical Science, 138-157.

[111] Reid, N. (1996). Likelihood and higher order approximations to tail areas: A review and annotated bibliography. Canadian Journal of Statistics, 24(2), 141-166.

[112] Schervish, M. J. (1995). Theory of statistics. Springer.

[113] Scholz, F. (2007). Applications of the Noncentral t–Distribution. Stat 498B, Industrial Statistics.

[114] Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk*. The journal of finance, 19(3), 425-442.

[115] Sharpe, W. F. (1966). Mutual fund performance. Journal of business, 119-138.

[116] Shephard, N. G. (1991). From characteristic function to distribution function: a simple framework for the theory. Econometric theory, 7(04), 519-529.

[117] Skovgaard, I. (1986). Successive improvement of the order of ancillarity. Biometrika, 73(2), 516-519.

[118] Skovgaard, I. M. (1987). Saddlepoint expansions for conditional distributions. Journal of Applied Probability, 875-887.

[119] Skovgaard, I. M. (1990). On the density of minimum contrast estimators. The Annals of Statistics, 18(2), 779-789.

[120] Tricomi, F. G., & Erdélyi, A. (1951). The asymptotic expansion of a ratio of gamma functions. Pacific J. Math, 1(1), 133-142.

[121] Van Belle, G. (2011). Statistical rules of thumb (Vol. 699). John Wiley & Sons.

[122] Walck, C. (2007). Handbook on statistical distributions for experimentalists.

[123] Wald, A. (1941). Asymptotically most powerful tests of statistical hypotheses. The Annals of Mathematical Statistics, 12(1), 1-19.

[124] Wang, S. (1992). General saddlepoint approximations in the bootstrap. Statistics & probability letters, 13(1), 61-66.

[125] Wasserman, L. (2013). All of statistics: a concise course in statistical inference. Springer Science & Business Media.

[126] Wendel, J. G. (1961). The non-absolute convergence of Gil-Pelaez'inversion integral. The Annals of Mathematical Statistics, 32(1), 338-339.

[127] Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. The Annals of Mathematical Statistics, 9(1), 60-62.

[128] Wright, J. A., Yam, S. C. P., & Yung, S. P. (2011). A note on a test for the equality of multiple Sharpe ratios and its application on the evaluation of iShares. Technical report, 2012. URL http://www. sta. cuhk. edu. hk/scpy/Preprints/John% 20Wright/A% 20test% 20for% 20the% 20equality% 20of% 20multiple% 20Sharpe% 20ratios. pdf. to appear.

[129] Wright, J. A., Yam, S. C. P., & Yung, S. P. (2014). A test for the equality of multiple Sharpe ratios. The Journal of Risk, 16(4), 3.

Glossary of Notation

$O(\cdot)$	Order of Convergence
μ	Expected Return
σ	Standard Deviation
SR	Population Sharpe Ratio
r_f	Risk-free Rate of Return
P_t	Price of an Asset at Date t
R_t	Net Return
g_t	Gross Return
r_t	Log Return
ω_t	Brownian Motion Process
$N(\mu, \sigma^2)$	Normal Distribution with Mean μ and Variance σ^2
$LN(\mu, \sigma^2)$	Lognormal Distribution with Location μ and Scale σ
$T_\nu(\delta)$	Noncentral T Statistics with Degrees of Freedom ν and Noncentrality Value δ
χ^2_ν	Chi-square Distribution with Degrees of Freedom ν
$t_{1-\alpha,\nu}$	$1-\alpha$ Quartile of the Central T Distribution with ν Degrees of Freedom
$t_{1-\alpha,\nu}(\delta)$	$1-\alpha$ Quartile of the Noncentral T Distribution
$F(\cdot)$	Cumulative Distribution Function
$f(\cdot)$	Probability Density Function
$E[\cdot]$	Expected Value
$\Gamma(\cdot)$	Gamma Function
$Var(\cdot)$	Variance
m'_k	Uncentered k th Order Raw Moment
m_k	k th Order Central Moment
α_k	k th Order Standardized Moment
∇g	Gradient of Function g
z_α	Upper 100α percentile of the Standard Normal Distribution
κ_n	n th Order Cumulant
$\psi(\theta)$	Parameter of Interest
$I(\theta)$	Expected Fisher Full Information Matrix
$j(\theta)$	Observed Fisher Full Information Matrix
$\mathbb{R}^n$	Set of Ordered n-Tuples of Real Numbers
$\dim(\cdot)$	Dimension

$S(Y)$	Sufficient Statistic
$\eta(\theta)$	Natural Parameter
J	Jacobian of Transformation
$\lambda(\theta)$	Nuisance Parameter
$\mathcal{L}(\theta)$	Likelihood Function
$\ell(\theta)$	Log-likelihood Function
$s(\theta)$	Score Function
$j_{\lambda\lambda'}(\theta)$	Observed Nuisance Information Matrix
$S(\theta)$	Rao Statistic or Lagrange Multiplier Statistic
$q(\theta)$	Wald statistic
$R(\theta)$	Signed Log-likelihood Ratio Statistic
$\tilde{\ell}(\theta)$	Tilted Log-likelihood Function
α	Lagrange Multiplier
$\hat{\theta}_\psi$	Constrained MLE
$H(\theta, \alpha)$	Lagrangian Function
$\tilde{j}(\theta)$	Observed Information Matrix for Tilted Log-likelihood Function
$p(\theta)$	p-value Function
$M_X(t)$	Moment Generating Function
$C_X(t)$	Central Moment Generating Function
$\varphi_X(t)$	Characteristic Function
$K_X(t)$	Cumulant Generating Function
ρ_n	n th Order Standardized Cumulant
$H_n(x)$	Hermite polynomials
$\phi(\cdot)$	The Probability Density Function for Standard Normal Distribution
$\varphi(\theta)$	Canonical Parameter
$\chi(\theta)$	Scalar Parameter of Interest $\psi(\theta)$ in $\varphi(\theta)$ Scale
$\mathbf{V}$	Ancillary Directions
$Q(\theta)$	Standardized Maximum Likelihood Departure in Canonical Parameter $\varphi(\theta)$ scale
CP	Proportion of True Sharpe Ratio Falls within 95% Confidence Interval
LE	Proportion of True Sharpe Ratio Falls below the Lower Limit of 95% Confidence Interval
UE	Proportion of the true Sharpe Ratio Falls above the Upper Limit of 95% Confidence Interval
AB	Average Bias Defined as $AB = (\lvert LE - 0.025\rvert + \lvert UE - 0.025\rvert)/2$
SE	Standard Error
SY	Degree of Symmetry Defined as $SY = \max\left\{\frac{LE}{UE}, \frac{UE}{LE}\right\}$
ρ	Correlation coefficient (for Two Samples IID case) or Autocorrelated Coefficient (for AR(1) Structure)
SR_l	J.S. Sharpe Ratio